2014

Different-based methods in nonparametric regression models

Wenlin Dai
Hong Kong Baptist University

Difference-based Methods in Nonparametric Regression Models

DAI Wenlin

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

Principal Supervisor: Dr. TONG Tiejun

Hong Kong Baptist University

August 2014

Declaration

I hereby declare that this thesis represents my own work which has been done after registration for the degree of PhD at Hong Kong Baptist University, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

Signature: _____

Date: August 2014

Abstract

This thesis develops some new difference-based methods for nonparametric regression models.

The first part of this thesis focuses on the variance estimation for nonparametric models with various settings. In Chapter 2, a unified framework of variance estimator is proposed for a model with smooth mean function. This framework combines the higher order difference sequence with least squares method and greatly extends the literature, including most of existing methods as special cases. We derive the asymptotic mean squared errors and make both theoretical and numerical comparison for various estimators within the system. Based on the dramatic interaction of ordinary difference sequences and least squares method, we eventually find a uniformly satisfactory estimator for all the settings, solving the challenging problem of sequence selection. In Chapter 3, three methods are developed for the variance estimation in the repeated measurement setting. Both their asymptotic properties and finite sample performance are explored. The sequencing method is shown to be the most adaptive while the sample variance method and the partitioning method are shown to outperform in certain cases. In Chapter 4, we propose a pairwise regression method for estimating the residual variance. Specifically, we regress the squared difference between observations on the squared distance between design points, and then estimate the residual variance as the intercept. Unlike most existing difference-based estimators that require a smooth regression function, our method applies to regression models with jump discontinuities. And it also applies to the situations where the design points are unequally spaced.

The smoothness assumption of the nonparametric regression function is quite critical for the curve fitting and the residual variance estimation. The second part (Chapter 5) concentrates on the discontinuities detection for the mean function. In particular, we revisit the difference-based method in Müller and Stadtmüller (1999) and propose to improve it. To achieve the goal, we first reveal that their method is less efficient due to the inappropriate choice of the response variable in their linear

regression model. We then propose a new regression model for estimating the residual variance and the total amount of discontinuities simultaneously. In both theory and simulations, we show that the proposed variance estimator has a smaller MSE compared to their estimator, whereas the efficiency of the estimators for the total amount of discontinuities remain unchanged. Finally, we construct a new test procedure for detection using the newly proposed estimations; and via simulation studies, we demonstrate that our new test procedure outperforms the existing one in most settings.

At the beginning of Chapter 6, a series of new difference sequences is defined to complete the span between the optimal sequence and the ordinary sequence. The variance estimators using proposed sequences are shown to be quite robust and achieve smallest mean square errors for most of general settings. Then, the difference-based methods for variance function estimation are generally discussed.

Keywords: Asymptotic normality, Difference-based estimator, Difference sequence, Jump point, Least square, Nonparametric regression, Pairwise regression, Repeated measurement, Residual variance

Acknowledgements

I would like to express my greatest gratitude to my supervisor, Dr. Tong Tiejun, for his patient guidance, encouragement and advice throughout my Ph.D study. I am so lucky to have a supervisor who cares so much about my work, and who responds to my questions and queries so promptly. This thesis would not have been possible without his constant support.

I would also like to thank all the faculty and staff in the Department of Mathematics for providing me such a supportive environment. In particular, my sincere thanks go to Professor Zhu Lixing and Dr. Peng Heng, for their invaluable advices and stimulating comments. Thanks also go to my friends who have encouraged me to pursuit my dream when I was lost, comforted me when I got dispirited, and accompanied me when I felt lonely. In addition, I would like to thank the following graduate students in the Department of Mathematics: Wang Tao, Chen Chi, Dong Kai, Guo Xu, Zhu Xuehu, Wu Qin, and also my football teammates for making my Ph.D life full of fun and cherished memories.

Last but not least, I would like to thank: my parents, Dai Zhiqiang and Wu Yufeng, for giving birth to me at the first place and for supporting me spiritually throughout my life; my sister, Dai Wenjing, for accompanying our parents when I was absent; and my wife, Zhang Xinying, for her constant love, understanding and accompany. And most importantly, thank her for promoting me to be a daddy. I owe her everything!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Nonparametric Regression Model

Nonparametric regression models are widely used to explore the relationship between the explanatory variable X and the response variable Y, and have been extensively studied in the past several decades. This type of models is superior for the *nonparametric* feature, which promises the flexibility of the regression function.

The estimation of the regression function is quite important for such models. A well fitted regression function is required for various purposes, including description of the relationship between two variables, prediction of observations at new experiment points, and substitution for missing values, and etc. Some important features (e.g. monotonicity, smoothness and unimodality) of the regression function or its derivatives are also of great interest in many situations (Bowman et al.; 1998; Müller and Stadtmüller; 1999; Ghosal et al.; 2000; Eggermont and LaRiccia; 2000). There is a large body of literature on the estimation of the mean function. Essentially, they estimates each response as a weighted average of all the observations and the weights are determined through different techniques. The kernel estimators, the local linear estimators and the smoothing spline estimators (Wand and Jones; 1995; Fan and Gijbels; 1996; Wahba; 1990), are most popularly used by researchers.

Apart from the mean function, the estimation of the residual variance σ^2 has also been recognized as an equally important problem and has attracted plenty of attention

(Dette et al.; 1998). An accurate yet economic estimator of σ^2 is required in most of aspects mentioned above for nonparametric regression, e.g., in choosing the amount of smoothing, in testing the goodness of fit, and in the construction of confidence intervals (Rice; 1984; Eubank and Spiegelman; 1990; Gasser et al.; 1991; Härdle and Tsybakov; 1997). The variance estimation methods can be generally divided into two classes: *the residual-based methods* and *the difference-based methods*. We will briefly introduce both types of methods in next two sections.

1.2 Residual-based Method

Consider the nonparametric regression model of the form

$$Y_i = g(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where Y_i are the observations, g is an unknown mean function, x_i are the design points, and ε_i are the independent and identically distributed (i.i.d.) random errors with mean zero and variance σ^2.

For residual-based methods, one first estimates the mean function g with some nonparametric approach instructed in Section 1.1. Then, the variance estimators are generated with the residuals based on that approximation of the mean function. For instance, if we fit model (1.1) using a kernel-based method, the residuals are $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ and we can estimate σ^2 by $\hat{\sigma}^2 = (n-\nu)^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = (n-\nu)^{-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where ν is the degrees of freedom for the fitted model. With an appropriate bandwidth, the minimum mean squared error (MSE) of $\hat{\sigma}^2$ is given as

$$\mathrm{MSE}(\hat{\sigma}^2) = \mathrm{var}(\hat{\sigma}^2) + (E(\hat{\sigma}^2) - \sigma^2)^2 = n^{-1}\mathrm{var}(\varepsilon^2) + o(n^{-1}). \tag{1.2}$$

Hall and Marron (1990) showed that this MSE is asymptotically optimal in a minimax sense. As pointed out in Dette et al. (1998), residual-based estimators depend heavily on the delicate choice of tuning parameters, and their practical applications are somewhat limited.

1.3 Difference-based Method

In practice, for some reasons such as in choosing the amount of smoothing or testing goodness of fit, one may need an estimate of σ^2 that is independent of the fitted mean function $\hat{g}$ (Eubank and Spiegelman; 1990; Ye; 1998; Wang; 2011). To achieve this, another popular category of estimators have also been developed in the literature. The essential idea in these methods is to use the differences between nearby observations to remove the trend in the mean function. The estimators constructed with this idea are generally named difference-based estimators.

The idea of difference-based estimators can date back to von Neumann (1941), and it was first introduced into nonparametric regression by Rice (1984) with form

$$\hat{\sigma}_{\mathrm{R}}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_i - Y_{i+1})^2.$$

Afterwards, Gasser et al. (1986) proposed a second order estimators for the sake of bias reduction. Hall et al. (1990) proposed a general form for the difference-based estimators with an arbitrary fixed order. More difference-based estimators are investigated in Buckley and Eagleson (1989); Seifert et al. (1993); Dette et al. (1998). Hall et al. (1991); Munk et al. (2005) explored difference-based variance estimators for multivariate cases.

Dette et al. (1998) pointed out that none of the fixed order difference-based estimators can achieve the asymptotically optimal rate 1.2 as that is achieved by the residual-based estimators (Hall and Marron; 1990). To improve the literature, Müller et al. (2003), Tong et al. (2008) and Du and Schick (2009) proposed covariate-matched U-statistic estimators for the residual variance. In addition, Müller and Stadtmüller (1999), Tong and Wang (2005) and Tong et al. (2013) proposed some least squares methods for estimating the residual variance, motivated by the fact that the Rice estimator is always positively biased. And these methods manage to obtain the asymptotically optimal rate 1.2.

In addition to the residual variance estimation, the difference-based methods can also be used in various areas, e.g. variance function estimation (Brown and Levine; 2007; Cai et al.; 2009), heteroscedasticity test (Dette and Munk; 1998), jump detec-

tion (Müller and Stadtmüller; 1999), partial linear model estimation (Yatchew; 1997; You et al.; 2010; Wang et al.; 2011).

1.4 Overall Structure

The rest of this thesis is organized as follows.

In Chapter 2, we consider the variance estimation problem in classic nonparametric regression model. A unified framework of variance estimator is proposed, combining the higher order difference sequence and least squares method. The framework greatly extends the literature, with most of existing methods as special cases. We derive the asymptotic mean squared errors and make both theoretical and numerical comparison of various estimators involved in the framework. Based on the dramatic interaction of ordinary difference sequences and least squares method, we eventually find a uniformly satisfactory estimator for all the settings.

In Chapter 3, we consider the residual variance estimation in nonparametric regression models with both balanced and unbalanced repeated measurement data. Specifically, we propose three difference-based methods: the sample variance method by using only the variation within design points, the partitioning method by using only the variation between design points, and the sequencing method by using both between and within variations. We investigate the statistical properties of the proposed estimators and establish the optimality of the sequencing estimator. In addition, we have made some suggestions for practical implementation through extensive simulation studies.

In Chapter 4, we proposed a pairwise regression method for estimating the residual variance in nonparametric regression models with jump discontinuities. The proposed method generalizes the existing methods from different points of view and has several important merits. In particular, is is superior for the flexibility in eliminating the effect of potential jumps in the mean function and the applicability in both equally and unequally design settings. Unlike the existing competitors, our method provides a direct way to estimate the residual variance without the estimations of mean function

and jump points.

In Chapter 5, we consider the detection problem whether the mean function is smooth or not. To achieve the goal, we first reveal that the method in Müller and Stadtmüller (1999) is less efficient due to the inappropriate choice of the response variable in their linear regression model. We then propose a new regression model for estimating the residual variance σ^2 and the total amount of discontinuity γ simultaneously, and also propose a new test procedure for detecting the smoothness. In both theory and simulations, we show that the proposed variance estimator has a smaller MSE compared to the estimator in Müller and Stadtmüller (1999). Consequently, the proposed new test procedure also improves the existing test in Müller and Stadtmüller (1999).

In Chapter 6, we introduce some future directions along with this thesis. In Section 6.1, we define a series of new difference sequences that include most existing sequences as special cases. Numerical results illustrate that the proposed sequences are quite robust and achieve the smallest MSEs for a wide range of settings. In future, we will investigate the asymptotic results of the new sequences and also suggest practical rules for the tuning parameter selection. In Sections 6.2, we discuss the difference-based estimation of variance function in general.

Chapter 2

A Unified Framework for Variance Estimation in Nonparametric Regression

2.1 Introduction

As we mentioned in Chapter 1, residual variance estimation is quite important in nonparametric regression models. Various difference based methods have been investigated during the past several decades. In this chapter, we first make a comprehensive review of the existing methods and then propose a unified framework of variance estimators.

Consider the nonparametric regression model of the form

$$Y_i = g(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where Y_i are the observations, g is an unknown mean function, x_i are the design points, and ε_i are the independent and identically distributed (i.i.d.) random errors with mean zero and variance σ^2.

Let $r > 0$ be an integer number and define $(d_0, \ldots, d_r)$ as a difference sequence

with

$$\sum_{j=0}^{r} d_j = 0 \qquad \text{and} \qquad \sum_{j=0}^{r} d_j^2 = 1. \tag{2.2}$$

We also assume that $d_0 d_r \neq 0$, $d_0 > 0$, and $d_j = 0$ for $j < 0$ and $j > r$. With the sequence $(d_0, \ldots, d_r)$, Hall et al. (1990) proposed the following estimator for σ^2,

$$\hat{\sigma}^2(r) = \frac{1}{n-r} \sum_{i=1}^{n-r} \left(\sum_{j=0}^{r} d_j Y_{j+i} \right)^2. \tag{2.3}$$

Estimators with form (2.3) are referred to as difference-based estimators with order r. When $r = 1$, the unique solution to constraint (2.2) is $(d_0, d_1) = (2^{-1/2}, -2^{-1/2})$ and it reduces to the first-order difference-based estimator in Rice (1984),

$$\hat{\sigma}^2_{\mathrm{R}} = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_i - Y_{i+1})^2, \tag{2.4}$$

an idea originated in von Neumann (1941). When $r \geq 2$, however, there are infinitely many solutions for the sequence $(d_0, \ldots, d_r)$ with constraint (2.2). Among the solutions, it is well known that the following two sequences are most widely used in practice: (i) the optimal sequence and (ii) the ordinary sequence.

The optimal sequence is obtained by minimizing the asymptotic MSE of the estimator with form (2.3). Under some smoothness condition on g, Hall et al. (1990) showed that the effect of g on the estimation bias is asymptotically negligible. Then to minimize the asymptotic MSE of the estimator is equivalent to minimizing its asymptotic variance. This leads to the optimal sequence satisfying $\sum_{j=0}^{r} d_j d_{j+i} = -1/2r$ for $1 \leq i \leq r$. In addition, it was shown that the optimal sequence of any order r is unique up to reversal of the sequence. For the special case of $r = 2$, we have $(d_0, d_1, d_2) = (0.809, -0.5, -0.309)$ and the corresponding estimator is

$$\hat{\sigma}^2_{\mathrm{opt}}(2) = \frac{1}{n-2} \sum_{i=1}^{n-2} (0.809 Y_i - 0.5 Y_{i+1} - 0.309 Y_{i+2})^2.$$

We refer to the estimator (2.3) with the optimal sequence as $\hat{\sigma}^2_{\mathrm{opt}}(r)$.

For small sample sizes, however, the bias term of difference-based estimators is often non-negligible, especially when the mean function g is rough. To reduce the

bias, researchers have also recommended to use the sequence

$$d_j = (-1)^j \binom{2r}{r}^{-1/2} \binom{r}{j}, \qquad j = 0, \ldots, r. \tag{2.5}$$

Note that sequence (2.5) is widely used in numerical differentiation and is so-called the ordinary sequence. With this sequence, the bias of estimator (2.3) vanishes for polynomials up to degree $r-1$. For the special case of $r = 2$, we have $(d_0, d_1, d_2) = (6^{-1/2}, -(2/3)^{1/2}, 6^{-1/2})$ and the resulting estimator is

$$\hat{\sigma}^2_{\text{ord}}(2) = \frac{1}{6(n-2)} \sum_{i=1}^{n-2} (Y_i - 2Y_{i+1} + Y_{i+2})^2,$$

which was proposed in Gasser et al. (1986). We refer to the estimator (2.3) with the ordinary sequence as $\hat{\sigma}^2_{\text{ord}}(r)$. Noting that $\hat{\sigma}^2_{\text{R}} = \hat{\sigma}^2_{\text{opt}}(1) = \hat{\sigma}^2_{\text{ord}}(1)$, the Rice estimator thus serves as a connecting point between the optimal estimators and the ordinary estimators.

A properly chosen difference sequence is quite important when we estimate the residual variance using difference based estimators, since it has greatly effect on the properties of the estimator. As shown in Dette et al. (1998), the asymptotic variance of $\hat{\sigma}^2_{\text{ord}}(r)$ is always larger than that of $\hat{\sigma}^2_{\text{opt}}(r)$. In particular, $\text{var}(\hat{\sigma}^2_{\text{ord}}(r))/\text{var}(\hat{\sigma}^2_{\text{opt}}(r)) \approx \sqrt{\pi r/2}$ when r is large. On the other hand, the asymptotic bias of $\hat{\sigma}^2_{\text{ord}}(r)$ is always smaller than that of $\hat{\sigma}^2_{\text{opt}}(r)$, by noting that $E(\hat{\sigma}^2_{\text{ord}}(r)) = \sigma^2 + O(n^{-2r})$ and $E(\hat{\sigma}^2_{\text{opt}}(r)) = \sigma^2 + O(n^{-2})$. In view of this trade-off, Dette et al. (1998) suggested to use the ordinary sequence if the sample size is small and the signal-to-noise ratio is large; otherwise, the optimal sequence should be used. Although very easy to implement, this rule of thumb can be useless in practice since the signal-to-noise ratio is rarely known. In addition, it is never known in practice when the sample size is large enough so that the bias term can be negligible. Some other criteria, such as the graphical method in Buckley and Eagleson (1989), have also been suggested in the literature. We note that most of these methods are quite subjective. Up to now, the choice of difference sequence remains arbitrary and a controversial issue in nonparametric regression. For instance, Munk and Dette (1998), Hall and Heckman (2000) and Shen and Brown (2006) used the Rice estimator. Härdle and Kneip (1999),

Munk et al. (2005), Einmahl and Van Keilegom (2008), Cai et al. (2009) and Dette and Hetzler (2009) used the ordinary estimators. Yatchew (1999), Brown and Levine (2007), Benko et al. (2009), Pendakur and Sperlich (2010) and Gijbels et al. (2010) used the optimal estimators.

The main goal of the chapter is to provide an ingenious solution for the very challenging difference sequence selection problem. To achieve this, we propose a unified framework for variance estimation that combines the higher-order difference sequence and the linear regression technique systematically. By this combination, we show that

(a) the unified framework generates a very large family of estimators that includes most existing estimators as special cases;

(b) the existing difference-based estimators are all asymptotically suboptimal in the proposed family of estimators; and

(c) in the unified framework, the ordinary sequence can be often used no matter if the sample size is small or if the signal-to-noise ratio is large.

The rest of the chapter is organized as follows. In Section 2.2, we propose the general methodology for estimating the residual variance. In addition, we draw some connections between the proposed estimator and some other estimators in the literature. In Section 2.3, we investigate the statistical properties of the proposed estimator and make a comprehensive comparison of various estimators based on their theoretical properties. In Section 2.4, we carry out extensive simulation studies to evaluate the performance of the proposed estimator and give our suggestions for practical use. In Section 2.5, we apply the proposed method to two real data examples for illustration. All technical proofs are provided in Section 2.6.

2.2 A unified framework

2.2.1 Methodology

The aforementioned difference-based estimators, including the optimal estimators and ordinary estimators, are popular in practice owing to their independence of curve fitting and the ease of implementation. Nevertheless, noting that

$$\mathrm{MSE}(\hat{\sigma}^2_{\mathrm{opt}}(r)) = \min_{d_0,\ldots,d_r} \mathrm{MSE}(\hat{\sigma}^2(r)) = n^{-1}\left(\mathrm{var}(\varepsilon^2) + r^{-1}\sigma^4\right) + o(n^{-1}),$$

none of the fixed-order difference-based estimators (2.3) can attain the asymptotically optimal rate of MSE (1.2), a property possessed usually by the residual-based estimators (Buckley et al.; 1988; Buckley and Eagleson; 1989; Hall and Marron; 1990) or the covariate-matched U-statistic estimators (Müller et al.; 2003; Tong et al.; 2008; Du and Schick; 2009).

To improve the literature, Tong and Wang (2005) and Tong et al. (2013) have taken another direction to improve the difference-based estimation. Specifically, they proposed a linear regression model that treats a class of first-order difference-based estimators as regressors and then estimated σ^2 as the intercept using least squares. Their work was inspired by the fact that the Rice estimator is always positively biased and so the proposed method was targeted to eliminate such bias. But in essence, their method also reduces the estimation variance dramatically and so achieves the asymptotically optimal rate (1.2). Note, however, that the linear regression method in Tong and Wang (2005) only used the first-order difference-based estimators as regressors. As a consequence, the practical performance of their estimator may not be satisfactory when n is small and g is rough, as reported in Table 2 of Tong and Wang (2005).

To make the linear regression method a more effective tool and also to tackle the sequence selection problem, we now propose a unified framework for estimating σ^2 that combines the linear regression method with higher-order difference sequences. For any order-r difference sequence $d = (d_0, \ldots, d_r)$ satisfying (2.2), we define

$$s_k(r) = \frac{1}{n-rk}\sum_{i=1}^{n-rk}\Big(\sum_{j=0}^{r} d_j Y_{i+jk}\Big)^2. \tag{2.6}$$

Note that $s_k(r)$ is a generalization of the classical difference-based estimators and it reduces to the estimators (2.3) when $k = 1$. Let $J(r) = (\sum_{j=0}^{r} j d_j)^2 \int_0^1 [g'(x)]^2 dx$. It is easy to verify that

$$E[s_k(r)] = \sigma^2 + \frac{k^2}{n^2} J(r) + o(\frac{k^2}{n^2}). \tag{2.7}$$

This shows that $s_k(r)$ are always positively biased for estimating σ^2.

To eliminate the bias term, we now apply a linear regression model to a collection of $s_k(r)$ and then estimate σ^2 as the intercept. Specifically, by letting $\alpha = \sigma^2$, $\beta = J(r)$ and $h_k = k^2/n^2$, we have the approximately linear regression model $s_k(r) \approx \alpha + h_k\beta$. Then for the given response values $s_k(r)$, $k = 1, \ldots, m$ with $m = o(n)$, we fit the regression model by minimizing the following weighted sum of squares

$$\sum_{k=1}^{m} w_k (s_k(r) - \alpha - h_k\beta)^2,$$

where $w_k = (n - rk)/N$ are the corresponding weights with $N = \sum_{k=1}^{m}(n - rk) = nm - rm(m + 1)/2$. This results in the estimator

$$\hat{\sigma}^2(r, m) = \hat{\alpha} = \sum_{k=1}^{m} b_k w_k s_k(r), \tag{2.8}$$

where $b_k = 1 - \bar{h}_w(h_k - \bar{h}_w)/(\sum_{k=1}^{m} w_k h_k^2 - \bar{h}_w^2)$ and $\bar{h}_w = \sum_{k=1}^{m} w_k h_k$.

We note that the weights w_k are assigned because each $s_k(r)$ involves $(n - rk)$ pairs of observations and the regression weighs equally for each pair. By doing this, we have a simplified form for the final estimator and have also taken in account the performance in finite sample size setting. Whereas for the asymptotic behavior, it can be shown that the estimator (2.8) is asymptotically equivalent to the estimator that minimizes the unweighted sum of squares $\sum_{k=1}^{m}(s_k(r) - \alpha - h_k\beta)^2$. In addition, we show in Section 2.6.1 that

Theorem 1. *For the equidistant design, we have (i) the proposed estimator $\hat{\sigma}^2(r, m)$ is an unbiased estimator for any difference sequence satisfying (2.2) when $g(x)$ is a linear function; and (ii) for the ordinary sequence, the estimator is an unbiased estimator for any polynomial $g(x)$ with order up to $p \leq r - 1$.*

2.2.2 Optimal Estimator

This section gives more exposure to the proposed estimator (2.8) and also defines the optimal estimator. As mentioned, by combining the linear regression method with higher-order difference sequences, we have proposed a general framework for estimating the residual variance in nonparametric regression. Specifically, by treating r and m as two tuning parameters, the estimator (2.8) provides a two-dimensional cone space for searching the optimal estimator. Let $\mathcal{S} = \{(r, m) : r = 1, 2, \dots; m = 1, 2, \dots\}$. We define the optimal estimator as $\hat{\sigma}^2_{\text{opt}} = \hat{\sigma}^2(r_{\text{opt}}, m_{opt})$ where

$$(r_{\text{opt}}, m_{opt}) = \underset{(r,m)\in\mathcal{S}}{\text{argmin}}\ E\left(\hat{\sigma}^2(r, m) - \sigma^2\right)^2. \tag{2.9}$$

Note that $(r_{\text{opt}}, m_{opt})$ are unknown in practice and need to be estimated.

It is obvious that the new framework includes most existing difference-based estimators as special cases. They are all located in the edge of the two-dimensional cone space. First, if we let $m = 1$ and $r = 1$, the new framework results in the first-order difference-based estimator in (2.4), which is located on the corner of the cone space. Now if we fix $m = 1$ and allow $r \geq 2$, the new framework results in the difference-based estimator in (2.3). Further, if we use the optimal sequence, it leads in the optimal estimator $\hat{\sigma}^2_{\text{opt}}(r)$; and if we use the ordinary sequence, it leads in the ordinary estimator $\hat{\sigma}^2_{\text{ord}}(r)$. On the other hand, if we fix $r = 1$ and allow $m \geq 2$, the new framework results in the least squares estimator in Tong and Wang (2005). Note also that the difference-based estimator in (2.3) finds the optimal tuning parameters $(r_{\text{opt}}, m_{opt})$ only in the subspace $\mathcal{S}_1 = \{(r, 1) : r = 1, 2, \dots\}$; whereas the least squares estimator finds them only in the subspace $\mathcal{S}_2 = \{(1, m) : m = 1, 2, \dots\}$. Neither of the existing methods is globally optimal in the new framework since $(r_{\text{opt}}, m_{opt})$ can also be located inside the cone space, i.e., in the space of $\mathcal{S} \setminus (\mathcal{S}_1 \cup \mathcal{S}_2)$.

In addition to generating a very large family of estimators, more importantly we will show that the new framework has great potential to tackle the challenging sequence selection problem. Specifically, in the next two sections we will show that a combination between the linear regression method and the ordinary sequence will be perfect, in which the linear regression method reduces mainly the estimation variance

and the ordinary sequence controls well in estimation bias. In contrary, a combination between the linear regression method and the optimal sequence is much less satisfactory. Then consequently, we may always recommend the use of the ordinary sequence in the proposed estimator (2.8), no matter if the sample size is small or if the signal-to-noise ratio is large.

It is also noteworthy that the proposed method can be interpreted from another aspect. In estimating the mean function at a given point, Cheng et al. (2007) and Paige et al. (2009) formed a linear combination of the local linear estimators evaluated at several nearby points as the final estimate. The linear combination therein was constructed in such a way that maximizes the variance reduction while remaining the asymptotic bias unchanged. To our knowledge, there is no existing work in the literature on variance reduction in nonparametric variance estimation. Note that $s_k(r)$ in (2.6) can be represented as a combination of several first-order estimators. We can, therefore, treat the proposed estimator (2.8) as a variance reduced estimator in comparison with the least squares estimator in Tong and Wang (2005).

2.3 Theoretical results

2.3.1 Asymptotic Variance and Bias

In what follows we study the statistical properties of the proposed estimator. We first derive the asymptotic mean of the estimator with various difference sequences. In Section 2.6.2 we show that

Theorem 2. *For the equidistant design, let r be fixed, $m \to \infty$ and $m/n \to 0$.*

(i) If the mean function $g(x)$ has a bounded second derivative, then for any order-r sequence we have

$$E(\hat{\sigma}^2(r,m)) = \sigma^2 + O(\frac{m^3}{n^3}).$$

(ii) If the ordinary sequence with fixed order $r \geq 2$ is used and the mean function

$g(x)$ has a bounded rth derivative, we have

$$E(\hat{\sigma}^2_{\text{ord}}(r,m)) = \sigma^2 + \binom{2r}{r}^{-2} \frac{3(1-r)m^{2r}}{(2r+1)(2r+3)n^{2r}} ||g^{(r)}||_2^2 + o(\frac{m^{2r}}{n^{2r}}),$$

where $||f||_2$ refers to the L_2-norm of the function f.

For the asymptotic variance, we present only the results for $r = 2$ for simplicity. Note that $r = 2$ is the minimum order that distinguishes the optimal sequence and the ordinary sequence. Using a higher order difference sequence will yield more tedious derivation and also more complex solutions; yet the comparison results remain the same. In addition, we note that an $r \geq 3$ order is rarely recommended in practice, even for $m = 1$ (Dette et al.; 1998).

For $r = 2$, the proposed estimator of σ^2 is given as $\hat{\sigma}^2(2,m) = \sum_{k=1}^{m} b_k w_k s_k(2)$. For ease of notation, let $Y = (Y_1, \ldots, Y_n)^T$, E be the set of positive even integers, I be the indicator function, and $\sum_{k=1}^{0} b_k = 0$. It is easy to verify that $\hat{\sigma}^2(2,m)$ can be expressed as a quadratic form. Specifically, we have

$$\hat{\sigma}^2(2,m) = Y^T DY / \text{tr}(D),$$

where $D = (d_{ij})_{n\times n}$ is an $n \times n$ matrix with diagonal elements

$$d_{ii} = d_0^2 \sum_{k=1}^{\min(m,[\frac{n-i}{2}])} b_k + d_1^2 \sum_{k=1}^{\min(m,n-i,i-1)} b_k + d_2^2 \sum_{k=1}^{\min(m,[\frac{i-1}{2}])} b_k, \qquad i = 1, \ldots, n,$$

and off-diagonal elements

$$d_{ij} = \begin{cases} d_0 d_1 b_k I_{(1\leq k\leq m)} + d_0 d_2 b_{k/2} I_{(k\in E)} & 1 \leq |i-j| = k \leq 2m, 1 \leq \min(i,j) \leq k, \\ -d_1^2 b_k I_{(1\leq k\leq m)} + d_0 d_2 b_{k/2} I_{(k\in E)} & 1 \leq |i-j| = k \leq 2m, k+1 \leq i,j \leq n-k, \\ d_1 d_2 b_k I_{(1\leq k\leq m)} + d_0 d_2 b_{k/2} I_{(k\in E)} & 1 \leq |i-j| = k \leq 2m, n-k+1 \leq \max(i,j) \leq n, \\ 0 & \text{otherwise.} \end{cases}$$

$[x]$ refers to the largest integer smaller than or equal to x.

To calculate the asymptotic variance, we use the following expression:

$$\begin{aligned} \text{var}(\hat{\sigma}^2) &= \frac{1}{\text{tr}(D)^2} \left[4\sigma^2 g^T D^2 g + 4 g^T (D \text{diag}(D) u) \sigma^3 \gamma_3 \right. \\ &\quad \left. + \sigma^4 \text{tr}\{\text{diag}(D)^2\}(\gamma_4 - 3) + 2\sigma^4 \text{tr}(D^2) \right], \end{aligned}$$

where $g = (g(x_1), \ldots, g(x_n))^T$, diag$(D)$ denotes the diagonal matrix of the diagonal element of D, $u = (1, \ldots, 1)^T$ and $\gamma_i = E(\varepsilon^i/\sigma^i)$ for $i = 3$ and 4. In Section 2.6.3 we show that

Theorem 3. *Assume that g has a bounded second derivative. For any $m \to \infty$ and $m = o(n)$, we can get the following results*

$$\text{var}(\hat{\sigma}^2(2, m)) = \frac{1}{n}\text{var}(\varepsilon^2) + \frac{A_1}{mn}\sigma^4 + \frac{A_2 m}{n^2}\text{var}(\varepsilon^2) + o(\frac{1}{nm}) + o(\frac{m}{n^2}), \qquad (2.10)$$

where $A_1 = [\frac{9}{4} + 9d_1^2(d_1^2 - \frac{1}{2})]$ and $A_2 = [\frac{9}{56} + \frac{165}{448}d_1^2(1 - d_1^2)]$.

The above results are for arbitrary difference sequences of order $r = 2$, which may be neither optimal nor ordinary. It is shown that the proposed estimator attains the asymptotically optimal rate (1.2), which is not achieved by the original difference-based estimators. With the optimal bandwidth, $m_{\text{opt}} = \sqrt{A_1\sigma^4/A_2\text{var}(\varepsilon^2)}n^{1/2}$, the optimal variance is of the form

$$\text{var}_{\text{opt}}(\hat{\sigma}^2(2, m)) = \text{var}(\varepsilon^2)n^{-1} + \sqrt{A_1 A_2 \sigma^2 \text{var}(\varepsilon^2)}n^{-3/2} + o(n^{-3/2}).$$

For optimal sequence, $d_{\text{opt}} = ((1 + \sqrt{5})/4, -1/2, (1 - \sqrt{5})/4)$, $A_1(d_{\text{opt}}) \approx 1.69$ and $A_2(d_{\text{opt}}) \approx 0.23$. For ordinary sequence, $d_{\text{ord}} = 6^{-1/2}(1, -2, 1)$, $A_1(d_{\text{ord}}) \approx 3.25$ and $A_2(d_{\text{ord}}) \approx 0.24$. Hence, for the same bandwidth m, the asymptotic variance of $\hat{\sigma}^2_{\text{ord}}(2, m)$ should be a little larger than that of $\hat{\sigma}^2_{\text{opt}}(2, m)$, but the difference between the two estimators gets smaller as the sample size increases.

Remark. Through a procedure similar with the proof of Theorem 3, we can show that $\hat{\sigma}^2(r, m)$ also achieves the optimal rate (1.2) for arbitrary order-r sequences.

2.3.2 Comparison

In this section, we make a comprehensive comparison of all the existing difference-based variance estimators, so as to demonstrate the effect of r and m to the estimators and eventually give an end to the controversial question about sequence selection.

In Table (2.1), we list the asymptotic relative variance (rVAR) and the order of squared bias (SQB) of various estimators included in our unified framework. For the

sake of simplicity, we concentrate on the normal error case. For non-normal error cases, the comparison results are quite similar. The rVAR are derived by multiplying the asymptotic variance with $n/(2\sigma^4)$.

Table 2.1: Asymptotic variance and squared bias.

	Original			Combined with Least Square		
Unified	$\hat{\sigma}^2(1,1)$	$\hat{\sigma}^2_{\text{opt}}(r,1)$	$\hat{\sigma}^2_{\text{ord}}(r,1)$	$\hat{\sigma}^2(1,m)$	$\hat{\sigma}^2_{\text{opt}}(r,m)$	$\hat{\sigma}^2_{\text{ord}}(r,m)$
Existing	$\hat{\sigma}^2_{\text{R}}$	$\hat{\sigma}^2_{\text{opt}}(r)$	$\hat{\sigma}^2_{\text{ord}}(r)$	$\hat{\sigma}^2_{\text{TW}}(m)$	-	-
rVAR	$3/2$	$1+1/2r$	$\sqrt{\pi r/2}$	1	1	1
SQB	$O(1/n^4)$	$O(1/n^4)$	$O(1/n^{4r})$	$O(m^6/n^6)$	$O(m^6/n^6)$	$O(m^{4r}/n^{4r})$

We begin our discussion with the three original estimators. First of all, none of them achieve the asymptotically optimal rate for the variance term, since the order r is always finite. Compare $\hat{\sigma}^2(1,1)$ with $\hat{\sigma}^2_{\text{opt}}(r,1)$ and $\hat{\sigma}^2_{\text{ord}}(r,1)$, we find that as r gets larger, rVAR decreases for the optimal estimators while increases for the ordinary estimators instead. At the same time, the order of SQB remains the same for the optimal estimators and goes down very quickly for the ordinary estimators. Dette et al. (1998) proposed an accurate approximation of mean square error under some regular conditions. Specifically, they approximate $\text{MSE}[\hat{\sigma}^2_{\text{opt}}(r,1)]$ as

$$C^2(r)||g'||_2^4/n^4 + (2+1/r)\sigma^4/n + 4C^2(r)\sigma^2||g''||_2^2/n^5, \tag{2.11}$$

where $C(r) = (2r+1)(r+1)/12$. Their approximation for $\text{MSE}[\hat{\sigma}^2_{\text{ord}}(r,1)]$ is

$$\binom{2r}{r}^{-2}\left[||g^{(r)}||_2^4/n^{4r} + 2\binom{4r}{2r}\sigma^4/n + 4\sigma^2||g^{(2r)}||_2^2/n^{4r+1}\right]. \tag{2.12}$$

The first part is the square bias term and the other two parts are variance term. Note that $C(r)$ in (2.11) is an increasing function of the sequence order r, hence SQB of the optimal estimators actually increases quickly as r raises though its order remains

the same. A high order r will result in large SQB for the ordinary estimators while generate large variance for the optimal estimators. Hence in practice, estimators with order $r > 3$ are rarely recommended.

Then, we focus on the estimators generated through least square method. First of all, they all achieve the asymptotic optimal rate for the variance. Then for SQB, the effect of r is similar with that for the original estimators. That is say as r grows, SQB increases for $\hat{\sigma}^2_{\text{opt}}(r, m)$ while decreases for $\hat{\sigma}^2_{\text{ord}}(r, m)$. We also observed such patterns in our numerical studies.

Finally, we compare the original estimators with their corresponding derivatives, so as to access the effect of least square method or equivalently the bandwidth m. Obviously, the asymptotical variance of all the original estimators are improved to be optimal no matter which kind of sequence is applied. Through least square regression, the order of SQB for $\hat{\sigma}^2_{\text{ord}}(r, m)$ is higher than that of $\hat{\sigma}^2_{\text{ord}}(r, m)$ for any $m \to \infty$ and $m = o(n)$. And it also gets higher for $\hat{\sigma}^2(1, m)$ and $\hat{\sigma}^2_{\text{opt}}(r, m)$ if $m = n^t$ and $t > 1/3$. From this aspect, we may say that the main benefit brought by least square method is optimizing the asymptotic variance, although it was first applied to reduce the positive bias of $\hat{\sigma}^2_{\text{R}}$ in Tong and Wang (2005).

Now we make some suggestions for practical implementation based on the theoretical results in Table 2.1. Asymptotically, the mean squared error are equivalent with the variance, and the least square estimators are obviously better choices. With a proper bandwidth m, each of them will generate optimal mean squared error. For finite sample sizes, the mean squared error may be greatly effected by the squared bias as shown in (2.11). An estimator with smaller squared bias will be more stable and hence more preferred. Among all the estimators, $\hat{\sigma}^2_{\text{ord}}(r, 1)$ and $\hat{\sigma}^2_{\text{ord}}(r, m)$ maintain better control for the squared bias. Combine the two results together, we get the conclusion that $\hat{\sigma}^2_{\text{ord}}(r, m)$ is the only choice appropriate for all the situations. So, we recommend $\hat{\sigma}^2_{\text{ord}}(r, m)$ for practical implementation.

2.4 Simulation studies

In this section, we provide a criterion for tuning parameters selection and then conduct a simulation study to access the finite sample performance of proposed estimator under various settings.

Before that, we give a more direct illustration for the behavior of $\hat{\sigma}^2(r, m)$ with respect to the change of r or m. We tried plenty of settings and got various plots illustrating the change of relative mean squared error (RMSE) of proposed estimator against the two parameters. Two examples are presented in Figure 2.1.

First, we can see that the RMSE vary significantly against the change of two parameters, that is say the choice of (r, m) is critical for $\hat{\sigma}^2(r, m)$. Also, the minimum value of RMSE are not always on the marginal area in these cases, that is say the optimal tuning parameters $(r_{\text{opt}}, m_{\text{opt}})$ may be located in the space of $\mathcal{S} \setminus (\mathcal{S}_1 \cup \mathcal{S}_2)$.

Then, we look at the marginal area. The lines $m = 1$ show us the RMSE of the original estimators, $\hat{\sigma}^2_{\text{opt}}(r, 1)$ and $\hat{\sigma}^2_{\text{ord}}(r, 1)$, against the order of difference r, respectively. We can see that the RMSE of $\hat{\sigma}^2_{\text{opt}}$ decreases as r increases, whereas RMSE of $\hat{\sigma}^2_{\text{ord}}$ increases instead, which are consistent with what suggested in Hall et al. (1990). The line $r = 1$ is corresponding with the RMSE for $\hat{\sigma}^2_{\text{TW}}$ in Tong and Wang (2005) and Tong et al. (2013).

2.4.1 Selection of Tuning Parameters

A good choice of (r, m) is important for the performance of the estimate of σ^2. An effective tuning parameter selection procedure is required. For the order of difference, we consider $r = 1,\ 2$ and 3 which is popularly used in literature (Gasser et al.; 1986; Eagleson; 1989; Dette et al.; 1998; Tong and Wang; 2005). For the choice of bandwidth m, Tong and Wang (2005) suggested $m_t = n^{1/3}$ and $m_s = n^{1/2}$ as two options. Besides, they also provided an adaptive bandwidth selection procedure based on a cross-validation method.

As the CV method, they divided the whole data set into V disjoint subsamples,

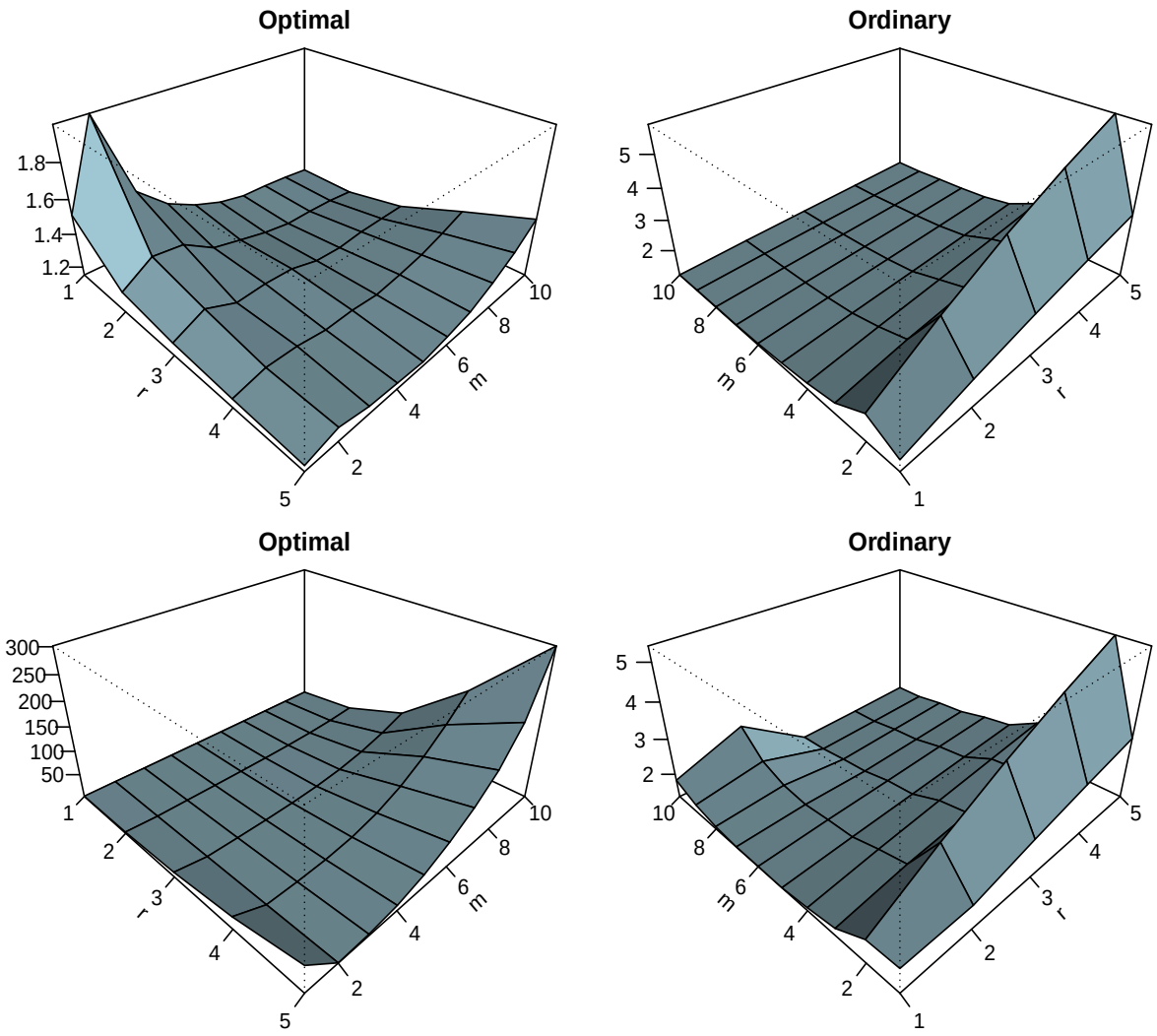

Figure 2.1: $n = 100$. The left two plots: $\mathrm{RMSE}[\hat{\sigma}^2_{\mathrm{opt}}(r,m)]$; the right ones $\mathrm{RMSE}[\hat{\sigma}^2_{\mathrm{ord}}(r,m)]$. Upper: $g(x) = 5\sin(\pi x)$ and $\sigma^2 = 4$; Lower: $g(x) = 5\sin(2\pi x)$ and $\sigma^2 = 4$. The RMSE are based on 1000 Monte Carlo runs and calculated with $\mathrm{MSE}(\hat{\sigma}^2)n/(2\sigma^4)$.

$S_1, \ldots, S_V$ and then select the $m = m_{\mathrm{CV}}$ minimizing

$$\mathrm{CV}(m) = \sum_{v=1}^{V} [\hat{\sigma}^2(m) - \hat{\sigma}^2_v(m)]^2,$$

where $\hat{\sigma}^2_v$ denotes the estimated variance on the whole sample except for S_v.

The CV method is computational expensive when selecting r and m simultaneously, especially for large sample sizes. Hence, we suggest another effective method

named *plateau method* proposed in Müller and Stadtmüller (1999) p. 318 for moderate and large sample size cases. Here, we choose the following criterion, which approximates the local variation of the estimator,

$$(\hat{r}, \hat{m}) = \arg\min_{r,m} \left\{ \frac{1}{2m_r+1} \sum_{i=[m/r]-m_r}^{[m/r]+m_r} [\hat{\sigma}^2(r,i)]^2 - \left[\frac{1}{2m_r+1} \sum_{i=[m/r]-m_r}^{[m/r]+m_r} \hat{\sigma}^2(r,i) \right]^2 \right\},$$

where $m_r = [m_0/r]$ and $m_0 = \max([n/50], 2)$, so that the largest span keeps the same for all sequences with different orders.

From Figure 2.2, we can find that $\hat{\sigma}^2(r,m)$ stays around the true value of the residual variance within some range of bandwidth m, and then moves away monotonically. That is say within this area, the estimator is relatively stable. The criterion 2.4.1 selects the pair of parameters corresponding with the smallest approximated mean squared error.

2.4.2 Simulation Study

In this section, we conduct a simulation study to investigate the performance of proposed estimator and also make a numerical comparison with some existing estimators.

We apply the following settings. Four mean functions are chosen as $g(x) = 5x$ and $g(x) = 5sin(w\pi x)$, $w = 1,\ 2$ and 4, with different oscillation levels. Sample size, $n = 25$, 100 and 500, corresponding with small, moderate and large sample sizes respectively. The random errors ε are generated from $N(0, \sigma^2)$ with $\sigma = 0.2, 0.5$ and 2. We calculate six estimators for each setting, e.i., $\hat{\sigma}^2_{\mathrm{R}}$, $\hat{\sigma}^2_{\mathrm{GSJ}}$, $\hat{\sigma}^2_{\mathrm{HKT}}$, $\hat{\sigma}^2_{\mathrm{TW}}$, $\hat{\sigma}^2_{\mathrm{opt}}(r,m)$ and $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$.

The cross-validation criterion is used to determine the tuning parameters of $\hat{\sigma}^2_{\mathrm{TW}}$ for all sample sizes and also for $\hat{\sigma}^2_{\mathrm{opt}}(r,m)$ and $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ when $n = 25$. When $n = 100$ and 500, the plateau method is applied for $\hat{\sigma}^2_{\mathrm{opt}}(r,m)$ and $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ instead. The range of parameters are set as follows. For $\hat{\sigma}^2_{\mathrm{TW}}$, we choose m from $(1,\ldots,5)$ as $n = 25$, $(1,\ldots,15)$ as $n = 100$ and $(1,\ldots,30)$ as $n = 500$. For $\hat{\sigma}^2_{\mathrm{opt}}(r,m)$ and $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$, we choose (r,m) from $(1,1),\ldots,(1,5)$, $(2,1)$, $(2,2)$ and $(3,1)$ as $n = 25$; $(1,1),\ldots,(1,15)$, $(2,1),\ldots,(2,7)$ and $(3,1)\ldots(3,5)$ as $n = 100$; $(1,1),\ldots,(1,30)$, $(2,1),\ldots,(2,15)$ and $(3,1)\ldots(3,10)$ as $n = 500$.

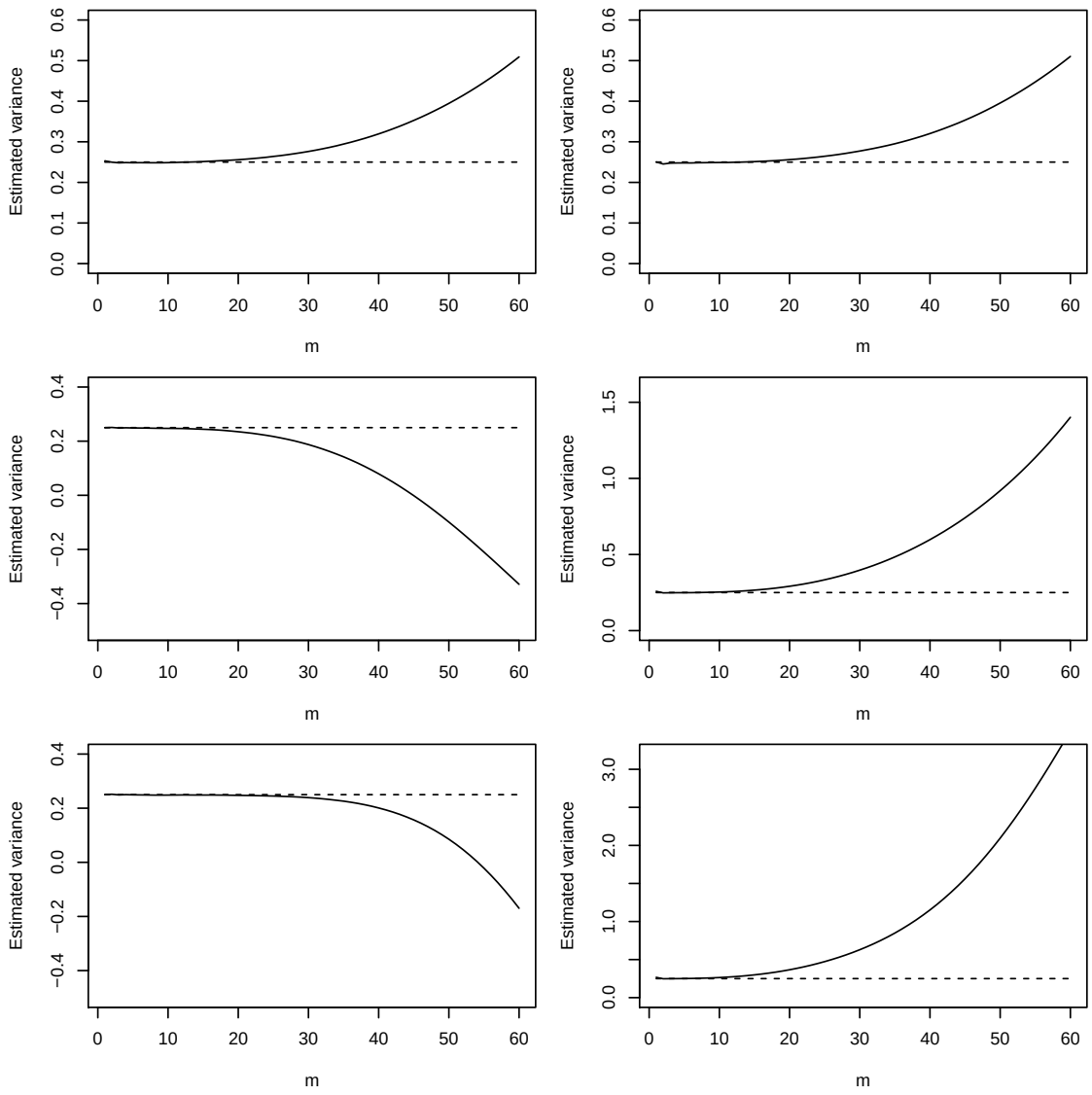

Figure 2.2: The change of $\hat{\sigma}^2$ along with the increase of bandwidth m. The value of $\hat{\sigma}^2$ (solid lines) are based on average of 100 times simulation results. The left: ordinary sequence. The right: optimal sequence. $n = 500$, $\sigma^2 = 0.25$ (dashed lines) and $g(x) = 5sin(4\pi x)$. From top to bottom: $r = 1$, 2 and 3.

With selected turning parameters, we get estimates denoted with $\hat{\sigma}^2_{\mathrm{TW}}(\mathrm{CV})$, $\hat{\sigma}^2_{\mathrm{opt}}(\mathrm{CV})$ and $\hat{\sigma}^2_{\mathrm{ord}}(\mathrm{CV})$. We repeat this procedure for 1000 times and compute the mean squared error (MSE) for each estimator. For the convenience of comparison, we scale the MSE by multiplying $n/(2\sigma^4)$, and report the RMSE in in Tables 2.2, 2.3 and 2.4.

Among all the estimators, $\hat{\sigma}^2_{\mathrm{GSJ}}$ and $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ are relatively robust across various situations, $\hat{\sigma}^2_{\mathrm{R}}$ and $\hat{\sigma}^2_{\mathrm{TW}}$ are less stable, $\hat{\sigma}^2_{\mathrm{HKT}}$ and $\hat{\sigma}^2_{\mathrm{opt}}(r,m)$ are extremely sensitive to simulation settings. Estimators based on the same type of sequences share the same properties for the robustness. Then, we compare the original estimators with the least square estimators under the same setting and find that the original estimators are all improved by the least square technique for most of settings.

Table 2.2: Relative mean squared errors for the six investigated estimators. $n = 500$.

n	σ	$g(x)$	$\hat{\sigma}^2_{\mathrm{R}}$	$\hat{\sigma}^2_{\mathrm{GSJ}}$	$\hat{\sigma}^2_{\mathrm{HKT}}$	$\hat{\sigma}^2_{\mathrm{TW}}$(CV)	$\hat{\sigma}^2_{\mathrm{opt}}$(CV)	$\hat{\sigma}^2_{\mathrm{ord}}$(CV)
500	0.2	Linear	1.47	1.91	1.20	1.20	1.04	1.03
		$w = 1$	1.47	1.91	1.25	1.23	1.11	1.09
		$w = 2$	1.61	1.91	2.13	1.26	1.20	1.15
		$w = 4$	3.88	1.91	16.2	1.48	2.12	1.17
	0.5	Linear	1.47	1.91	1.20	1.20	1.03	1.02
		$w = 1$	1.47	1.91	1.20	1.20	1.06	1.05
		$w = 2$	1.47	1.91	1.22	1.22	1.10	1.10
		$w = 4$	1.53	1.91	1.58	1.29	1.18	1.17
	2	Linear	1.47	1.91	1.20	1.21	1.02	1.02
		$w = 1$	1.47	1.91	1.20	1.20	1.02	1.02
		$w = 2$	1.47	1.91	1.20	1.19	1.03	1.04
		$w = 4$	1.47	1.91	1.20	1.19	1.06	1.08

For large sample sizes, we observe the following results for most of cases, $\mathrm{RMSE}[\hat{\sigma}^2_{\mathrm{opt}}(r,m)] \simeq \mathrm{RMSE}[\hat{\sigma}^2_{\mathrm{ord}}(r,m)] < \mathrm{RMSE}[\hat{\sigma}^2_{\mathrm{TW}}] \simeq \mathrm{RMSE}[\hat{\sigma}^2_{\mathrm{HKT}}] < \mathrm{RMSE}[\hat{\sigma}^2_{\mathrm{R}}] < \hat{\mathrm{RMSE}}[\sigma^2_{\mathrm{GSJ}}]$. When sample size n is moderate, the above relationship holds for smooth mean functions and large σ. Under opposite situations, $\hat{\sigma}^2_{\mathrm{R}}$ and $\hat{\sigma}^2_{\mathrm{HKT}}$ get quite unstable and even fail to give reasonable results for some settings, e.g., $(n = 100;\ \sigma = 0.2;\ w = 2, 4)$. While $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ still retains its excellent performance.

Table 2.3: Relative mean squared errors for the six investigated estimators. $n = 100$.

n	σ	$g(x)$	$\hat{\sigma}^2_{\mathrm{R}}$	$\hat{\sigma}^2_{\mathrm{GSJ}}$	$\hat{\sigma}^2_{\mathrm{HKT}}$	$\hat{\sigma}^2_{\mathrm{TW}}(\mathrm{CV})$	$\hat{\sigma}^2_{\mathrm{opt}}(\mathrm{CV})$	$\hat{\sigma}^2_{\mathrm{ord}}(\mathrm{CV})$
100	0.2	Linear	1.48	1.90	1.46	1.27	1.16	1.15
		$w = 1$	2.48	1.90	8.02	1.44	1.41	1.32
		$w = 2$	19.5	1.90	113	4.70	1.74	1.57
		$w = 4$	297	1.90	1803	2.87	2.12	2.01
	0.5	Linear	1.46	1.90	1.23	1.28	1.14	1.13
		$w = 1$	1.47	1.90	1.36	1.30	1.27	1.28
		$w = 2$	1.85	1.90	3.88	1.36	1.36	1.35
		$w = 4$	8.75	1.90	46.5	2.65	1.74	1.70
	2	Linear	1.46	1.90	1.23	1.28	1.14	1.14
		$w = 1$	1.46	1.90	1.23	1.29	1.16	1.14
		$w = 2$	1.46	1.90	1.23	1.27	1.14	1.17
		$w = 4$	1.47	1.90	1.35	1.34	1.27	1.33

As sample size is small, all the estimators get more sensitive to the oscillation level of the mean function and the value of the residual variance. Only $\hat{\sigma}^2_{\mathrm{ord}}(r, m)$ provides reasonable mean squared error for all the settings. Though RMSE$[\hat{\sigma}^2_{\mathrm{ord}}]$ is a little larger when the mean function is not oscillating and $\sigma = 2$, but still $\hat{\sigma}^2_{\mathrm{ord}}$ is more preferable for its reliability.

In conclusion, we recommend $\hat{\sigma}^2_{\mathrm{ord}}$ for practical implementation, no matter if the sample size is small or if the signal-to-noise ratio is large.

2.4.3 Robustness of the method

In practical application, some assumptions about Model (2.1) may be violated. In this section, we consider two kinds of variation, *non-equidistant design* and *non-smooth mean function*. For each case, we conduct simulation studies to access the robustness of both existing and newly proposed methods.

Table 2.4: Relative mean squared errors for the six investigated estimators. $n = 25$.

n	σ	$g(x)$	$\hat{\sigma}^2_{\mathrm{R}}$	$\hat{\sigma}^2_{\mathrm{GSJ}}$	$\hat{\sigma}^2_{\mathrm{HKT}}$	$\hat{\sigma}^2_{\mathrm{TW}}(\mathrm{CV})$	$\hat{\sigma}^2_{\mathrm{opt}}(\mathrm{CV})$	$\hat{\sigma}^2_{\mathrm{ord}}(\mathrm{CV})$
25	0.2	Linear	4.32	1.85	20.3	2.78	2.37	1.76
		$w = 1$	70.8	1.86	397	31.3	12.8	2.08
		$w = 2$	1108	2.40	6150	15.4	216	2.20
		$w = 4$	17295	145	90828	261	9396	2.66
	0.5	Linear	1.48	1.90	1.46	1.19	1.36	1.56
		$w = 1$	3.10	1.85	11.5	2.08	1.82	2.01
		$w = 2$	29.4	1.84	159	23.3	6.22	2.06
		$w = 4$	442	5.22	2308	172	241	2.19
	2	Linear	1.46	1.90	1.23	1.26	1.34	1.51
		$w = 1$	1.43	1.85	1.36	1.24	1.32	1.53
		$w = 2$	1.51	1.85	1.91	1.24	1.38	1.61
		$w = 4$	3.10	1.84	10.4	2.59	2.42	2.52

Non-equidistant design

We consider two kinds of non-equidistant design

(1) α percent design points are generated from $\mathrm{Unif}(0, 1)$ and the other $1-\alpha$ percent are generated from $\mathrm{Unif}(0.4, 0.6)$, including one cluster at the interval $(0.4, 0.6)$, $\alpha = 0.2, 0.4, 0.6$ and 0.8.

(2) All the design points are generated from $\mathrm{Beta}(2, 2)$.

Two mean functions are employed, $g_1(x) = 5x$ and $g_2(x) = 5\sin(2\pi x)$. Random errors are generated from $N(0, 0.25)$. Sample size is chosen as $n = 100$. The bandwidth choices are the same with that from Section 4.2. We report RMSE of the estimators in Table 2.5.

When the mean function is not oscillating, $g_1(x) = 5x$, the first four estimators are relatively robust to the violation of equidistant design, while $\hat{\sigma}^2_{\mathrm{opt}}(\mathrm{CV})$ and $\hat{\sigma}^2_{\mathrm{ord}}(\mathrm{CV})$ fail to keep robustness when $\alpha = 0.2$, in which case a big cluster of data appears.

As the mean function gets oscillating, $g_2(x) = 5\sin(2\pi x)$, all the estimators get more sensitive to the distribution of design points, among which $\hat{\sigma}^2_{\mathrm{GSJ}}$ is the most stable one.

Table 2.5: Robustness of the six investigated estimators to non-equidistant design points.

$g(x)$	Design	$\hat{\sigma}^2_{\mathrm{R}}$	$\hat{\sigma}^2_{\mathrm{GSJ}}$	$\hat{\sigma}^2_{\mathrm{HKT}}$	$\hat{\sigma}^2_{\mathrm{TW}}$(CV)	$\hat{\sigma}^2_{\mathrm{opt}}$(CV)	$\hat{\sigma}^2_{\mathrm{ord}}$(CV)
	equidistant	1.46	1.90	1.23	1.28	1.14	1.13
	$\alpha = 0.8$	1.42	1.83	1.29	1.29	1.17	1.22
	$\alpha = 0.6$	1.44	1.84	1.33	1.31	1.28	1.27
$g_1(x)$	$\alpha = 0.4$	1.42	1.82	1.33	1.36	1.45	1.41
	$\alpha = 0.2$	1.61	2.04	1.47	1.50	2.69	2.48
	Beta	1.37	1.79	1.18	1.27	1.13	1.18
	equidistant	1.85	1.90	3.87	1.36	1.36	1.35
	$\alpha = 0.8$	3.12	2.16	7.46	2.20	3.48	1.89
	$\alpha = 0.6$	3.62	2.25	8.92	2.41	4.96	1.96
$g_2(x)$	$\alpha = 0.4$	4.77	2.38	12.2	2.81	42.8	5.95
	$\alpha = 0.2$	8.18	3.19	16.8	5.43	48.0	45.6
	Beta	2.95	2.08	5.90	2.22	8.13	3.04

Non-smooth mean function

We consider two kinds of mean functions with one jump point at $x = 0.5$: $g_1(x) = 5x + cI(x > 0.5)$ and $g_2(x) = 5\sin(2\pi x) + cI(x > 0.5)$, where $c = 0.5, 1, 2$ and 4. The design points are equidistant. Other settings are the same with that for *non-equidistant* cases. RMSE of the estimators are listed in Table 2.6.

From Table 2.6, we can see that all the estimators are not robust to violation of the smooth mean function assumption. As the jump magnitude increases, RMSE of these estimators keep raising. From this point of view, an effective difference-based method should be further developed for models involving such kind of mean functions.

Table 2.6: Robustness of the six investigated estimators to non-smooth mean functions.

$g(x)$	c	$\hat{\sigma}^2_{\mathrm{R}}$	$\hat{\sigma}^2_{\mathrm{GSJ}}$	$\hat{\sigma}^2_{\mathrm{HKT}}$	$\hat{\sigma}^2_{\mathrm{TW}}(\mathrm{CV})$	$\hat{\sigma}^2_{\mathrm{opt}}(\mathrm{CV})$	$\hat{\sigma}^2_{\mathrm{ord}}(\mathrm{CV})$
	0	1.46	1.90	1.23	1.28	1.14	1.13
	0.5	1.47	1.92	1.25	1.29	1.14	1.17
$g_1(x)$	1	1.53	1.96	1.36	1.31	1.28	1.30
	2	1.97	2.21	2.35	1.58	3.17	2.64
	4	7.50	4.89	15.1	4.35	23.3	9.02
	0	1.85	1.90	3.88	1.36	1.36	1.35
	0.5	1.86	1.92	3.70	1.38	1.42	1.35
$g_2(x)$	1	1.98	1.96	3.90	1.50	1.55	1.42
	2	2.69	2.22	5.59	2.75	2.46	1.76
	4	8.67	4.89	19.5	4.45	11.8	6.25

2.5 Real applications

We apply the recommended estimator to two real data sets for illustration.

The first data are daily measurements of the temperature from the weather Stations in Inuvik, N.W.T., Canada, which are analyzed in Ramsay and Silverman (1997). We get the data from MD*base. The data are plotted in Figure (2.3). There are totally $n = 365$ pairs of observations in the data set. We apply plateau method to choose the bandwidth m and order of sequence r. m ranges from 1 to 50 and $r = 1,\ 2,\ 3$. The selected tuning parameters are $(r, m) = (3, 10)$ and $\hat{\sigma}^2_{\mathrm{ord}}(3, 10) = 0.43$.

The second data are Sea Level Pressure record of September 26-27, 2010, from Tsing Yi(North) station, Hong Kong. The data are recorded at a 10-minutes interval and totally 288 pairs of data are observed. The data are plotted in Figure (2.5). For this data, we choose bandwidth from $1, \ldots, 30$ and order $r = 1,\ 2,\ 3$. The selected

tuning parameters are $(r, m) = (3, 5)$ and $\hat{\sigma}^2_{\text{ord}}(3, 5) = 0.011$.

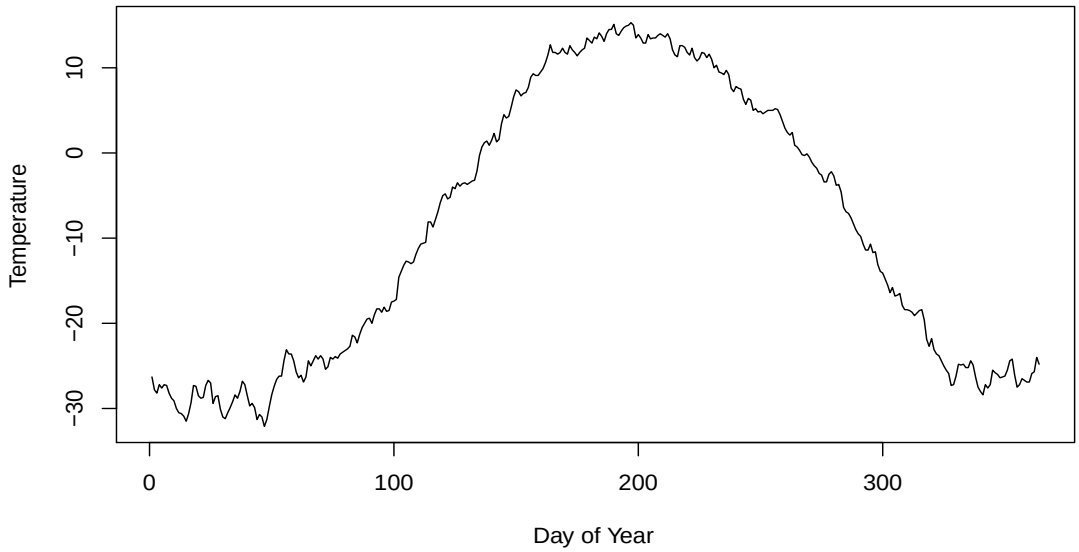

Figure 2.3: Daily Temperature

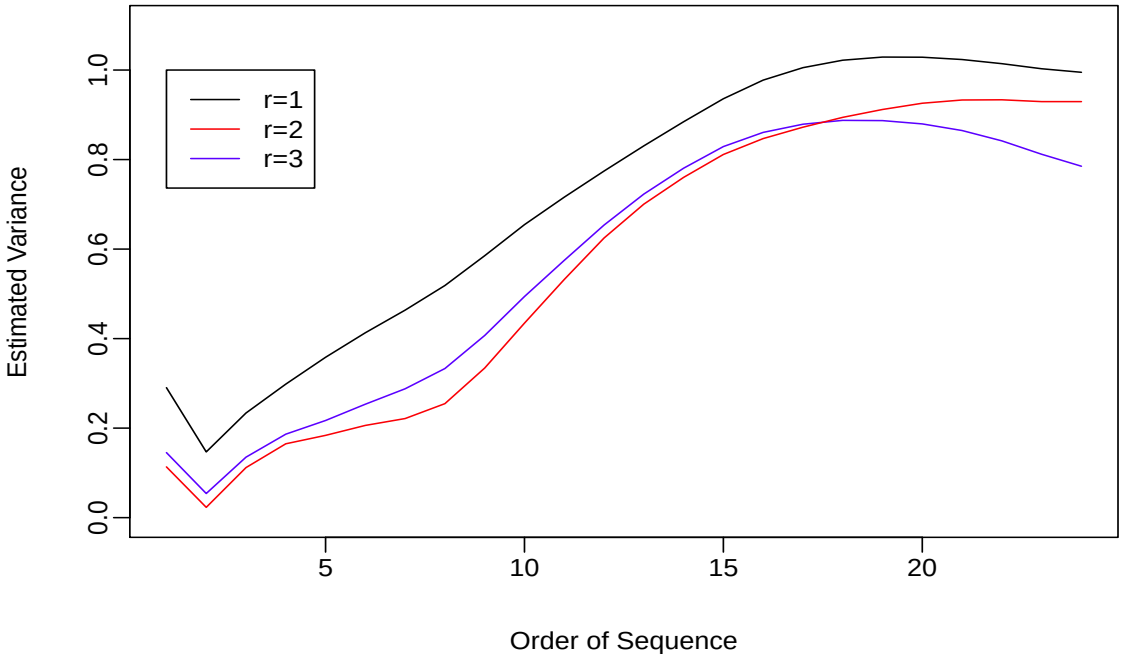

Figure 2.4: Estimated variance for Daily Temperature Data.

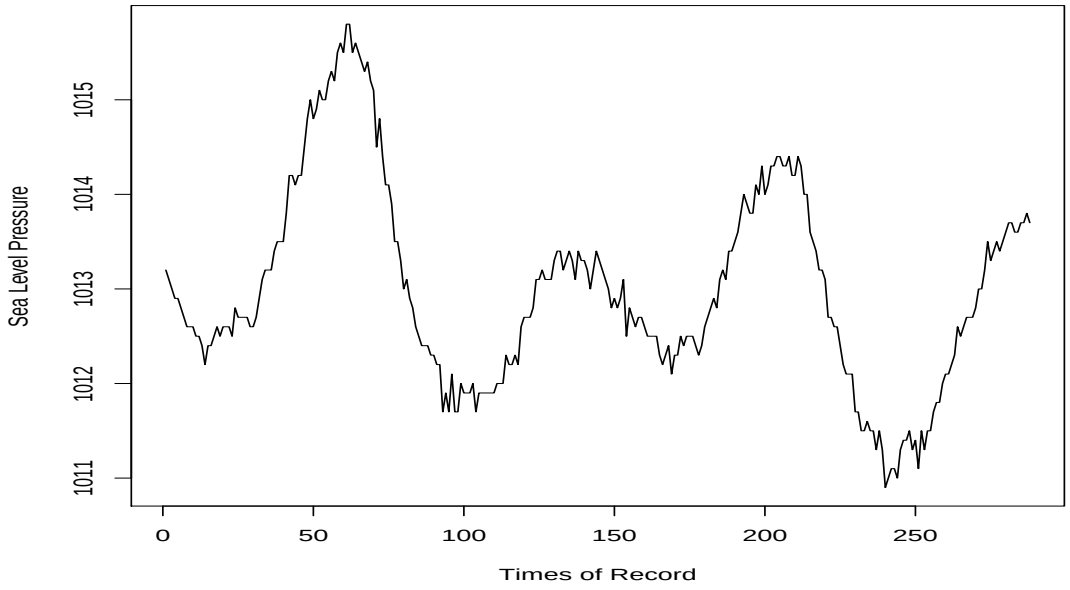

Figure 2.5: Sea Level Pressure.

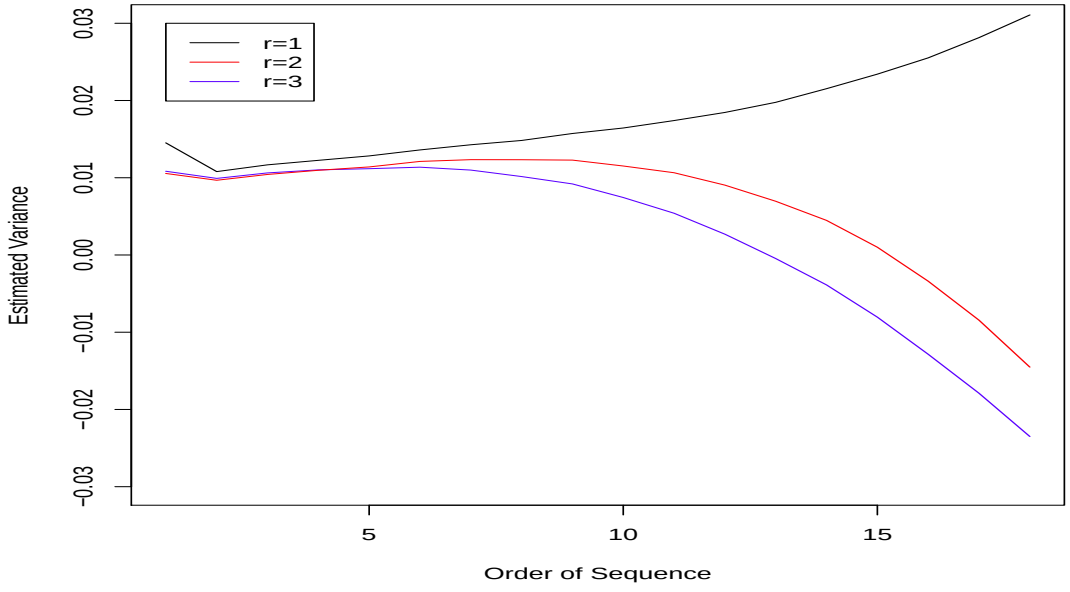

Figure 2.6: Estimated variance Sea Level Pressure

The Sea Level Pressure data appear more oscillating than Daily Temperature data and our parameter selection criterion suggests a smaller bandwidth for the more oscillating one, which is corresponding with our conclusions in Section 2.3 and 2.4. Also from Figure (2.4) and (2.6), we find, for both sets of data, the selected tuning parameters locate close to the inflexion point, where the estimated variance increases or decreases rapidly afterwards.

2.6 Proofs

This section provides technical proofs for Theorem 1-3.

2.6.1 Proof of Theorem 1

It is easy to verify the following results,

$$\sum_{k=1}^{m} b_k w_k = 1 \text{ and } \sum_{k=1}^{m} b_k w_k h_k = 0. \tag{2.13}$$

(i) Assume that $g(x) = ax + b$. We have

$$\begin{aligned} E(s_k(r)) &= \frac{1}{n-kr} E\left[\sum_{i=1}^{n-kr}(\sum_{j=0}^{r} d_j Y_{i+jk})^2\right] \\ &= \sigma^2 + \frac{1}{n-kr}\sum_{i=1}^{n-kr}\left(\sum_{j=0}^{r} d_j (a\frac{(i+jk)}{n} + b)\right)^2 \\ &= \sigma^2 + \frac{1}{n-kr}\sum_{i=1}^{n-kr}\left(ak\sum_{j=0}^{r} d_j j/n\right)^2 \\ &= \sigma^2 + a^2(\sum_{j=0}^{r} d_j j)^2 h_k. \end{aligned}$$

This leads to

$$\begin{aligned} E(\hat{\sigma}^2(r,m)) &= \sum_{k=1}^{m} b_k w_k \left[\sigma^2 + a^2(\sum_{j=0}^{r} d_j j)^2 h_k\right] \\ &= \sigma^2 + a^2(\sum_{j=0}^{r} d_j j)^2 \sum_{k=1}^{m} b_k w_k h_k \\ &= \sigma^2. \end{aligned}$$

(ii) Assume that $g(x)$ is an order-p polynomial with $p \leq r-1$. Let $g_i^{(p)}$ denote the p-th order derivative at the design point x_i, then we have

$$\begin{aligned} E(s_k(r)_{\text{ord}}) &= \sigma^2 + \frac{1}{n-kr}\sum_{i=1}^{n-kr}\left(\sum_{j=0}^{r} d_j g_{i+jk}\right)^2 \\ &= \sigma^2 + \frac{1}{n-kr}\sum_{i=1}^{n-kr}\left(d_0 g_i + d_1 \sum_{j=0}^{p} g_i^{(j)}\frac{(k/n)^j}{j!} + \cdots + d_r \sum_{j=0}^{p} g_i^{(j)}\frac{(jk/n)^j}{j!}\right)^2 \\ &= \sigma^2 + \frac{1}{n-kr}\sum_{i=1}^{n-kr}\left(C_0 g_i + C_1 g_i' \frac{k}{n} + \cdots + C_p g_i^{(p)}\frac{k^p}{n^p}\right)^2, \end{aligned}$$

where we denote $C_i(d)$ with C_i for simplicity of presentation and $C_0 = \sum_{j=0}^{r} d_j$ and $C_i = \sum_{j=0}^{r} j^i d_j / i!$ for $i = 1, \ldots, r$. When $(d_0, \ldots, d_r)$ is chosen as the order-r ordinary sequence, we know that $C_i = 0$ for $0 \leq i \leq p \leq r-1$. Hence, $E(s_k(r)_{\text{ord}}) = \sigma^2$ and so $E(\hat{\sigma}^2(r,m)_{\text{ord}}) = \sigma^2$.

2.6.2 Proof of Theorem 2

For $s_k(r)$, we have

$$\begin{aligned} E(s_k(r)) &= \frac{1}{n-kr} E\left[\sum_{i=1}^{n-kr}(\sum_{j=0}^{r} d_j Y_{i+jk})^2\right] \\ &= \sigma^2 + \frac{1}{n-kr}\sum_{i=1}^{n-kr}\left(C_1 g_i' k/n + O(k^2/n^2)\right)^2 \\ &= \sigma^2 + C_1^2 h_k \int_0^1 [g'(x)]^2 dx + O(\frac{k^3}{n^3}). \end{aligned}$$

This leads to

$$\begin{aligned} E(\hat{\sigma}^2(r,m)) &= \sum_{k=1}^{m} b_k w_k \left[\sigma^2 + C_1^2 h_k \int_0^1 [g'(x)]^2 dx + O(\frac{k^3}{n^3})\right] \\ &= \sigma^2 + C_1^2 \int_0^1 [g'(x)]^2 \sum_{k=1}^{m} b_k w_k h_k + O(\frac{m^3}{n^3}) \\ &= \sigma^2 + O(\frac{m^3}{n^3}) \end{aligned}$$

For the order-r ordinary sequence, we have $C_1 = C_2 = \cdots = C_{r-1} = 0$. Subject to

Dette et al. (1998), when order-r ordinary sequence is employed, $C_r = \binom{2r}{r}^{-1}$. Hence,

$$\begin{aligned} E(s_k(r)) &= \frac{1}{n-kr} E\left[\sum_{i=1}^{n-kr} (\sum_{j=0}^{r} d_j Y_{i+jk})^2\right] \\ &= \sigma^2 + \frac{1}{n-kr}\sum_{i=1}^{n-kr}\left(C_1 g_i' k/n + \cdots + C_r g_i^{(r)} k^r/n^r + o(k^r/n^r)\right)^2 \\ &= \sigma^2 + C_r^2 \frac{k^{2r}}{n^{2r}} \int_0^1 [g^{(r)}(x)]^2 dx + o(\frac{k^{2r}}{n^{2r}}). \end{aligned}$$

Further, we have

$$\begin{aligned} E(\hat{\sigma}^2(r,m)) &= \sum_{k=1}^{m} b_k w_k \left[\sigma^2 + C_r^2 \frac{k^{2r}}{n^{2r}} \int_0^1 [g^{(r)}(x)]^2 dx + o(\frac{k^{2r}}{n^{2r}})\right] \\ &= \sigma^2 + C_r^2 \int_0^1 [g^{(r)}(x)]^2 dx \sum_{k=1}^{m} b_k w_k \frac{k^{2r}}{n^{2r}} + o(\frac{m^{2r}}{n^{2r}}) \\ &= \sigma^2 + \frac{3C_r^2(1-r)m^{2r}}{(2r+1)(2r+3)n^{2r}} \int_0^1 [g^{(r)}(x)]^2 dx + o(\frac{m^{2r}}{n^{2r}}), \end{aligned}$$

where we use the fact that

$$\begin{aligned} \sum_{k=1}^{m} b_k w_k \frac{k^{2r}}{n^{2r}} &= \sum_{k=1}^{m} w_k \frac{k^{2r}}{n^{2r}} - \frac{\bar{h}_w}{\sum_{k=1}^{m} w_k h_k^2 - \bar{h}_w^2} \sum_{k=1}^{m} (h_k - \bar{h}_w) \frac{k^{2r}}{n^{2r}} \\ &= \frac{1}{(2r+1)} \frac{m^{2r}}{n^{2r}} (1 + o(1)) - \frac{15}{4}\left[\frac{1}{(2r+3)} - \frac{1}{3(2r+1)}\right] \frac{m^{2r}}{n^{2r}} (1 + o(1)) \\ &= \frac{3(1-r)}{(2r+1)(2r+3)} \frac{m^{2r}}{n^{2r}} (1 + o(1)). \end{aligned}$$

2.6.3 Proof of Theorem 3

We first introduce three lemmas. Lemma 1 is derived with some simple algebra. Lemma 2 is derived by the results in Lemma 1. The proof of Lemma 3 is provided.

Lemma 1. *Assume that $m \to \infty$ and $m/n \to 0$. We have*

(a) $\sum_{k=1}^{m} h_k = \frac{m^3}{3n^2} + \frac{m^2}{2n^2} + o(\frac{m^2}{n^2})$.

(b) $\bar{h}_w = \frac{1}{Nn^2}[\frac{1}{3}nm^3 + \frac{1}{2}nm^2 - \frac{1}{2}m^4 + o(nm^2) + o(m^4)]$.

(c) $\frac{1}{N} = \frac{1}{mn} + \frac{1}{n^2} + o(\frac{1}{n^2})$.

(d) $\sum_{k=1}^{m} w_k h_k^2 = \frac{1}{Nn^4}[\frac{1}{5}nm^5 + \frac{1}{2}nm^4 - \frac{1}{3}m^6 + o(nm^4) + o(m^6)]$.

(e) $\sum_{k=1}^{m} h_k^2 = \frac{m^5}{5n^4} + \frac{m^4}{2n^4} + o(\frac{m^4}{n^4})$.

(f) $\Delta = \sum_{k=1}^{m} w_k h_k^2 - \bar{h}_w^2 = \frac{1}{N^2 n^4}[\frac{4}{45} n^2 m^6 + \frac{1}{6} n^2 m^5 - \frac{1}{5} nm^7 + o(nm^7) + o(n^2 m^5)]$.

Lemma 2. *Assume that $m \to \infty$ and $m/n \to 0$. We have*

(a) $\sum_{k=1}^{m} b_k = m - \frac{5m^2}{8n} + o(\frac{m^2}{n})$.

(b) $\sum_{k=1}^{l-1} b_k = \frac{9}{4} l - \frac{5l^3}{4m^2} + o(l) + O(1), \quad 1 \le l \le m$.

(c) $\sum_{k=1}^{[(l-1)/2]} b_k = \frac{9}{8} l - \frac{5l^3}{32m^2} + o(l) + O(1), \quad 1 \le l \le 2m$.

(d) $\sum_{k=1}^{m} b_k^2 = \frac{9}{4} m + o(m)$.

(d) $\sum_{k=1}^{m/2} b_k b_{2k} = \frac{9}{8} m + o(m)$.

(e) $\sum_{k=1}^{m} k b_k = O(m^2), \quad 1 \le l \le m$.

(f) $\sum_{k=1}^{l-1} k b_k = O(l^2), \quad 1 \le l \le m$.

(g) $\sum_{k=1}^{[(l-1)/2]} k b_k = O(l^2), \quad 1 \le l \le 2m$.

(h) $\sum_{k=1}^{m} k^2 b_k = o(m^3)$.

Lemma 3. *Under the same conditions as in Theorem 1, we have*

(a) $g^T D^2 g = O(\frac{m^5}{n^2})$.

(b) $g^T (D \mathrm{diag}(D) u) = O(\frac{m^4}{n})$.

(c) $\mathrm{tr}(D^2) = nm^2 - [\frac{103}{56} - \frac{165}{448} d_1^2 (1 - d_1^2)] m^3 + [\frac{9}{8} + \frac{9}{2}(d_1^2 - \frac{1}{2}) d_1^2] mn + o(m^3 + nm)$.

(d) $\mathrm{tr}\{\mathrm{diag}(D)^2\} = nm^2 - [\frac{103}{56} - \frac{165}{448} d_1^2 (1 - d_1^2)] m^3 + o(m^3)$.

Proof of Lemma 3. (a) Let $g_i = g(x_i)$, $g_i^{'} = g^{'}(x_i)$ and $g_i^{''} = g^{''}(x_i)$ for $i = 1, \ldots, n$. Noting that D is symmetric, we have $g^T D^2 g = g^T D^T D g = (Dg)^T Dg \triangleq p^T p$,

where $p = Dg = (p_1, p_2, \ldots, p_n)^T$. For $i \in [2m+1, n-2m]$, by Lemma 2(h) we have

$$\begin{aligned}
p_i &= g_i(d_0^2 + d_1^2 + d_2^2)\sum_{k=1}^{m} b_k + \sum_{k=1}^{2m}[(d_0d_1 + d_1d_2)b_k I_{1\le k\le m} + d_0d_2 b_{k/2} I_{k\in E}]g_{i-k} \\
&\quad + \sum_{k=1}^{2m}[(d_0d_1 + d_1d_2)b_k I_{1\le k\le m} + d_0d_2 b_{k/2} I_{k\in E}]g_{i+k} \\
&= -2g_i(d_0d_1 + d_1d_2 + d_0d_2)\sum_{k=1}^{m} b_k + (d_0d_1 + d_1d_2)\sum_{k=1}^{m} b_k(g_{i-k} + g_{i+k}) \\
&\quad + d_0d_2\sum_{k=1}^{m} b_k(g_{i-2k} + g_{i+2k}) \\
&= (d_0d_1 + d_1d_2)\sum_{k=1}^{m} b_k(g_{i-k} + g_{i+k} - 2g_i) + d_0d_2\sum_{k=1}^{m} b_k(g_{i-2k} + g_{i+2k} - 2g_i) \\
&= (d_0d_1 + d_1d_2)\sum_{k=1}^{m} b_k(\frac{k^2}{n^2}g_i'' + o(\frac{k^2}{n^2})) + 4d_0d_2\sum_{k=1}^{m} b_k(\frac{k^2}{n^2}g_i'' + o(\frac{k^2}{n^2})) \\
&= o(\frac{m^3}{n^2}).
\end{aligned}$$

For $i \in [1, m]$, by (e), (f) and (g) in Lemma 2, we have

$$\begin{aligned}
p_i &= g_i[d_0^2\sum_{k=1}^{m} b_k + d_1^2\sum_{k=1}^{i-1} b_k + d_2^2\sum_{k=1}^{[(i-1)/2]} b_k] + d_0d_1\sum_{k=[(i+1)/2]}^{i-1} b_k g_{i-k} \\
&\quad + (d_0d_1 + d_1d_2)\sum_{k=1}^{[(i-1)/2]} b_k g_{i-k} + (d_0d_1 + d_1d_2)\sum_{k=1}^{i-1} b_k g_{i+k} + d_0d_1\sum_{k=i}^{m} b_k g_{i+k} \\
&\quad + d_0d_2\sum_{k=1}^{[(i-1)/2]} b_k g_{i-2k} + d_0d_2\sum_{k=1}^{m} b_k g_{i+2k} \\
&= d_0d_2\sum_{k=1}^{[(i-1)/2]} b_k(g_{i-2k} - g_i) + d_1d_2\sum_{k=1}^{[(i-1)/2]} b_k(g_{i-k} - g_i) + d_0d_1\sum_{k=1}^{i-1} b_k(g_{i-k} - g_i) \\
&\quad + d_1d_2\sum_{k=1}^{i-1} b_k(g_{i+k} - g_i) + d_0d_1\sum_{k=1}^{m} b_k(g_{i+k} - g_i) + d_0d_2\sum_{k=1}^{m} b_k(g_{i+2k} - g_i) \\
&= O(\frac{m^2}{n}).
\end{aligned}$$

For $i \in [m+1, 2m]$, by (e) and (g) in Lemma 2, we have,

$$\begin{aligned}
p_i &= g_i[d_0^2\sum_{k=1}^{m} b_k + d_1^2\sum_{k=1}^{m} b_k + d_2^2 \sum_{k=1}^{[(i-1)/2]} b_k] + d_0d_1 \sum_{k=[(i+1)/2]}^{m} b_k g_{i-k} \\
&\quad +(d_0d_1 + d_1d_2) \sum_{k=1}^{[(i-1)/2]} b_k g_{i-k} + (d_0d_1 + d_1d_2)\sum_{k=1}^{m} b_k g_{i+k} \\
&\quad +d_0d_2 \sum_{k=1}^{[(i-1)/2]} b_k g_{i-2k} + d_0d_2 \sum_{k=1}^{m} b_k g_{i+2k} \\
&= (d_0d_1 + d_1d_2)\sum_{k=1}^{m} b_k(g_{i+k} - g_i) + d_0d_2 \sum_{k=1}^{m} b_k(g_{i+2k} - g_i) + d_0d_1 \sum_{k=1}^{m} b_k(g_{i-k} - g_i) \\
&\quad +d_0d_2 \sum_{k=1}^{[(i-1)/2]} b_k(g_{i-2k} - g_i) + d_1d_2 \sum_{k=1}^{[(i-1)/2]} b_k(g_{i-k} - g_i) \\
&= O(\frac{m^2}{n}).
\end{aligned}$$

Similarly, we have $p_i = O(m^2/n)$ for $i \in [n-2m+1, n]$. This leads to

$$g^T D^2 g = \sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{2m} p_i^2 + \sum_{i=2m+1}^{n-2m} p_i^2 + \sum_{i=n-2m+1}^{n} p_i^2 = O(\frac{m^5}{n^2}).$$

(b) By Lemma 2(a), 2(b) and 2(c), we have

$$\begin{aligned}
g^T(D\mathrm{diag}(D)u) &= p^T \mathrm{diag}(D)u \\
&= (\sum_{i=1}^{m} + \sum_{i=m+1}^{2m} + \sum_{i=1+2m}^{n-2m} + \sum_{i=n-2m+1}^{n-m} + \sum_{i=n-m+1}^{n})p_i d_{ii} \\
&= O(\frac{m^4}{n}).
\end{aligned}$$

(c) We divide $\mathrm{tr}(D^2)$ into five parts as follows:

$$\mathrm{tr}(D^2) = \Big(\sum_{i=1}^{m}\sum_{j=1}^{n} + \sum_{i=m+1}^{2m}\sum_{j=1}^{n} + \sum_{i=2m+1}^{n-2m}\sum_{j=1}^{n} + \sum_{i=n-2m+1}^{n-m}\sum_{j=1}^{n} + \sum_{i=n-m+1}^{n}\sum_{j=1}^{n}\Big) d_{ij}^2.$$

For the first part,

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}^2 &= \sum_{i=1}^{m} d_{ii}^2 + \sum_{i=1}^{m}\sum_{j=1,j\neq i}^{n} d_{ij}^2 \\
&= \sum_{i=1}^{m}[d_0^2\sum_{k=1}^{m} b_k + d_1^2\sum_{k=1}^{i-1} b_k + d_2^2\sum_{k=1}^{[(i-1)/2]}\ ]^2 + o(m^3) \\
&= \sum_{i=1}^{m}[d_0^2(m - \frac{5m^2}{8n} + o(\frac{m^2}{n})) + d_1^2(\frac{9}{4}l - \frac{5l^3}{4m^2} + o(l) + O(1)) \\
&\quad + d_2^2(\frac{9}{8}l - \frac{5l^3}{32m^2} + o(l) + O(1))]^2 + o(m^3) \\
&= (d_0^4 + \frac{11}{14}d_1^4 + \frac{2545}{7168}d_2^4 + \frac{13}{8}d_0^2d_1^2 + \frac{233}{224}d_1^2d_2^2 + \frac{67}{64}d_0^2d_2^2)m^3 + o(m^3).
\end{aligned}$$

For the second part,

$$\begin{aligned}
\sum_{i=m+1}^{2m}\sum_{j=1}^{n} d_{ij}^2 &= \sum_{i=1}^{m} d_{ii}^2 + \sum_{i=1}^{m}\sum_{j=1,j\neq i}^{n} d_{ij}^2 \\
&= \sum_{i=1}^{m}[(d_0^2 + d_1^2)\sum_{k=1}^{m} b_k + d_2^2\sum_{k=1}^{[(i-1)/2]}\ ]^2 + o(m^3) \\
&= \sum_{i=1}^{m}[(d_0^2 + d_1^2)(m - \frac{5m^2}{8n} + o(\frac{m^2}{n})) + d_2^2(\frac{9}{8}l - \frac{5l^3}{32m^2} + o(l) + O(1))]^2 + o(m^3) \\
&= (d_0^4 + d_1^4 + \frac{8719}{7168}d_2^4 + 2d_0^2d_1^2 + \frac{141}{64}d_1^2d_2^2 + \frac{141}{64}d_0^2d_2^2)m^3 + o(m^3).
\end{aligned}$$

For the third part,

$$\begin{aligned}
\sum_{i=2m+1}^{n-2m}\sum_{j=1}^{n} d_{ij}^2 &= (n-4m)\{(\sum_{k=1}^{m} b_k)^2 + 2\sum_{k=1}^{2m}[(d_0d_1 + d_1d_2)b_kI_{(1\leq k\leq m)} + d_0d_2b_{k/2}I_{(k\in E)}]^2\} \\
&= (n-4m)\Big[m^2 - \frac{5m^3}{4n} + \frac{m^3}{n} + 2(\frac{9}{16} + \frac{9}{4}d_0d_1^2d_2)m + o(m)\Big] \\
&= nm^2 - \frac{21}{4}m^3 + (\frac{9}{8} + \frac{9}{2}d_0d_1^2d_2)nm + o(nm) + o(m^3),
\end{aligned}$$

where

$$\begin{aligned}
&\sum_{k=1}^{2m}[(d_0d_1 + d_1d_2)b_kI_{(1\leq k\leq m)} + d_0d_2b_{k/2}I_{(k\in E)}]^2 \\
= &\ \big[(d_0d_1 + d_1d_2)^2 + d_0^2d_2^2\big]\sum_{k=1}^{m} b_k^2 + 2d_0d_2(d_0d_1 + d_1d_2)\sum_{k=1}^{[m/2]} b_{2k}b_k \\
= &\ \big[d_0^2d_1^2 + d_1^2d_2^2 + d_0^2d_2^2 + 2d_0d_1^2d_2 + d_0^2d_1d_2 + d_0d_1d_2^2\big](\frac{9}{4}m + o(m)) \\
= &\ (\frac{9}{16} + \frac{9}{4}d_0d_1^2d_2)m + o(m).
\end{aligned}$$

For the fourth part,

$$\sum_{i=n-m+1}^{n}\sum_{j=1}^{n} d_{ij}^2 = (\frac{2545}{7168}d_0^4 + \frac{11}{14}d_1^4 + d_2^4 + \frac{233}{224}d_0^2d_1^2 + \frac{13}{8}d_1^2d_2^2 + \frac{67}{64}d_0^2d_2^2)m^3 + o(m^3).$$

For the fifth part,

$$\sum_{i=n-2m+1}^{n}\sum_{j=1}^{n} d_{ij}^2 = (\frac{8719}{7168}d_0^4 + d_1^4 + d_2^4 + \frac{141}{64}d_0^2d_1^2 + 2d_1^2d_2^2 + \frac{141}{64}d_0^2d_2^2)m^3 + o(m^3).$$

Finally, we have

$$\begin{aligned}
\mathrm{tr}(D^2) &= nm^2 - \frac{21}{4}m^3 + (\frac{9}{8} + \frac{9}{2}d_0d_1^2d_2)nm + o(nm) + o(m^3) \\
&\quad +(\frac{25}{7}(d_0^4 + d_1^4 + d_2^4) + \frac{3077}{448}d_0^2d_1^2 + \frac{13}{2}d_0^2d_1^2 + \frac{3077}{448}d_1^2d_2^2)m^3 + o(m^3) \\
&= nm^2 - \left[\frac{103}{56} - \frac{165}{448}d_1^2(1-d_1^2)\right]m^3 + \left[\frac{9}{8} + \frac{9}{2}(d_1^2 - \frac{1}{2})d_1^2\right]mn + o(m^3 + nm).
\end{aligned}$$

(d) By the results in the derivation of $\mathrm{tr}(D^2)$, we have

$$\begin{aligned}
\mathrm{tr}\{\mathrm{diag}(D)^2\} &= (\sum_{i=1}^{m} + \sum_{i=m+1}^{2m} + \sum_{i=1+2m}^{n-2m} + \sum_{i=n-2m+1}^{n-m} + \sum_{i=n-m+1}^{n})d_{ii}^2 \\
&= nm^2 - [\frac{103}{56} - \frac{165}{448}d_1^2(1-d_1^2)]m^3 + o(m^3).
\end{aligned}$$

Proof of Theorem 3. By Lemma 3, the variance of $\hat{\sigma}^2$ is given by

$$\begin{aligned}
\mathrm{var}(\hat{\sigma}^2(2,m)) &= \frac{1}{\mathrm{tr}(D^2)}[4\sigma^2 g^T D^2 g + 4g^T(D\mathrm{diag}(D)u)\sigma^3\gamma_3 \\
&\quad +\sigma^4 tr(\mathrm{diag}(D)^2)(\gamma_4 - 3) + 2\sigma^4\mathrm{tr}(D^2)] \\
&= \frac{1}{N^2}\Big\{O(\frac{m^5}{n^2}) + O(\frac{m^4}{n}) \\
&\quad +(\mathrm{var}(\varepsilon^2) - 2\sigma^4)\big(nm^2 - [\frac{103}{56} - \frac{165}{448}d_1^2(1-d_1^2)]m^3 + o(m^3)\big) \\
&\quad +2\sigma^4\big(nm^2 - [\frac{103}{56} - \frac{165}{448}d_1^2(1-d_1^2)]m^3 \\
&\quad +[\frac{9}{8} + \frac{9}{2}(d_1^2 - \frac{1}{2})d_1^2]mn + o(m^3) + o(nm)\big)\Big\} \\
&= \frac{1}{n}\mathrm{var}(\varepsilon^2) + A_1\frac{\sigma^4}{mn} + A_2\frac{m}{n^2}\mathrm{var}(\varepsilon^2) + o(\frac{1}{nm}) + o(\frac{m}{n^2}),
\end{aligned}$$

where A_1 are A_2 are the same as in Theorem 3. This proves the theorem.

Chapter 3

Difference-based Variance Estimation in Nonparametric Regression with Repeated Measurements

3.1 Introduction

Repeated measurements are commonly available in many statistical problems. For example, the temperature at a certain location is often recorded at various times, a physical feature is sometimes evaluated by different technicians to reduce evaluation bias, nutrition intake of an individual on a certain day is routinely measured by several different formats to reduce reporting error. Thus, how to take advantage of the repeated measurements and develop a variance estimator that has the same advantage of not requiring a mean estimation is of great importance.

Consider the nonparametric regression model with repeated measurements,

$$Y_{ij} = f(x_i) + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m, \tag{3.1}$$

where Y_{ij} are observations, x_i are design points, f is an unknown mean function, and ε_{ij} are i.i.d. random errors with mean zero and variance σ^2.

Despite the rich literature on difference-based variance estimation for model (2.1), very little attention has been paid to model (3.1) with $m \geq 2$. Gasser et al. (1986) encountered the multiple measurements issue, but they decided to order the data sequentially and treat them as if they came from different design points. Thus, the multiple measurements feature is ignored. This is quite a pity, since intuitively the repeated measurements contain different type of information, and this new information should be taken into account in constructing estimators. We suspect that one reason very few work is available for treating multiple observations in difference-based variance estimation literature is that it is not easy to combine the between-design-point difference and the within-design-point difference properly. In addition, even if a certain new treatment is proposed, it is not straightforward to analyze how effective this treatment is in theory. For example, it is difficult to know if the treatment has optimal large sample property, in other words, it is difficult to know if a better method can be found in treating the multiple measurements, either within the difference-based method family or overall. In this work, we will fill this literature in both aspects. Specifically, we will propose three new difference-based methods to utilize the multiple measurements, respectively the sample variance method, the partitioning method and the sequencing method. We analyze these methods and illustrate the practical advantages of each method under different data structures and/or model assumptions.

The rest of the chapter is organized as the following. In Section 3.2, we propose three difference-based methods for estimating σ^2 in nonparametric regression with balanced repeated measurements: the sample variance method, the partitioning method, and the sequencing method. We also explore their asymptotic properties, especially for the proposed sequencing estimator, where we derive its asymptotic mean squared error (MSE), its optimal bandwidth and its asymptotic normality. Extensive simulation studies are conducted in Section 3.3 to evaluate and compare the finite sample performance of the proposed estimators to the residual-based estimator. We then extend the methods to the nonparametric regression models with unbalanced repeated measurement data in Section 3.4. Two real data examples are analyzed in

Section 3.5 to demonstrate the practical usefulness of the proposed methods, and we conclude the chapter in Section 3.6 with a brief discussion. All the technical proofs are provided in Section 3.7.

3.2 Main Results

To estimate σ^2 in model (3.1), a naive approach is to evade the issue of repeated measurements by taking average of the observations at each design point. Assume that $x_i = i/n$ for all i. Let the averaged observations be

$$\bar{Y}_i = \frac{1}{m}\sum_{j=1}^{m} Y_{ij} = f(x_i) + \frac{1}{m}\sum_{j=1}^{m} \varepsilon_{ij}, \qquad i = 1, \ldots, n.$$

Given that ε_{ij} are i.i.d. random errors with variance σ^2, we have $\mathrm{Var}(\bar{Y}_i) = \sigma^2/m$. Then to estimate σ^2, we multiply the sequence $\bar{Y}_i$ by $\sqrt{m}$ and then apply Tong and Wang (2005)'s method to the new sequence to get the estimation. We name this estimator the averaging estimator, written as $\hat{\sigma}^2_{\text{naive}}$. At the asymptotically optimal bandwidth $h_{opt} = \{28n\sigma^4/\mathrm{Var}(\varepsilon^2)\}^{1/2}$ in Tong and Wang (2005), the MSE of $\hat{\sigma}^2_{\text{naive}}$ is

$$\mathrm{MSE}\{\hat{\sigma}^2_{\text{naive}}(h_{opt})\} = \frac{1}{n}\mathrm{Var}(\varepsilon^2) + O(n^{-3/2}). \tag{3.2}$$

The number of repeats m does not appear in (3.2), hence the naive method clearly does not take advantage of the repeated measurements. Specifically, by taking averages, this method sacrifices the information contained in the repeated measurements for simplicity. Further, multiplying the average sequence by $\sqrt{m}$ enlarged the mean function. As a consequence, the trend in the mean function is less negligible in finite sample settings. The analysis on the naive method above indicates that in nonparametric regression with repeated measurements, there are two types of information we can use and should probably treat differently: (i) the variation within design points, and (ii) the variation between design points.

In what follows, we propose three new methods for estimating σ^2 in nonparametric regression with repeated measurements. The first method is the sample variance method where only the variation within design points is used. The second method

proposed is the partitioning method where only the variation between design points is used. Whereas our third method, the sequencing method, uses both types of variations. The statistical properties of all three methods will be investigated.

3.2.1 Sample Variance Method

Our first method aims to intelligently use the existence of repeated measurements for the variance estimation. Let $s_i^2 = \sum_{j=1}^{m}(Y_{ij} - \bar{Y}_i)^2/(m-1)$ be the sample variance of the repeated measurements at the ith design point, $i = 1, \ldots, n$. Given that $Y_{i1}, \ldots, Y_{im}$ are i.i.d. random variables, we have $E(s_i^2) = \sigma^2$. Note also that $s_1^2, \ldots, s_n^2$ are independent of each other. We define the sample variance estimator of σ^2 as

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2.$$

It is clear that $\hat{\sigma}_1^2$ is an unbiased estimator of σ^2. By Rose and Smith (2002), we have

$$\mathrm{Var}\left(s_i^2\right) = \frac{1}{m}\mathrm{Var}(\varepsilon^2) + \frac{2}{m(m-1)}\sigma^4.$$

This leads to

$$\mathrm{MSE}(\hat{\sigma}_1^2) = \mathrm{Var}(\hat{\sigma}_1^2) + \mathrm{Bias}^2(\hat{\sigma}_1^2) = \frac{1}{mn}\mathrm{Var}(\varepsilon^2) + \frac{2}{m(m-1)n}\sigma^4.$$

Note that Hall and Marron (1990) showed that the residual-based estimators can achieve an optimal estimation variance $\mathrm{Var}(\varepsilon^2)/(nm)$, hence $\hat{\sigma}_1^2$ is not optimal. When m is large, the discrepancy can be very small. Specifically, when $m \to \infty$, the second term in the above display is negligible and $\hat{\sigma}_1^2$ is asymptotically the best unbiased estimator of σ^2. However, when m is small, the sample variance estimator is clearly suboptimal. In this work, we are particularly interested in the scenario where m is fixed but n is large.

An especially nice feature of $\hat{\sigma}_1^2$ is that it is completely free from any assumptions on the mean function f. This makes the sample variance estimator robust in the most general nonparametric settings, especially when the mean function is nonsmooth, noncontinuous or highly oscillating so that other difference-based methods fail to perform well. Finally, the sample variance estimator is extremely easy to implement in practice.

3.2.2 Partitioning Method

Our second method is a partitioning method. We first partition the observations Y_{ij} into m groups according to the following sampling-based algorithm.

(i) Sample one observation from the set $\{Y_{i1}, \ldots, Y_{im}\}$ for each i to form the first response group $G(1) = \{Y_{1g_1}, \ldots, Y_{ng_1}\}$.

(ii) Sample one observation from the remaining set $\{Y_{i1}, \ldots, Y_{im}\} \setminus \{Y_{ig_1}\}$ for each i to form the second response group $G(2) = \{Y_{1g_2}, \ldots, Y_{ng_2}\}$.

(iii) Repeat Step (ii), until we obtain the last response group $G(m) = \{Y_{1g_m}, \ldots, Y_{ng_m}\}$.

We then apply Tong and Wang (2005)'s method to each group $G(j)$ to get the estimates $\hat{\sigma}^2_{(j)}$, $j = 1, \ldots, m$. The final estimator is defined as

$$\hat{\sigma}^2_2 = \frac{1}{m} \sum_{j=1}^{m} \hat{\sigma}^2_{(j)}.$$

We refer to it as the partitioning estimator of variance.

Under the model assumption, the groups $G(1), \ldots, G(m)$ are independent of each other. Therefore, the estimators $\sigma^2_{(j)}$ are also independent of each other. Then with the optimal bandwidth $h_{opt} = \{28n\sigma^4/\mathrm{Var}(\varepsilon^2)\}^{1/2}$, the MSE of $\hat{\sigma}^2_2$ is given as

$$\mathrm{MSE}\{\hat{\sigma}^2_2(h_{opt})\} = \frac{1}{nm}\mathrm{Var}(\varepsilon^2) + \frac{9\sqrt{7}}{28mn^{3/2}}\sigma^2\{\mathrm{Var}(\varepsilon^2)\}^{1/2} + o(\frac{1}{n^{3/2}}). \tag{3.3}$$

This shows that the proposed partitioning estimator achieves the same asymptotically optimal estimation variance as that for the residual-based estimators.

3.2.3 Sequencing Method

We first order the observations as $\{Y_{11}, \ldots, Y_{1m}, \ldots, Y_{n1}, \ldots, Y_{nm}\}$ and relabel the indices as $l = 1, 2, \ldots, nm$. With this notation, model (3.1) can be written as

$$Z_l = f(t_l) + \epsilon_l, \qquad l = 1, \ldots, nm, \tag{3.4}$$

where $\{Z_1, Z_2, \ldots, Z_{nm}\} = \{Y_{11}, \ldots, Y_{1m}, \ldots, Y_{n1}, \ldots, Y_{nm}\}$, $\{t_1, t_2, \ldots, t_{nm}\} = \{x_1, \ldots, x_1, \ldots, x_n, \ldots, x_n\}$, and $\{\epsilon_l, \epsilon_2, \ldots, \epsilon_{nm}\} = \{\varepsilon_{11}, \ldots, \varepsilon_{1m}, \ldots, \varepsilon_{n1}, \ldots, \varepsilon_{nm}\}$.

For model (3.4), we define the lag-p Rice estimator

$$\hat{\sigma}^2_{\mathrm{R}}(p) = \frac{1}{2(nm-p)} \sum_{l=p+1}^{nm} (Z_l - Z_{l-p})^2, \qquad \text{for } p = 1, \ldots, nm-1.$$

Note that the first m lag-p Rice estimators only use differences of the identical or consecutive design points, i.e., none of the $f(x_i) - f(x_{i-r})$ terms with $r \geq 2$ are involved in the first m lag-p Rice estimators. We thus combine them and define a new Rice-type estimator using the weighted average of the first m lag-p Rice estimators,

$$\begin{aligned}
\hat{\sigma}^2_{\mathrm{Rt}} &= \frac{1}{m^2 n - m(m+1)/2} \sum_{p=1}^{m} (nm-p)\hat{\sigma}^2_{\mathrm{R}}(p) \\
&= \frac{1}{2m^2 n - m(m+1)} \left\{ \sum_{k=1}^{m-1} \sum_{i=1}^{n} \sum_{j=k+1}^{m} (Y_{ij} - Y_{i,j-k})^2 + \sum_{k=1}^{m} \sum_{i=2}^{n} \sum_{j=1}^{k} (Y_{ij} - Y_{i-1,m-k+j})^2 \right\},
\end{aligned}$$

where the weight for $\hat{\sigma}^2_{\mathrm{R}}(p)$ is assigned because the lag-p Rice estimator uses $(nm-p)$ pairs of data.

Some algebra yields

$$\begin{aligned}
E(\hat{\sigma}^2_{\mathrm{Rt}}) &= \sigma^2 + \frac{1}{2m^2 n - m(m+1)} \sum_{k=1}^{m} \sum_{i=2}^{n} \sum_{j=1}^{k} \{f(x_i) - f(x_{i-1})\}^2 \\
&= \sigma^2 + \frac{m(m+1)/2}{2m^2 n - m(m+1)} \sum_{i=2}^{n} \{f(x_i) - f(x_{i-1})\}^2 .
\end{aligned}$$

This reveals that the Rice-type estimator $\hat{\sigma}^2_{\mathrm{Rt}}$ is always positively biased, unless f is a constant function. Suppose that f has a bounded first derivative. By the Taylor expansion we have

$$E(\hat{\sigma}^2_{\mathrm{Rt}}) = \sigma^2 + \frac{(n-1)m(m+1)}{n^2\{2m^2 n - m(m+1)\}} J + o(\frac{1}{n^2}), \tag{3.5}$$

where $J = \int_0^1 \{f'(x)\}^2 dx/2$. To eliminate the bias term in (3.5), we further define the lag-r Rice-type estimators

$$\begin{aligned}
\hat{\sigma}^2_{\mathrm{Rt}}(r) &= \frac{1}{c_r} \sum_{p=(r-1)m+1}^{rm} (nm-p)\hat{\sigma}^2_{\mathrm{R}}(p) \qquad (3.6) \\
&= \frac{1}{2c_r} \left\{ \sum_{k=1}^{m-1} \sum_{i=r}^{n} \sum_{j=k+1}^{m} (Y_{ij} - Y_{i-r+1,j-k})^2 + \sum_{k=1}^{m} \sum_{i=r+1}^{n} \sum_{j=1}^{k} (Y_{ij} - Y_{i-r,m-k+j})^2 \right\},
\end{aligned}$$

where $r = 1, 2, n-1$, and $c_r = \sum_{p=(r-1)m+1}^{rm}(nm-p) = m^2n - rm^2 + m(m-1)/2$. By definition, $\hat{\sigma}^2_{\mathrm{Rt}} = \hat{\sigma}^2_{\mathrm{Rt}}(1)$. Similar calculation at any fixed $r = o(n)$ yields

$$\begin{aligned} & E\left\{\hat{\sigma}^2_{\mathrm{Rt}}(r)\right\} \\ = \; & \sigma^2 + \frac{1}{2c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}\left\{f(x_i) - f(x_{i-r+1})\right\}^2 + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}\left\{f(x_i) - f(x_{i-r})\right\}^2\right] \\ = \; & \sigma^2 + Jd_r + o(r^2/n^2), \end{aligned} \tag{3.7}$$

where

$$d_r = \frac{m\left\{(m-1)(n-r+1)(r-1)^2 + (m+1)(n-r)r^2\right\}}{2c_r n^2}. \tag{3.8}$$

The relation in (3.7) indicates that the lag-r Rice-type estimator $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ has a linear relationship with the quantity d_r. Taking advantage of this relation, we fit a linear regression model by treating $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ as the response variable and d_r as the covariate, and estimate σ^2 as the intercept of the linear model.

We choose the first b pairs of $\{d_r, \hat{\sigma}^2_{\mathrm{Rt}}(r)\}$ to perform the regression, where $b = o(n)$. The choice of b will be discussed later. In performing the linear regression estimation, because $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ involves c_r pairs of data, we assign weight $w_r = c_r/s_b$ to the rth observation, where $s_b = \sum_{r=1}^{b} c_r = m^2nb - m^2b(b+1)/2 + m(m-1)b/2$. We then minimize the weighted sum of squares $\sum_{r=1}^{b} w_r\{\hat{\sigma}^2_{\mathrm{Rt}}(r) - \alpha - \beta d_r\}^2$ to fit the linear model

$$\hat{\sigma}^2_{\mathrm{Rt}}(r) = \alpha + \beta d_r + e_r, \qquad r = 1, \ldots, b. \tag{3.9}$$

For ease of notation, let $\bar{\sigma}^2_w = \sum_{r=1}^{b} w_r\hat{\sigma}^2_{\mathrm{Rt}}(r)$ and $\bar{d}_w = \sum_{r=1}^{b} w_r d_r$. Then the sequencing estimator of σ^2 is given as

$$\hat{\sigma}^2_3 = \hat{\alpha} = \bar{\sigma}^2_w - \hat{\beta}\bar{d}_w, \tag{3.10}$$

where $\hat{\beta} = \sum_{r=1}^{b} w_r\hat{\sigma}^2_{\mathrm{Rt}}(r)(d_r - \bar{d}_w)/\sum_{r=1}^{b} w_r(d_r - \bar{d}_w)^2$ is the fitted slope. In Section 3.7.1 we prove that

Theorem 4. *For the equally spaced design, $\hat{\sigma}^2_3$ is an unbiased estimator of σ^2 when f is a linear function, regardless of the choice of b.*

In what follows we establish further statistical properties of the sequencing estimator $\hat{\sigma}_3^2$. For notational convenience, we let $\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1m}, \ldots, Y_{n1}, \ldots, Y_{nm})^T$, $\boldsymbol{f} = \{f(x_1), \ldots, f(x_1), \ldots, f(x_n), \ldots, f(x_n)\}^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \ldots, \varepsilon_{1m}, \ldots, \varepsilon_{n1}, \ldots, \varepsilon_{nm})^T$. Then $\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$. Also let $\boldsymbol{u} = (1, \ldots, 1)^T$, $\gamma_i = E(\varepsilon^i/\sigma^i)$ for $i = 3, 4$, and assume that $\gamma_4 > 1$.

Quadratic Form Representation

Let $\tau_0 = 0$ and $\tau_r = 1 - \bar{d}_w(d_r - \bar{d}_w)/\sum_{r=1}^b w_r(d_r - \bar{d}_w)^2$, $r = 1, \ldots, b$. By (3.10),

$$\hat{\sigma}_3^2 = \sum_{r=1}^{b} \tau_r w_r \hat{\sigma}_{\mathrm{Rt}}^2(r) = \frac{1}{2s_b} \sum_{r=1}^{b} \left\{ \tau_r \sum_{p=(r-1)m+1}^{rm} \sum_{l=p+1}^{nm} (Z_l - Z_{l-p})^2 \right\}.$$

With some algebra, we can write $\hat{\sigma}_3^2$ as

$$\hat{\sigma}_3^2 = \frac{1}{2s_b} \boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y},$$

where $\boldsymbol{D}$ is an $(nm) \times (nm)$ symmetric matrix with elements

$$\boldsymbol{D}_{ij} = \begin{cases} d_{ii}(a), & (a-1)m < i = j \le am \text{ with } a = 1, \ldots, n, \\ -\tau_a, & (a-1)m < |i-j| \le am \text{ with } a = 1, \ldots, b, \\ 0, & \text{otherwise,} \end{cases}$$

where $d_{ii}(a) = m\sum_{r=1}^b \tau_r + m\sum_{r=0}^{a-1} \tau_r + \{i - 1 - (a-1)m\}\tau_a$ for $a = 1, \ldots, b$; $d_{ii}(a) = 2m\sum_{r=1}^b \tau_r$ for $a = b+1, \ldots, n-b$; and $d_{ii}(a) = m\sum_{r=1}^b \tau_r + m\sum_{r=0}^{n-a} \tau_r + (am - i)\tau_{n+1-a}$ for $a = n-b+1, \ldots, n$.

Note that $\boldsymbol{D}$ depends on the design points only. By letting $f = 0$, we have

$$E(\hat{\sigma}_3^2) = \frac{1}{2s_b} E(\boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y}) = \frac{1}{2s_b} E(\boldsymbol{\varepsilon}^T \boldsymbol{D} \boldsymbol{\varepsilon}) = \frac{\sigma^2}{2s_b} \mathrm{tr}(\boldsymbol{D}),$$

Now because of Theorem 4, $\hat{\sigma}_3^2$ is unbiased when $f = 0$, we have $\mathrm{tr}(\boldsymbol{D}) = 2s_b$. This shows that the proposed sequencing estimator possesses a quadratic form,

$$\hat{\sigma}_3^2 = \boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y} / \mathrm{tr}(\boldsymbol{D}). \tag{3.11}$$

Asymptotic MSE and Optimal Bandwidth

The quadratic form representation (3.11) of σ_3^2 enables us to take advantage of the existing results in Dette et al. (1998) and directly obtain

$$\begin{aligned}\mathrm{MSE}(\hat{\sigma}_3^2) &= \big[(\boldsymbol{f}^T\boldsymbol{D}\boldsymbol{f})^2 + 4\sigma^2\boldsymbol{f}^T\boldsymbol{D}^2\boldsymbol{f} + 4\boldsymbol{f}^T\{\boldsymbol{D}\cdot\mathrm{diag}(\boldsymbol{D})\boldsymbol{u}\}\sigma^3\gamma_3 \\ &\quad + \sigma^4(\gamma_4-3)\mathrm{tr}[\mathrm{diag}(\boldsymbol{D})^2] + 2\sigma^4\mathrm{tr}(\boldsymbol{D}^2)\big]/\{\mathrm{tr}(\boldsymbol{D})\}^2, \qquad (3.12)\end{aligned}$$

where $\mathrm{diag}(\boldsymbol{D})$ denotes the diagonal matrix of the diagonal elements of $\boldsymbol{D}$. The first term in (3.12) represents the squared bias, and the last four terms represent the variance term of the estimator. In the case when the random errors are normally distributed, $\gamma_3 = 0$ and $\gamma_4 = 3$ so that the third and fourth terms vanish.

Theorem 5. *Assume that f has a bounded second derivative. For the equally spaced design with $b \to \infty$ and $b/n \to 0$, we have*

$$\mathrm{Bias}(\hat{\sigma}_3^2) = O(b^3n^{-3}), \qquad (3.13)$$

$$\mathrm{Var}(\hat{\sigma}_3^2) = \frac{\mathrm{Var}(\varepsilon^2)}{mn} + \frac{9\sigma^4}{4m^2nb} + \frac{9b\mathrm{Var}(\varepsilon^2)}{112mn^2} + o\{(nb)^{-1} + bn^{-2}\}, \qquad (3.14)$$

$$\mathrm{MSE}(\hat{\sigma}_3^2) = \frac{\mathrm{Var}(\varepsilon^2)}{mn} + \frac{9\sigma^4}{4m^2nb} + \frac{9b\mathrm{Var}(\varepsilon^2)}{112mn^2} + o\{(nb)^{-1} + bn^{-2}\} + O(b^6n^{-6}). \qquad (3.15)$$

Theorem 5 indicates that $\hat{\sigma}_3^2$ is a consistent estimator of σ^2, and its MSE reaches the asymptotically optimal rate (Dette et al.; 1998). By (3.15), the asymptotically optimal bandwidth in terms of minimizing the MSE is given as

$$b_{opt} = \left\{\frac{28n\sigma^4}{m\mathrm{Var}(\varepsilon^2)}\right\}^{1/2}. \qquad (3.16)$$

It is interesting to point out that b_{opt} does not depend on the mean function f. We also note that b_{opt} is a decreasing function of m. Substituting (3.16) into (3.15) leads to

$$\mathrm{MSE}\{\hat{\sigma}_3^2(b_{opt})\} = \frac{1}{nm}\mathrm{Var}(\varepsilon^2) + \frac{9\sqrt{7}}{28m^{3/2}n^{3/2}}\sigma^2\{\mathrm{Var}(\varepsilon^2)\}^{1/2} + o(1/n^{3/2}). \qquad (3.17)$$

Comparing (3.3) and (3.17), we have $\mathrm{MSE}\{\hat{\sigma}_3^2(b_{opt})\} < \mathrm{MSE}\{\hat{\sigma}_2^2(h_{opt})\}$ for any $m \geq 2$. This implies that the sequencing estimator behaves asymptotically better than the partitioning estimator in the presence of repeated measurements. Note also that $b_{opt} = h_{opt}/m^{1/2}$. When $m = 1$, $b_{opt} = h_{opt}$ and the two estimators are identical.

Adaptive Choice of Bandwidth

For simplicity, we use normal random errors to illustrate the choice of bandwidth in the finite sample situation. When the errors are not normal, the only additional complexity is to estimate the ratio $\gamma_4 = \mathrm{Var}(\varepsilon^2)/\sigma^4$; all other aspects of the bandwidth selection procedure remain the same as in the normal error case.

For normal random errors, $\mathrm{Var}(\varepsilon^2) = 2\sigma^4$ so that b_{opt} is simplified as $(14n/m)^{1/2}$, which does not depend on the smoothness of the mean function and the magnitude of residual variance. We caution here that the above b_{opt} applies for large n only. When n is small or when f is rough, the performance of b_{opt} is sometimes not satisfactory in practice. This was also observed in Tong and Wang (2005) for $m = 1$. This is because some higher order terms ignored in the calculation of the asymptotic MSE for the estimator (3.15) indeed depend on the smoothness of the function. Consequently, we need a smaller bandwidth to diminish the impact of the mean function in the finite sample case. Simulation studies (not shown) indicate that the bandwidth choices for $m = 1$ in Tong and Wang (2005) often work well for $m \geq 2$, as long as m is not too large (say, $m \leq 20$). In summary, we suggest to use (i) $b_s = n^{1/2}$ for large n, and (ii) $b_t = n^{1/3}$ for small n or for rough f. In the remainder of this chapter, we take the integer part of b_s and b_t whenever necessary.

A cross validation (CV) strategy can also be applied to select the bandwidth. Specifically, we first split the whole data set into V disjoint subsamples $\{S_1, \ldots, S_V\}$, and then select $b = b_{\mathrm{CV}}$ that minimizes $\mathrm{CV}(b) = \sum_{v=1}^{V} \{\hat{\sigma}_3^2(b) - \hat{\sigma}_{3,v}^2(b)\}^2$, where $\hat{\sigma}_3^2(b)$ and $\hat{\sigma}_{3,v}^2(b)$ are the estimates of σ^2 based on the whole sample $\cup_{i=1}^{V} S_i$ and the subsample $\cup_{i \neq v} S_i$ with bandwidth b, respectively. Note that the design points in $\cup_{i \neq v} S_i$ are not equally spaced on $[0, 1]$. Thus to compute $\hat{\sigma}_{3,v}^2(b)$, we need to use the formula developed in the general design. Finally, the CV method requires much more expensive computation compared to b_s and b_t.

Asymptotic Normality

We have the following asymptotic normality for the Rice-type estimators $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ in (3.6) and for the sequencing estimator $\hat{\sigma}^2_3$ in (3.11). Let $\xrightarrow{\mathcal{D}}$ denote convergence in distribution.

Theorem 6. *Assume that f has a bounded second derivative and $E(\varepsilon^4)$ is finite. Then for any $r = n^\vartheta$ with $0 \le \vartheta < 1/2$, the lag-r Rice-type estimator satisfies*

$$\sqrt{n}\{\hat{\sigma}^2_{Rt}(r) - \sigma^2\} \xrightarrow{\mathcal{D}} N\{0, (\gamma_4 - 1 + 1/m)\sigma^4/m\} \qquad \text{as } n \to \infty.$$

Theorem 7. *Assume that f has a bounded second derivative and $E(\varepsilon^{4+2\delta})$ is finite for some δ in $(0, 1)$. Then for any $b = n^\vartheta$ with $0 < \vartheta < 1/2$, the sequencing estimator $\hat{\sigma}^2_3$ satisfies*

$$\sqrt{n}(\hat{\sigma}^2_3 - \sigma^2) \xrightarrow{\mathcal{D}} N\{0, (\gamma_4 - 1)\sigma^4/m\} \qquad \text{as } n \to \infty.$$

Proofs of Theorems 6 and 7 are given in Sections 3.7.3 and 3.7.4, respectively. Theorem 6 indicates that $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ has the same asymptotic property as "the m-order optimal difference-based estimator" proposed in Hall et al. (1990). Given that $\mathrm{E}(\varepsilon^{4+2\delta})$ is finite for some δ in $(0, 1)$, Theorems 6 and 7 show that the sequencing estimator is more efficient than the Rice-type estimators for any fixed m. Specifically, the efficiency of $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ relative to $\hat{\sigma}^2_3$ is given as $(\gamma_4 - 1)/(\gamma_4 - 1 + 1/m)$, which is an increase function of m. When the random errors are normally distributed, the relative efficiency reduces to $2m/(2m + 1)$. As illustration, the relative efficiency is 66.7% when $m = 1$, 80% when $m = 2$, and 90.9% when $m = 5$. Finally, $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ and $\hat{\sigma}^2_3$ become asymptotically equivalent as $m \to \infty$.

Theorem 7 can be easily used to construct confidence intervals for σ^2. For example, when $mn > (\gamma_4 - 1)z^2_{\alpha/2}$, an approximate $1 - \alpha$ confidence interval for σ^2 is

$$[\hat{\sigma}^2_3/\{1 + z_{\alpha/2}\sqrt{(\gamma_4 - 1)/mn}\},\ \hat{\sigma}^2_3/\{1 - z_{\alpha/2}\sqrt{(\gamma_4 - 1)/mn}\}],$$

where z_α is the upper α-th percentile of the standard normal distribution. For normal data, the parameter $\gamma_4 = 3$ so the confidence interval is fully specified. In general, γ_4 needs to be replaced by an estimate.

Generalized Least Squares Estimator

In constructing the Rice-type estimators at different lags, we have used the same observations to form different pairs. Thus, our linear regression model (3.9) concerns correlated data. When the responses are correlated, the proper way of performing linear regression is the generalized least squares (GLS) method, where the optimal weighting matrix is the inverse variance-covariance matrix of the observations. In our problem, the variance-covariance matrix is found to have very special property. Specifically, in Section 3.7.4 we prove the following results.

Lemma 4. *Assume that f has a bounded second derivative and $E(\varepsilon^4)$ is finite. Then for any $b = n^{\vartheta}$ with $0 \le \vartheta < 1/2$, the variance-covariance matrix of $\{\hat{\sigma}^2_{Rt}(1), \ldots, \hat{\sigma}^2_{Rt}(b)\}$ has leading order $\Sigma = (\sigma_{pr})_{b\times b}$, where $\sigma_{pp} = (\gamma_4 - 1 + 1/m)\sigma^4/(mn)$ for any $1 \le p \le b$ and $\sigma_{rp} = \sigma_{pr} = (\gamma_4 - 1)\sigma^4/(mn)$ for any $1 \le r < p \le b$.*

Lemma 4 states that the leading order of the variance-covariance matrix has the same value on the diagonal, even though each diagonal element corresponds to a different lag. In addition, the off-diagonal elements are also identical, hence the matrix Σ is compound symmetric. These properties yield great simplification of GLS. Specifically, let $\boldsymbol{z} = \{\hat{\sigma}^2_{\mathrm{Rt}}(1), \ldots, \hat{\sigma}^2_{\mathrm{Rt}}(r)\}^T$, $\boldsymbol{\beta} = (\alpha, \beta)^T$, $\boldsymbol{e} = (e_1, \ldots, e_b)^T$, $\boldsymbol{d} = (d_1, \ldots, d_b)^T$, and $X = (\boldsymbol{u}, \boldsymbol{d})$ be the design matrix. With these notations, the linear model (3.9) is equivalent to $\boldsymbol{z} = X\boldsymbol{\beta} + \boldsymbol{e}$, and to the first order, the optimal GLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\boldsymbol{z}.$$

From Lemma 4, we have $\Sigma = (\gamma_4 - 1 + 1/m)\sigma^4\{(1-\rho)I + \rho\boldsymbol{u}\boldsymbol{u}^T\}/(mn)$, where $\rho = (\gamma_4 - 1)/(\gamma_4 - 1 + 1/m)$ and I is the identity matrix. Due to the compound symmetry structure of Σ and the fact that the first column of X is $\boldsymbol{u}$, it is not difficult to show that $(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\boldsymbol{z} = (X^TX)^{-1}X^T\boldsymbol{z}$ (McElroy (1967) and Kariya and Kurata (2004)). This implies that the optimal GLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ is in fact the same as the ordinary least squares estimator $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ to the first order. In

other words, the simplest OLS is already the most efficient way of perform the linear regression.

However, the sequencing method is not the optimal GLS or OLS. In fact, the estimator $\hat{\sigma}_3^2$ is a GLS with a special weighting strategy. Specifically, let $W = \text{diag}(w_1, \ldots, w_b)$ be the weight matrix and write the weighted least squares (WLS) estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\hat{\alpha}_{\text{WLS}}, \hat{\beta}_{\text{WLS}})^T = (X^T W^{-1} X)^{-1} X^T W^{-1} \boldsymbol{z}.$$

Then the sequencing estimator corresponds to the intercept estimation of $\hat{\boldsymbol{\beta}}_{\text{WLS}}$, i.e., $\hat{\sigma}_3^2 = \hat{\alpha}_{\text{WLS}}$. It is not difficult to see that when $n \to \infty$, the weights w_i converge to a constant uniformly. Thus, $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ is asymptotically the same as $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ and hence is also optimal. The reason we propose WLS instead of the simplest OLS to form the sequencing estimator is based on small sample consideration. When n is not too large, WLS takes into account the higher order difference of the variabilities at different lags hence it adapts better to the data and tends to have more stable numerical performance.

3.3 Simulation Studies

We now conduct simulation studies to evaluate the finite sample performance of the aforementioned estimators: the naive estimator $\hat{\sigma}_{\text{naive}}^2$, the sample variance estimator $\hat{\sigma}_1^2$, the partitioning estimator $\hat{\sigma}_2^2$, and the sequencing estimator $\hat{\sigma}_3^2$. For comparison, we also include a residual-based estimator, where we use the cubic smoothing spline to estimate the mean function and then use the squared residuals to estimate the variance. During the procedure, the smoothing parameter is selected via the generalized cross validation, and the resulting variance estimator is written as $\hat{\sigma}_{\text{SS}}^2$.

We consider the following two mean functions:

$$f_1(x) = 10x(1-x),$$
$$f_2(x) = 3x\sin(4\pi x),$$

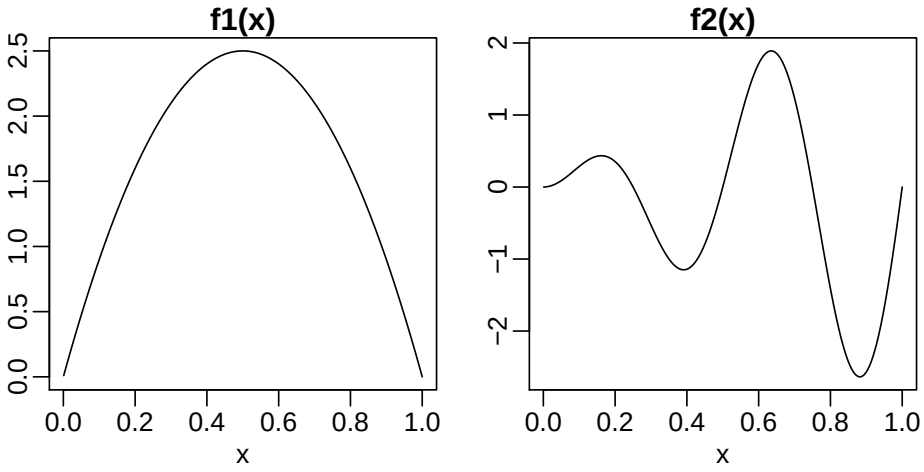

Figure 3.1: The mean functions $f_1(x)$ and $f_2(x)$, where $0 \leq x \leq 1$.

where f_1 is a low-frequency function and f_2 is an irregular high-frequency function (see Figure 3.1). The coefficients 10 in f_1 and 3 in f_2 are chosen so that the two mean functions have similar amplitudes. We set the design points $x_i = i/n$ and simulate ε_{ij} independently from $N(0, \sigma^2)$. For each mean function, we consider $n = 30$ and 200 corresponding to small and large sample sizes respectively, and $\sigma^2 = 0.25$ and 4 corresponding to small and large variances respectively. Further, we choose $m = 2, 3, 4, 5$ and 10 to represent different levels of repeated measurements. In total, we have 40 combinations of simulation settings.

We choose the bandwidths $b_s = n^{1/2}$ and $b_t = n^{1/3}$ for both $\hat{\sigma}_2^2$ and $\hat{\sigma}_3^2$. The corresponding estimators are referred to as $\hat{\sigma}_2^2(b_t)$, $\hat{\sigma}_2^2(b_s)$, $\hat{\sigma}_3^2(b_t)$ and $\hat{\sigma}_3^2(b_s)$ respectively. The CV method can also be used for estimating σ^2, and we find it generally performs as well as b_t and b_s. However, because CV is computationally more expensive, we do not recommend it and hence do not present its corresponding results in the remainder of the content. For $\hat{\sigma}_{\text{naive}}^2$, we use the bandwidth $b_t = n^{1/3}$ throughout the simulations. Also note that the quadratic matrix $\boldsymbol{D}$ is not guaranteed to be positive definite. This means that $\hat{\sigma}_{\text{naive}}^2$, $\hat{\sigma}_2^2(b_t)$, $\hat{\sigma}_2^2(b_s)$, $\hat{\sigma}_3^2(b_t)$ and $\hat{\sigma}_3^2(b_s)$ may take negative estimates, though it happens very rarely in our simulations. We replace negative estimates by zero in the calculation of the relative mean squared errors.

We repeat the simulation $10,000$ times for each setting. The relative mean squared

errors, $(mn)\text{MSE}/(2\sigma^4)$, are reported in Table 3.1 for $n = 30$ and in Table 3.2 for $n = 200$. Based on the simulation results, we summarize the findings below. (i) The sequencing estimator $\hat{\sigma}_3^2(b_s)$ or $\hat{\sigma}_3^2(b_t)$ exhibits the best performance in all but one setting; it even outperforms the residual-based estimator $\hat{\sigma}_{\text{SS}}^2$ when an appropriate bandwidth is used. (ii) The relative performance of $\hat{\sigma}_3^2(b_s)$ and $\hat{\sigma}_3^2(b_t)$ depends on the smoothness of f, the sample size n and the signal-to-noise ratio. In general, $\hat{\sigma}_3^2(b_s)$ performs slightly better than $\hat{\sigma}_3^2(b_t)$ for most settings; whereas for small n and rough f, $\hat{\sigma}_3^2(b_t)$ is much better than $\hat{\sigma}_3^2(b_s)$. (iii) The sequencing estimator always performs better than the partitioning estimator. Specifically, $\hat{\sigma}_3^2(b_s)$ always outperforms $\hat{\sigma}_2^2(b_s)$ and $\hat{\sigma}_3^2(b_t)$ always outperforms $\hat{\sigma}_2^2(b_t)$. (iv) The sample variance estimator $\hat{\sigma}_1^2$ does not suffer from the bias term caused by the lack of smoothness of f and the large signal-to-noise ratio. As a consequence, it outperforms all other methods when n is small (30), σ^2 is small (0.25) and f is rough (f_2). (v) The naive estimator $\hat{\sigma}_{\text{naive}}^2$ is always the worst among all the estimators. (vi) When m increases, all the proposed estimators, except the naive estimator, have a decreased relative MSE. In particular, the MSE of $\hat{\sigma}_1^2$ decreases dramatically as m increases. When $m = 10$, $\hat{\sigma}_1^2$ always performs well and is among the best of all the estimators. This demonstrates again the importance of extracting the information contained in the repeated measurements.

3.4 Nonparametric Regression with Unbalanced Repeated Measurements

In this section, we consider the nonparametric regression model with unbalanced repeated measurements,

$$Y_{ij} = f(x_i) + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m_i,$$

where Y_{ij}, x_i, f and ε_{ij} are defined as before. We assume that m_i are not all the same, where $m_i = 1$ represents a single observation at the ith design point.

f	σ^2	m	$\hat{\sigma}^2_{\text{naive}}$	$\hat{\sigma}^2_1$	$\hat{\sigma}^2_2(b_t)$	$\hat{\sigma}^2_2(b_s)$	$\hat{\sigma}^2_3(b_t)$	$\hat{\sigma}^2_3(b_s)$	$\hat{\sigma}^2_{\text{SS}}$
f_1	0.25	2	3.17	2.07	1.57	1.45	1.31	1.28	1.80
		3	4.92	1.55	1.61	1.50	1.24	1.25	1.58
		4	6.84	1.37	1.64	1.57	1.19	1.24	1.42
		5	8.67	1.25	1.61	1.58	1.15	1.22	1.31
		10	21.15	1.14	1.64	1.79	1.12	1.28	1.17
	4	2	3.09	2.07	1.55	1.33	1.28	1.18	1.53
		3	4.65	1.55	1.58	1.33	1.21	1.14	1.37
		4	6.29	1.37	1.61	1.35	1.16	1.11	1.26
		5	7.70	1.25	1.57	1.32	1.12	1.07	1.19
		10	15.83	1.14	1.59	1.34	1.08	1.07	1.14
f_2	0.25	2	8.79	2.07	2.98	16.42	2.18	11.08	2.69
		3	22.84	1.55	3.63	23.71	2.21	13.63	2.17
		4	49.36	1.37	4.30	31.03	2.29	16.28	1.90
		5	90.14	1.25	4.90	38.27	2.37	18.97	1.71
		10	658.93	1.14	8.02	74.46	2.98	33.04	1.42
	4	2	3.17	2.07	1.57	1.43	1.30	1.26	2.16
		3	4.78	1.55	1.60	1.46	1.23	1.22	1.77
		4	6.62	1.37	1.63	1.49	1.18	1.19	1.55
		5	8.21	1.25	1.60	1.51	1.14	1.17	1.43
		10	19.23	1.14	1.62	1.66	1.10	1.21	1.27

Table 3.1: Relative mean squared errors for the seven estimators under various settings with $n = 30$.

f	σ^2	m	$\hat{\sigma}^2_{\text{naive}}$	$\hat{\sigma}^2_1$	$\hat{\sigma}^2_2(b_t)$	$\hat{\sigma}^2_2(b_s)$	$\hat{\sigma}^2_3(b_t)$	$\hat{\sigma}^2_3(b_s)$	$\hat{\sigma}^2_{\text{SS}}$
f_1	0.25	2	2.55	2.02	1.29	1.12	1.15	1.07	1.13
		3	3.89	1.48	1.27	1.08	1.07	1.03	1.06
		4	5.08	1.31	1.27	1.11	1.06	1.03	1.06
		5	6.24	1.25	1.27	1.11	1.05	1.03	1.05
		10	12.86	1.10	1.24	1.08	1.01	1.01	1.02
	4	2	2.55	2.02	1.29	1.11	1.15	1.07	1.10
		3	3.89	1.48	1.27	1.08	1.07	1.02	1.05
		4	5.08	1.31	1.27	1.10	1.06	1.03	1.05
		5	6.23	1.25	1.27	1.10	1.05	1.03	1.04
		10	12.83	1.10	1.24	1.07	1.01	1.00	1.01
f_2	0.25	2	2.56	2.02	1.29	1.36	1.15	1.28	1.40
		3	3.90	1.48	1.27	1.45	1.08	1.31	1.25
		4	5.13	1.31	1.28	1.59	1.07	1.41	1.21
		5	6.30	1.25	1.27	1.70	1.06	1.48	1.17
		10	13.24	1.10	1.24	2.23	1.01	1.83	1.09
	4	2	2.55	2.02	1.28	1.11	1.15	1.07	1.23
		3	3.89	1.48	1.27	1.08	1.08	1.02	1.14
		4	5.08	1.31	1.27	1.11	1.06	1.03	1.12
		5	6.23	1.25	1.27	1.10	1.05	1.03	1.10
		10	12.83	1.10	1.24	1.08	1.01	1.00	1.05

Table 3.2: Relative mean squared errors for the seven estimators under various settings with $n = 200$.

3.4.1 Methodology

We first point out that when m_i's are not identical, the averaged observations $\bar{Y}_i$ no longer have homogeneous variances. Instead, $\text{Var}(\bar{Y}_i) = \sigma^2/m_i$. Consequently, the naive method is no longer applicable for unbalanced repeated measurements, simply because Tong and Wang (2005) does not apply to heterogeneous variances. For the other three proposed methods, we derive the following corresponding results. Their numerical performance will be studied in Section 3.4.2.

The sample variance method still yields a valid estimator. For the ith design point, let $s_i^2 = \sum_{j=1}^{m_i}(Y_{ij} - \bar{Y}_i)^2/(m_i - 1)$ when $m_i \geq 2$ and $s_i^2 = 0$ when $m_i = 1$. The sample variance estimator is then

$$\tilde{\sigma}_1^2 = \frac{1}{M-n}\sum_{i=1}^{n}(m_i - 1)s_i^2,$$

where $M = \sum_{i=1}^{n} m_i$. We note that $\tilde{\sigma}_1^2$ is an unbiased estimator for σ^2. In the special case when $m_i \geq 2$ and are identical, $\tilde{\sigma}_1^2$ reduces to $\hat{\sigma}_1^2$.

The partitioning method continues to work for unbalanced repeated measurements with slight modification. First, we sample one observation Y_{1g_1} from the set $\{Y_{11}, \ldots, Y_{1m_1}\}$, one observation Y_{2g_1} from the set $\{Y_{21}, \ldots, Y_{2m_2}\}$, $\ldots$, and one observation Y_{ng_1} from the set $\{Y_{n1}, \ldots, Y_{nm_n}\}$. Second, we apply Tong and Wang (2005)'s method on the selected group $G(1) = \{Y_{1g_1}, \ldots, Y_{ng_1}\}$ to get one estimate $\hat{\sigma}_{(1)}^2$. We then repeat the process B times and estimate σ^2 by

$$\tilde{\sigma}_2^2 = \frac{1}{B}\sum_{j=1}^{B}\hat{\sigma}_{(j)}^2.$$

Unlike the partitions in Section 3.2.2, the groups $G(1), \ldots, G(B)$ are not fully separated so that the estimators $\sigma_{(1)}^2, \ldots, \sigma_{(B)}^2$ may not be independent of each other. In the special case when m_i are all the same, it can be shown that $\tilde{\sigma}_2^2$ and $\hat{\sigma}_2^2$ are asymptotically equivalent as $B \to \infty$. In general, the larger the B value, the closer performance between $\hat{\sigma}_2^2$ and $\tilde{\sigma}_2^2$. We suggest to choose a B value that is at least larger than $\max\{m_1, \ldots, m_n\}$ in practice.

The sequencing method can also be adjusted to apply for unbalanced repeated measurements. Let $d_{i_1 i_2} = (x_{i_2} - x_{i_1})^2$ be the squared distances between design

points x_{i_2} and x_{i_1}, and $S_{i_1 i_2} = \{s_{i_1(j_1)i_2(j_2)} = (Y_{i_2 j_2} - Y_{i_1 j_1})^2/2 : j_1 = 1, \ldots, m_{i_1},\ j_2 = 1, \ldots, m_{i_2}\}$ be the set of size $m_{i_1} m_{i_2}$ for the half squared differences associated with $d_{i_1 i_2}$. We collect all $d_{i_1 i_2}$ values so that $d_{i_1 i_2} \leq (b/n)^2$, and let $A = \{(i_1, i_2) : d_{i_1 i_2} \leq (b/n)^2,\ 1 \leq i_1 \leq i_2 \leq n\}$. Correspondingly, we collect all the $s_{i_1(j_1)i_2(j_2)}$ values for each $(i_1, i_2) \in A$. Now for each paired data $\{(d_{i_1 i_2}, s_{i_1(j_1)i_2(j_2)}) : (i_1, i_2) \in A,\ j_1 = 1, \ldots, m_{i_1},\ j_2 = 1, \ldots, m_{i_2}\}$, we fit a simple regression model $s_{i_1(j_1)i_2(j_2)} = \alpha + d_{i_1 i_2}\beta + \eta_{i_1 i_2}$ by least squares and then estimate σ^2 by the fitted intercept,

$$\tilde{\sigma}_3^2 = \tilde{\alpha} = \frac{1}{NT_2 - T_1^2} \sum_{(i_1, i_2) \in A} (T_2 - T_1 d_{i_1 i_2}) \sum_{j_1=1}^{m_{i_1}} \sum_{j_2=1}^{m_{i_2}} s_{i_1(j_1)i_2(j_2)}.$$

where $N = \sum_A m_{i_1} m_{i_2}$, $T_1 = \sum_A m_{i_1} m_{i_2} d_{i_1 i_2}$ and $T_2 = \sum_A m_{i_1} m_{i_2} d_{i_1 i_2}^2$. When m_i are all the same, it is easy to verify that $\tilde{\sigma}_3^2$ is equivalent to a least squares estimator with

$$\tilde{s}_r = \frac{1}{2m^2(n-r)} \sum_{i=r+1}^{n} \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} (Y_{i,j_2} - Y_{i-r,j_1})^2$$

as the dependent variable and $\tilde{d}_r = r^2/n^2$ as the independent variable. Our analytical and simulation studies (not shown) indicate that under equally spaced and balanced design, $\tilde{\sigma}_3^2$ and $\hat{\sigma}_3^2$ are equivalent asymptotically and similar in finite sample performance. Note that $\tilde{\sigma}_3^2$ also works for unequally spaced designs. In view of this, we claim that $\tilde{\sigma}_3^2$ generalizes the sequencing estimator $\hat{\sigma}_3^2$ not only from balanced repeated measurements to unbalanced repeated measurements, but also from equally spaced designs to unequally spaced designs.

3.4.2 A Simulation Study

We now study the finite sample performance of the proposed estimators under the unbalanced repeated measurement setting. The estimators considered for comparison are $\tilde{\sigma}_1^2$, $\tilde{\sigma}_2^2(b_t)$, $\tilde{\sigma}_2^2(b_s)$, $\tilde{\sigma}_3^2(b_t)$, $\tilde{\sigma}_3^2(b_s)$, and $\hat{\sigma}_{\text{SS}}^2$, where B is set to be 50 for the estimator $\tilde{\sigma}_2^2$. We consider the mean functions $f_1(x)$ and $f_2(x)$, the sample sizes $n = 30$ and $n = 200$, and the residual variances $\sigma^2 = 0.25$ and 4 as in Section 3.3. The design points are $x_i = i/n$ and ε_{ij} are simulated independently from $N(0, \sigma^2)$. For the different

measurements repetitions, we set $m_i = r$ if $i = 5k + r$, where k is a non-negative integer and r is an integer in [1, 5]. In total, there are a total of $3n$ observations.

We repeat the simulation $10,000$ times for each setting, and report in Table 3.3 the relative mean squared errors, i.e., $(3n)\text{MSE}/(2\sigma^4)$. Based on the simulation results, we summarize the following findings. First, $\tilde{\sigma}_3^2(b_s)$ or $\tilde{\sigma}_3^2(b_t)$ performs the best in all but one settings, where f is rough (f_2), σ^2 is small (0.25) and n is small. In this case, $\tilde{\sigma}_1^2$ works the best. The comparative performance of $\tilde{\sigma}_3^2(b_s)$ and $\tilde{\sigma}_3^2(b_t)$ is similar to that of $\hat{\sigma}_3^2(b_s)$ and $\hat{\sigma}_3^2(b_t)$ in the balanced repeated measurement setting. Second, the sequencing estimator always outperforms the partitioning estimator, regardless what bandwidth is used. Third, the finite sample performance of the sequencing estimator is superior to that of the residual-based estimator in all settings except for the case $(n, f, \sigma^2) = (30, f_2, 0.25)$.

Finally, it is interesting to compare the simulation results in Table 3.3 with those for the setup $m = 3$ in Tables 3.1 and 3.2, where $m = 3$ reflects the average number of repeated measurements. (i) For the sample variance method, we note that $\tilde{\sigma}_1^2$ works as well as or even better than $\hat{\sigma}_1^2$ in most settings. This indicates that $\tilde{\sigma}_1^2$ successfully adapts to the unbalanced measurements by putting more weights (i.e., $m_i - 1$) on the more accurate sample variances obtained from large m_i values. (ii) The partitioning method is less efficient when the repeated measurement data are unbalanced. We believe that this is caused by the finite choice of B. (iii) For the sequencing method, we note that $\tilde{\sigma}_3^2$ and $\hat{\sigma}_3^2$ are comparable in most settings. This indicates that the pair-wise adjustment of the sequencing method successfully generalizes the methodology and works effectively for the unbalanced repeated measurement settings.

3.5 Real Application

The first data set was from a study of fetal size conducted in Chitty et al. (1993) and Royston and Altman (1994), where 167 ultrasonographic fetal measurements were made from the 12th week of gestation onwards with two variables: *age* as the gesta-

n	f	σ^2	$\tilde{\sigma}_1^2$	$\tilde{\sigma}_2^2(b_t)$	$\tilde{\sigma}_2^2(b_s)$	$\tilde{\sigma}_3^2(b_t)$	$\tilde{\sigma}_3^2(b_s)$	$\hat{\sigma}_{\text{SS}}^2$
30	f_1	0.25	1.46	1.79	1.81	1.25	1.21	1.49
		4	1.46	1.75	1.62	1.16	1.10	1.31
	f_2	0.25	1.46	3.78	24.22	4.98	16.75	2.06
		4	1.46	1.77	1.75	1.22	1.20	1.68
200	f_1	0.25	1.49	1.60	1.48	1.08	1.05	1.08
		4	1.49	1.58	1.46	1.09	1.04	1.07
	f_2	0.25	1.48	1.59	1.82	1.09	1.44	1.26
		4	1.48	1.58	1.46	1.09	1.04	1.15

Table 3.3: Relative mean squared errors for the six estimators under various settings with unbalanced repeated measurements.

tional age in weeks and *length* as the mandible length in mm. The data set is named as "Mandible" and can be downloaded from the R package "lmtest". In this study, we treat *age* as the independent variable and *length* as the response variable and plot their relationship in the left panel of Figure 3.2. The fitted curve through smoothing spline shows a nonlinear relationship between the two variables. In addition, we assume a homogeneous variance along with the regression given that the heterogeneity in the variation is not obvious. To estimate the residual variance, we apply the three proposed estimators together with the residual-based estimator. The resulting estimates are: $\tilde{\sigma}_1^2 = 6.04$, $\tilde{\sigma}_2^2(b_t) = 7.89$, $\tilde{\sigma}_2^2(b_s) = 8.39$, $\tilde{\sigma}_3^2(b_t) = 5.71$, $\tilde{\sigma}_3^2(b_s) = 5.48$ and $\hat{\sigma}_{\text{SS}}^2 = 5.51$, where $\tilde{\sigma}_2^2(b_t)$ and $\tilde{\sigma}_2^2(b_s)$ are computed using $B = 50$. We note that the sequencing estimates and the residual-based estimate are very similar, and the sample variance estimate is also not too different from them. However, the partitioning estimates $\tilde{\sigma}_2^2(b_t)$ and $\tilde{\sigma}_2^2(b_s)$ yield quite different values. This coincides with the results in Section 3.3 that the sequencing method is always more reliable than the partitioning method.

The second data set was reported by University of Oxford via the department of statistics consulting service (Venables and Ripley; 2002). The data were collected

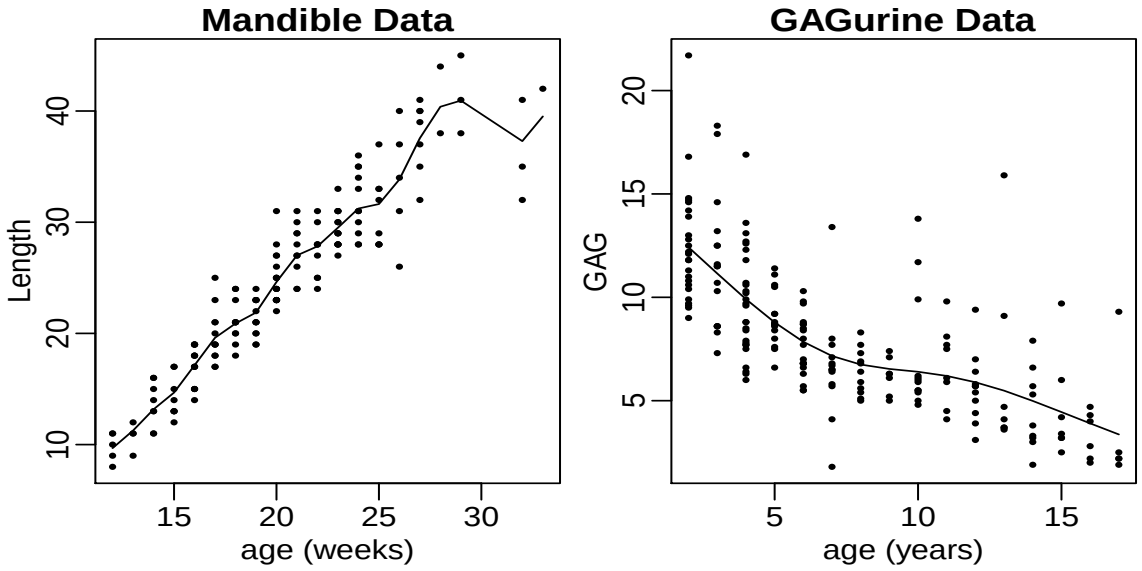

Figure 3.2: The Mandible data (left panel) and the GAGurine data (right panel) together with the fitted curves by smoothing spline.

on the concentration of a chemical GAG in the urine, and the aim of the study was to produce a chart to help a pediatrician to assess if a child's GAG concentration is normal or not. The data set is in the data frame "GAGurine" and it can be downloaded in the R package "MASS". The following two variables are included: *age* as the child age in years and *GAG* as the concentration of GAG. To estimate the residual variance, we use all the 167 children aged from 2 to 17 years. From the scatter plot and its fitted curve, we observe a slightly nonlinear pattern between the two variables. In addition, a constant variance assumption seems not unreasonable and therefore we adopt this assumption in this study. For the proposed methods and the residual-based method, the estimated variances are: $\tilde{\sigma}_1^2 = 5.89$, $\tilde{\sigma}_2^2(b_t) = 5.10$, $\tilde{\sigma}_2^2(b_s) = 5.41$, $\tilde{\sigma}_3^2(b_t) = 5.87$, $\tilde{\sigma}_3^2(b_s) = 5.97$, and $\hat{\sigma}_{\text{SS}}^2 = 5.57$, where $\tilde{\sigma}_2^2(b_t)$ and $\tilde{\sigma}_2^2(b_s)$ are computed using $B = 50$. Overall, we note that there is not large discrepancy among these estimates. Since for small n and large m, our simulation indicates that the sample variance estimator usually performs the best, we can evaluate the performance of other estimators by inspecting the difference from $\tilde{\sigma}_1^2$. To this end, the two sequencing estimators are again the winner against other methods including the residual-based estimator.

3.6 Conclusion

We have proposed three difference-based methods for estimating the residual variance in nonparametric regression with repeated measurements: the sample variance method by using only the variation within design points, the partitioning method by using only the variation between design points, and the sequencing method by using both between and within variations. We have investigated the statistical properties of the proposed estimators for fixed m and large n and have established the optimality of the sequencing estimator. When n is fixed while m is large, it is easily seen that the sample variance estimator is an efficient estimator and is recommended in practice. We further conducted extensive simulation studies to assess the finite sample performance, and applied two real data examples to demonstrate their application. In terms of implementation specifics, for large n, we recommend the sequencing estimator with the bandwidth $b_t = n^{1/3}$ when f is rough; otherwise, we recommend the sequencing estimator with bandwidth $b_s = n^{1/2}$. For small n, we recommend the sample variance estimator when f is rough or when m is large; otherwise we recommend the sequencing estimator with the bandwidth $b_t = n^{1/3}$. We have also extended the proposed difference-based methods to handle unbalanced repeated measurement settings and found them work well in practice. Further work might be needed to investigate the statistical properties under the unbalanced design.

Finally, we mention that the difference-based methods have been extended to more general settings, e.g., to multivariate covariates models (Hall et al.; 1991; Kulasekera and Gallagher; 2002; Munk et al.; 2005; Bock et al.; 2007; Liitiäinen et al.; 2010) and to semiparametric regression models (Xu and You; 2007; Wang et al.; 2011). Note also that a constant variance assumption may not be realistic in practice and the difference-based methods have been applied to the variance function estimation in the literature (Müller and Stadtmüller; 1993; Levine; 2006; Brown and Levine; 2007; Cai et al.; 2009). Further research is warranted in these directions when repeated measurements are presented.

3.7 Proofs

This section provides technical proofs for Theorem 4-7.

3.7.1 Proof of Theorem 4

Assume that the linear function is $f(x) = \mu + \delta x$. For ease of notation, denote $f_i = f(x_i)$, $i = 1, \ldots, n$. Then

$$\begin{aligned} E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] &= \sigma^2 + \frac{\delta^2}{2c_r}\left\{\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}\frac{(r-1)^2}{n^2} + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}\frac{r^2}{n^2}\right\} \\ &= \sigma^2 + \frac{\delta^2}{2c_r n^2}\left[(n-r+1)(r-1)^2\sum_{k=1}^{m-1}(m-k) + (n-r)r^2\sum_{k=1}^{m}k\right] \\ &= \sigma^2 + \frac{1}{2}d_r\delta^2. \end{aligned}$$

Note that $\sum_{r=1}^{b} w_r = 1$ and $\bar{d}_w = \sum_{r=1}^{b} w_r d_r$. We have

$$E(\bar{\sigma}^2_w) = \sum_{r=1}^{b} w_r E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] = \sigma^2 + \frac{1}{2}\delta^2\bar{d}_w. \tag{3.18}$$

Further, we have

$$E(\hat{\beta}) = \frac{\sum_{r=1}^{b} w_r(d_r - \bar{d}_w)E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right]}{\sum_{r=1}^{b} w_r(d_r - \bar{d}_w)^2} = \frac{(\delta^2/2)\left(\sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2\right)}{\sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2} = \frac{1}{2}\delta^2, \tag{3.19}$$

where $\sum_{r=1}^{b} w_r(d_r - \bar{d}_w)^2 = \sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2$ and

$$\begin{aligned} \sum_{r=1}^{b} w_r(d_r - \bar{d}_w)E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] &= \sum_{r=1}^{b} w_r d_r E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] - \bar{d}_w E(\bar{\sigma}^2_w) \\ &= \sigma^2\bar{d}_w + \frac{1}{2}\delta^2\sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w\left(\sigma^2 + \frac{1}{2}\delta^2\bar{d}_w\right) \\ &= \frac{1}{2}\delta^2\left(\sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2\right). \end{aligned}$$

Finally, by (3.18) and (3.19), we have $E(\hat{\sigma}^2_3) = E(\bar{\sigma}^2_w) - E(\hat{\beta})\bar{d}_w = \sigma^2$. This finishes the proof.

3.7.2 Proof of Theorem 5

Proof of (3.13): Instead of using the formula $\text{Bias}(\hat{\sigma}_3^2) = \boldsymbol{f}^T D \boldsymbol{f}/\text{tr}(D)$, we calculate this quantity directly from (3.10) which gives a more accurate approximation. For ease of notation, denote $f_i = f(x_i)$, $f_i' = f'(x_i)$, and $f_i'' = f''(x_i)$, $i = 1, \ldots, n$. Similarly as Section 3.7.1, we have

$$\begin{aligned} E\left[\hat{\sigma}_{\text{Rt}}^2(r)\right] &= \sigma^2 + \frac{m}{2c_r}\left[(m-1)\sum_{i=r}^{n}(f_i - f_{i-r+1})^2 + (m+1)\sum_{i=r+1}^{n}(f_i - f_{i-r})^2\right] \\ &= \sigma^2 + \frac{m}{2c_r}\left\{(m-1)\sum_{i=r}^{n}\left[\frac{(r-1)^2}{n^2}(f_i')^2 + O(\frac{(r-1)^3}{n^3})\right]\right. \\ &\quad \left. +(m+1)\sum_{i=r+1}^{n}\left[\frac{r^2}{n^2}(f_i')^2 + O(\frac{r^3}{n^3})\right]\right\} \\ &= \sigma^2 + Jd_r + O(\frac{r^3}{n^3}). \end{aligned}$$

Consequently, we have $E(\bar{\sigma}_w^2) = \sum_{r=1}^{b} w_r E\left[\hat{\sigma}_{\text{Rt}}^2(r)\right] = \sigma^2 + J\bar{d}_w + O(b^3/n^3)$. In addition, it is easy to verify that

$$\bar{d}_w = \frac{1}{s_b}\sum_{r=1}^{b} c_r d_r = \frac{b^2}{3n^2} - \frac{b^3}{12n^3} - \frac{(m-1)b}{2mn^2} + o(\frac{b^3}{n^3}) + o(\frac{b}{n^2}), \tag{3.20}$$

and

$$\sum_{r=1}^{b} w_r (d_r - \bar{d}_w)^2 = \sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2 = \frac{4b^4}{45n^4} + o(\frac{b^4}{n^4}). \tag{3.21}$$

This leads to

$$\begin{aligned} \sum_{r=1}^{b} w_r (d_r - \bar{d}_w) E\left[\hat{\sigma}_{\text{Rt}}^2(r)\right] &= \sum_{r=1}^{b} w_r d_r \left[\sigma^2 + Jd_r + O(\frac{r^3}{n^3})\right] - \bar{d}_w\left[\sigma^2 + J\bar{d}_w + O(\frac{b^3}{n^3})\right] \\ &= J\left(\sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2\right) + O(\frac{b^5}{n^5}). \end{aligned}$$

Finally, we have

$$\begin{aligned} E(\hat{\sigma}_3^2) &= E(\bar{\sigma}_w^2) - \frac{\bar{d}_w}{\sum_{r=1}^{b} w_r (d_r - \bar{d}_w)^2}\sum_{r=1}^{b} w_r (d_r - \bar{d}_w) E\left[\hat{\sigma}_{\text{Rt}}^2(r)\right] \\ &= \left[\sigma^2 + J\bar{d}_w + O(\frac{b^3}{n^3})\right] - \left[J\bar{d}_w + O(\frac{b^3}{n^3})\right] \\ &= \sigma^2 + O(\frac{b^3}{n^3}). \end{aligned}$$

To achieve the variance of $\hat{\sigma}_3^2$, we need Lemmas 5 and 6.

Lemma 5. *For the equally spaced design with $b \to \infty$ and $b/n \to 0$, we have*

(i) $\sum_{r=1}^{b} \tau_r = b - \frac{5b^2}{16n} + o(\frac{b^2}{n}) + o(1)$.

(ii) $\sum_{r=1}^{b} \tau_r^2 = \frac{9}{4}b + o(b)$.

(iii) $\sum_{r=1}^{b} r\tau_r = \frac{3}{16}b^2 + o(b^2)$.

(iv) $\sum_{r=1}^{b} r^2\tau_r = o(b^3)$.

(v) $\sum_{r=1}^{i} \tau_r = \frac{9}{4}i - \frac{5i^3}{4b^2} + o(i) + o(\frac{i^3}{b^2})$, $1 \le i \le b$.

(vi) $\sum_{r=1}^{i} r\tau_r = \frac{9}{8}i^2 - \frac{15i^4}{16b^2} + o(i^2) + o(\frac{i^4}{b^2})$, $1 \le i \le b$.

Proof. (i) Let $\eta = \bar{d}_w / \sum_{r=1}^{b} w_r(d_r - \bar{d}_w)^2$. Then $\tau_r = 1 - \eta(d_r - \bar{d}_w)$. By (3.20) and (3.21), we have $\eta = 15n^2/(4b^2) + o(n^2/b^2)$ and $\sum_{r=1}^{b}(d_r - \bar{d}_w) = \sum_{r=1}^{b} d_r - b\bar{d}_w = b^4/(12n^3) + o(b^4/n^3) + o(b^2/n^2)$. This leads to

$$\sum_{r=1}^{b} \tau_r = b - \eta \sum_{r=1}^{b}(d_r - \bar{d}_w) = b - \frac{5b^2}{16n} + o(\frac{b^2}{n}) + o(1).$$

(ii) By $\sum_{r=1}^{b} d_r = b^3/(3n^2) + o(b^3/n^2)$ and $\sum_{r=1}^{b} d_r^2 = b^5/(5n^4) + o(b^5/n^4)$, we have $\sum_{r=1}^{b}(d_r - \bar{d}_w)^2 = \sum_{r=1}^{b} d_r^2 - 2\bar{d}_w \sum_{r=1}^{b} d_r + b\bar{d}_w^2 = 4b^5/(45n^4) + o(b^5/n^4)$. This leads to

$$\sum_{r=1}^{b} \tau_r^2 = b - 2\eta \sum_{r=1}^{b}(d_r - \bar{d}_w) + \eta^2 \sum_{r=1}^{b}(d_r - \bar{d}_w)^2 = \frac{9}{4}b + o(b).$$

(iii) Note that $\sum_{r=1}^{b} rd_r = b^4/(4n^2) + o(b^4/n^2)$. We have

$$\sum_{r=1}^{b} r\tau_r = (1 + \eta\bar{d}_w) \sum_{r=1}^{b} r - \eta \sum_{r=1}^{b} rd_r = \frac{3}{16}b^2 + o(b^2).$$

(iv) Note that $\sum_{r=1}^{b} r^2 d_r = b^5/(5n^2) + o(b^5/n^2)$. We have

$$\sum_{r=1}^{b} r^2 \tau_r = (1+\eta\bar{d}_w)\sum_{r=1}^{b} r^2 - \eta\sum_{r=1}^{b} r^2 d_r = o(b^3).$$

(v) Note that $\sum_{r=1}^{i} d_r = i^3/(3n^2) + o(i^3/n^2)$. For any $1 \le i \le b$, we have

$$\sum_{r=1}^{i} \tau_r = (1+\eta\bar{d}_w)i - \eta\sum_{r=1}^{i} d_r = \frac{9}{4}i - \frac{5i^3}{4b^2} + o(i) + o(\frac{i^3}{b^2}).$$

(vi) Note that $\sum_{r=1}^{i} r d_r = i^4/(4n^2) + o(i^4/n^2)$. For any $1 \le i \le b$, we have

$$\sum_{r=1}^{i} r\tau_r = (1+\eta\bar{d}_w)\sum_{r=1}^{i} r - \eta\sum_{r=1}^{i} r d_r = \frac{9}{8}i^2 - \frac{15i^4}{16b^2} + o(i^2) + o(\frac{i^4}{b^2}).$$

Lemma 6. *Under the same conditions as in Theorem 5, we have*

(i) $\boldsymbol{f}^T D^2 \boldsymbol{f} = O(\frac{b^5}{n^2}) + O(\frac{b^2}{n})$.

(ii) $\boldsymbol{f}^T [D \cdot \mathrm{diag}(D)\boldsymbol{u}] = O(\frac{b^4}{n}) + O(b^2)$.

(iii) $\mathrm{tr}[\mathrm{diag}(D)^2] = 4m^3 n b^2 - \frac{103m^3}{28} b^3 + o(b^3)$.

(iv) $\mathrm{tr}(D^2) = 4m^3 n b^2 - \frac{103m^3}{28} b^3 + \frac{9m^2}{2} nb + o(b^3) + o(nb)$.

Proof. (i) Noting that the matrix D is symmetric, we have

$$\boldsymbol{f}^T D^2 \boldsymbol{f} = \boldsymbol{f}^T D^T D \boldsymbol{f} = (D\boldsymbol{f})^T D\boldsymbol{f} = \boldsymbol{\xi}^T \boldsymbol{\xi},$$

where $\boldsymbol{\xi} = D\boldsymbol{f} = (\xi_1, \ldots, \xi_{nm})^T$. Let $l = (i-1)m + j$, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Note that f has a bounded second derivative. When $i \in [b+1, n-b]$,

by Lemma 5 (i), (iii) and (iv), we have

$$\begin{aligned}
\xi_l &= (j-1)\sum_{r=1}^{b}\tau_r\left[\frac{r-1}{n}f_i' - \frac{(r-1)^2}{2n^2}f_i'' + o(\frac{r^2}{n^2})\right] \\
&\quad +(m-j+1)\sum_{r=1}^{b}\tau_r\left[\frac{r}{n}f_i' - \frac{r^2}{2n^2}f_i'' + o(\frac{r^2}{n^2})\right] \\
&\quad -(m-j)\sum_{r=1}^{b}\tau_r\left[\frac{r-1}{n}f_i' + \frac{(r-1)^2}{2n^2}f_i'' + o(\frac{r^2}{n^2})\right] \\
&\quad -j\sum_{r=1}^{b}\tau_r\left[\frac{r}{n}f_i' + \frac{r^2}{2n^2}f_i'' + o(\frac{r^2}{n^2})\right] \\
&= \frac{m-2j+1}{n}f_i'\sum_{r=1}^{b}\tau_r - \frac{1}{2n^2}f_i''\sum_{r=1}^{b}\tau_r[2mr^2 - 2(m-1)r + (m-1)] + o(\frac{b^3}{n^2}) \\
&= O(\frac{b}{n}) + o(\frac{b^3}{n^2}).
\end{aligned}$$

When $i \in [1, b]$, by Lemma 5 (i), (iii), (v) and (vi), we have

$$\begin{aligned}
\xi_l &= (j-1)\sum_{r=1}^{i}\tau_r\left[\frac{r-1}{n}f_i' + o(\frac{r^2}{n^2})\right] + (m-j+1)\sum_{r=0}^{i-1}\tau_r\left[\frac{r}{n}f_i' + o(\frac{r^2}{n^2})\right] \\
&\quad -(m-j)\sum_{r=1}^{b}\tau_r\left[\frac{r-1}{n}f_i' + o(\frac{r^2}{n^2})\right] - j\sum_{r=1}^{b}\tau_r\left[\frac{r}{n}f_i' + o(\frac{r^2}{n^2})\right] \\
&= O(\frac{b^2}{n}).
\end{aligned}$$

When $i \in [n-b+1, n]$, similar argument leads to $\xi_l = O(b^2/n)$. Finally,

$$\boldsymbol{f}^T D^2 \boldsymbol{f} = \boldsymbol{\xi}^T\boldsymbol{\xi} = \sum_{l=1}^{mb}\xi_l^2 + \sum_{l=mb+1}^{m(n-b)}\xi_l^2 + \sum_{l=m(n-b)+1}^{nm}\xi_l^2 = O(\frac{b^5}{n^2}) + O(\frac{b^2}{n}).$$

(ii) Note that $\boldsymbol{f}^T[D\cdot\mathrm{diag}(D)\boldsymbol{u}] = \boldsymbol{\xi}^T\cdot\mathrm{diag}(D)\boldsymbol{u}$. By part (i) and Lemma 5 (i) and (v), we have

$$\begin{aligned}
\boldsymbol{f}^T[D\cdot\mathrm{diag}(D)\boldsymbol{u}] &= \sum_{i=1}^{b}\sum_{j=1}^{m}\xi_{(i-1)m+j}\left[m\sum_{r=1}^{b}\tau_r + m\sum_{r=0}^{i-1}\tau_r + (j-1)\tau_i\right] \\
&\quad + \sum_{l=mb+1}^{m(n-b)}\xi_l\left(2m\sum_{r=1}^{b}\tau_r\right) \\
&\quad + \sum_{i=n-b+1}^{n}\sum_{j=1}^{m}\xi_{(i-1)m+j}\left[m\sum_{r=1}^{b}\tau_r + m\sum_{r=0}^{n-i}\tau_r + (m-j)\tau_i\right] \\
&= O(\frac{b^4}{n}) + O(b^2).
\end{aligned}$$

(iii) By Lemma 5 (i) and (v), we have

$$\begin{aligned}
\mathrm{tr}[\mathrm{diag}(D)^2] &= \sum_{l=mb+1}^{m(n-b)} \left(2m\sum_{r=1}^{b}\tau_r\right)^2 + 2\sum_{i=1}^{b}\sum_{j=1}^{m}\left[m\sum_{r=1}^{b}\tau_r + m\sum_{r=0}^{i-1}\tau_r + (j-1)\tau_i\right]^2 \\
&= 4m^3(n-2b)\left(b - \frac{5b^2}{16n} + o(\frac{b^2}{n}) + o(1)\right)^2 + 2m^3\sum_{i=1}^{b}\left[b + \frac{9}{4}i - \frac{5i^3}{4b^2} + o(b)\right]^2 \\
&= 4m^3nb^2 - \frac{103m^3}{28}b^3 + o(b^3).
\end{aligned}$$

(iv) By part (iii) and Lemma 5 (ii), we have

$$\begin{aligned}
\mathrm{tr}(D^2) &= \sum_{l=mb+1}^{m(n-b)}\left[\left(2m\sum_{r=1}^{b}\tau_r\right)^2 + 2m\sum_{r=1}^{b}\tau_r^2\right] \\
&\quad +2\sum_{i=1}^{b}\sum_{j=1}^{m}\left[\left(m\sum_{r=1}^{b}\tau_r + m\sum_{r=0}^{i-1}\tau_r + (j-1)\tau_i\right)^2 + m\sum_{r=1}^{b}\tau_r^2 + m\sum_{r=0}^{i-1}\tau_r^2 + (j-1)\tau_i^2\right] \\
&= \mathrm{tr}[\mathrm{diag}(D)^2] + 2m^2(n-2b)\sum_{r=1}^{b}\tau_r^2 + 2\sum_{i=1}^{b}\sum_{j=1}^{m}\left(m\sum_{r=1}^{b}\tau_r^2 + m\sum_{r=0}^{i-1}\tau_r^2 + (j-1)\tau_i^2\right) \\
&= \left[4m^3nb^2 - \frac{103m^3}{28}b^3 + o(b^3)\right] + 2m^2(n-2b)\left[\frac{9}{4}b + o(b)\right] + O(b^2) \\
&= 4m^3nb^2 - \frac{103m^3}{28}b^3 + \frac{9m^2}{2}nb + o(b^3) + o(nb).
\end{aligned}$$

Proof of (3.14): Note that the last four terms in (3.12) make up the variance of $\hat{\sigma}_3^2$. Note that $\sigma^4(\gamma_4 - 3) = \mathrm{Var}(\varepsilon^2) - 2\sigma^4$, $\mathrm{tr}(D) = 2s_b$, and $s_b = m^2nb - m^2b^2/2 + o(b^2)$. By Lemmas 5 and 6, we have

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}_3^2) &= \frac{1}{[\mathrm{tr}(D)]^2}\left\{4\sigma^2\boldsymbol{f}^TD^2\boldsymbol{f} + 4\boldsymbol{f}^T[D\cdot\mathrm{diag}(D)\boldsymbol{u}]\sigma^3\gamma_3\right. \\
&\quad \left.+\sigma^4(\gamma_4-3)\mathrm{tr}[\mathrm{diag}(D)^2] + 2\sigma^4\mathrm{tr}(D^2)\right\} \\
&= \frac{1}{4s_b^2}\left\{O(\frac{b^5}{n^2}) + O(\frac{b^2}{n}) + O(\frac{b^4}{n}) + O(b^2) + \sigma^4(\gamma_4-3)\left[4m^3nb^2 - \frac{103m^3}{28}b^3 + o(b^3)\right]\right. \\
&\quad \left.+2\sigma^4\left[4m^3nb^2 - \frac{103m^3}{28}b^3 + \frac{9m^2}{2}nb + o(b^3) + o(nb)\right]\right\} \\
&= \frac{1}{4s_b^2}\left[\left(4m^3nb^2 - \frac{103m^3}{28}b^3\right)\mathrm{Var}(\varepsilon^2) + 9m^2nb\sigma^4 + o(b^3) + o(nb)\right] \\
&= \frac{1}{mn}\mathrm{Var}(\varepsilon^2) + \frac{9}{4m^2nb}\sigma^4 + \frac{9b}{112mn^2}\mathrm{Var}(\varepsilon^2) + o(\frac{1}{nb}) + o(\frac{b}{n^2}).
\end{aligned}$$

Proof of (3.15): The MSE in (3.15) is an immediate result from (3.13) and (3.14).

3.7.3 Proof of Theorem 6

Let $f_i = f(x_i)$, $i = 1, \ldots, n$. To prove the asymptotic normality for $\hat{\sigma}^2_{\mathrm{Rt}}(r)$, we first partition it into three parts, $\hat{\sigma}^2_{\mathrm{Rt}}(r) = L_1 + L_2 + L_3$, where

$$L_1 = \frac{1}{2c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})^2 + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})^2\right],$$

$$L_2 = \frac{1}{c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k}) + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})\right],$$

$$L_3 = \frac{1}{2c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k})^2 + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})^2\right].$$

(i) Note that $J = O(1)$ and $d_r = O(r^2/n^2)$. For L_1, by Taylor series we have $L_1 = Jd_r + o(r^2/n^2) = O(r^2/n^2)$. This shows that $L_1 = o(n^{-1/2})$ when $r = n^{\vartheta}$ with $0 \le \vartheta < 3/4$.

(ii) For L_2, by Cauchy-Schwarz inequality we have

$$\begin{aligned}
L_2^2 &\le \frac{2}{c_r^2}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k})\right]^2 \\
&\quad + \frac{2}{c_r^2}\left[\sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})\right]^2 \\
&\le \frac{2}{c_r^2}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})^2\right]\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k})^2\right] \\
&\quad + \frac{2}{c_r^2}\left[\sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})^2\right]\left[\sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})^2\right].
\end{aligned}$$

This leads to

$$E(L_2^2) \le \frac{4\sigma^2}{c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})^2 + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})^2\right] = O(\frac{r^2}{n^2}).$$

This shows that $L_2 = o_p(n^{-1/2})$ for any $r = n^{\vartheta}$ with $0 \le \vartheta < 1/2$.

(iii) We represent the term L_3 as $L_3 = \sigma^2 + \sum_{i=r+1}^{n}\zeta_i(r)/(n-r) + O(1/n)$, where

$$\zeta_i(r) = \frac{1}{2m^2}\left[\sum_{k=1}^{m-1}\sum_{j=k+1}^{m}(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k})^2 + \sum_{k=1}^{m}\sum_{j=1}^{k}(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})^2\right] - \sigma^2. \quad (3.22)$$

We have $E(\zeta_i(r)) = 0$. Treat $\{\zeta_i(r), i = r+1, \ldots, n\}$ as a stochastic process. With some straightforward algebra, we have (a) for $r = 1$,

$$\mathrm{Cov}(\zeta_i(r), \zeta_l(r)) = \begin{cases} [(8m^2 - 3m + 1)\gamma_4 - (8m^2 - 15m + 1)]\sigma^4/(12m^3), & l - i = 0, \\ (4m^2 + 3m - 1)(\gamma_4 - 1)\sigma^4/(24m^3), & l - i = 1, \\ 0, & l - i \geq 2; \end{cases}$$

(b) for $r = 2$,

$$\mathrm{Cov}(\zeta_i(r), \zeta_l(r)) = \begin{cases} [(5m^2 + 1)\gamma_4 - (5m^2 - 12m + 1)]\sigma^4/(12m^3), & l - i = 0, \\ (4m^2 - 3m - 1)(\gamma_4 - 1)\sigma^4/(24m^3), & l - i = 1, \\ (m + 1)(\gamma_4 - 1)\sigma^4/(8m^2), & l - i = 2, \\ 0, & l - i \geq 3; \end{cases}$$

and (c) for any $r \geq 3$,

$$\mathrm{Cov}(\zeta_i(r), \zeta_l(r)) = \begin{cases} [(5m^2 + 1)\gamma_4 - (5m^2 - 12m + 1)]\sigma^4/(12m^3), & l - i = 0, \\ (m^2 - 1)(\gamma_4 - 1)\sigma^4/(24m^3), & l - i = 1, \\ (m - 1)(\gamma_4 - 1)\sigma^4/(8m^2), & l - i = r - 1, \\ (m + 1)(\gamma_4 - 1)\sigma^4/(8m^2), & l - i = r, \\ 0, & \text{otherwise.} \end{cases}$$

Note that the above covariances depend on i and l only through the difference $l - i$, regardless of the choice of r. This shows that for any given $r \geq 1$, $\{\zeta_i(r), i = r+1, \ldots, n\}$ is a strictly stationary sequence of random variables with mean zero and autocovariance function $C(\tau) = C(s, s+\tau) = \mathrm{Cov}(\zeta_s(r), \zeta_{s+\tau}(r))$. Also note that $\{\zeta_i(r), i = r+1, \ldots, n\}$ is an m-dependent sequence with $m = r$. Thus by Brockwell and Davis (1991), we have the following asymptotic normality for L_3,

$$\sqrt{n}(L_3 - \sigma^2) \xrightarrow{\mathcal{D}} N(0, \nu_r^2) \qquad \text{as } n \to \infty, \tag{3.23}$$

where $\nu_r^2 = C(0) + 2\sum_{\tau=1}^{r} C(\tau)$. For the covariance functions in (a)-(c), it is easy to verify that $\nu_1^2 = \nu_2^2 = \nu_r^2 = [\gamma_4/m - (m-1)/m^2]\sigma^4$ for any $r \geq 3$.

Finally, noting that $\hat{\sigma}^2_{\mathrm{Rt}}(r) = L_1 + L_2 + L_3 = L_3 + o_p(n^{-1/2})$, by (3.23) and Slutsky's theorem we have for any $r = n^{\vartheta}$ with $0 \leq \vartheta < 1/2$, $\sqrt{n}(\hat{\sigma}^2_{\mathrm{Rt}}(r) - \sigma^2) \xrightarrow{\mathcal{D}} N(0, [\gamma_4/m - (m-1)/m^2]\sigma^4)$ as $n \to \infty$.

3.7.4 Proof of Theorem 7

To prove Theorem 7, we need the following lemma which was originated from Whittle (1964).

Lemma 7. *Assume that the matrix* $A = (a_{ij})_{n\times n}$ *satisfies* $a_{ij} = a_{i-j}$ *and* $\sum_{-\infty}^{\infty} a_k^2 < \infty$. *Also assume that* $E(\varepsilon^2) = \sigma^2$ *and* $E(\varepsilon^{4+2\delta})$ *is finite for some* δ *in* $(0, 1)$. *Then*

$$\frac{1}{n}\boldsymbol{\varepsilon}^T A\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i-j}\varepsilon_i\varepsilon_j \xrightarrow{\mathcal{D}} N(a_0\sigma^2, \sigma_A^2), \quad \text{as } n \to \infty,$$

where $\sigma_A^2 = (\gamma_4 - 3)a_0^2\sigma^4/n + 2\sigma^4\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i-j}^2/n^2$.

Proof of Theorem 7:

By $\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$ and $\mathrm{tr}(D) = 2s_b$, we have

$$\hat{\sigma}_3^2 = \frac{1}{2s_b}\boldsymbol{f}^T D\boldsymbol{f} + \frac{1}{s_b}\boldsymbol{f}^T D\boldsymbol{\varepsilon} + \frac{1}{2s_b}\boldsymbol{\varepsilon}^T D\boldsymbol{\varepsilon}. \tag{3.24}$$

(i) For the first term in (3.24), noting that it corresponds to the bias term $E(\hat{\sigma}_3^2)$, By Theorem 5 we have $\boldsymbol{f}^T D\boldsymbol{f}/(2s_b) = O(b^3/n^3) = o(n^{-1/2})$ for any $b = n^{\vartheta}$ with $0 < \vartheta < 5/6$.

(ii) For the second term in (3.24), by Lemma 6 and the fact $s_b = O(nb)$ we have

$$E\left(\boldsymbol{f}^T D\boldsymbol{\varepsilon}/s_b\right)^2 = (\boldsymbol{f}^T D^2\boldsymbol{f})\sigma^2/s_b^2 = O(b^3/n^4) + O(1/n^3).$$

This implies that $\boldsymbol{f}^T D\boldsymbol{\varepsilon}/s_b = o_p(n^{-1/2})$ for any $b = o(n)$.

(iii) Now we derive the asymptotic normality for the last term in (3.24). Let $(mn/2s_b)D = C - H$, where $C = (c_{ij})_{n\times n}$ is an $(mn) \times (mn)$ matrix with elements

$$c_{ij} = \begin{cases} m^2 n\sum_{r=1}^{b}\tau_r/s_b, & 1 \le i = j \le mn, \\ -mn\tau_a/(2s_b), & (a-1)m < |i-j| \le am \text{ with } a = 1, \ldots, b, \\ 0, & \text{otherwise}, \end{cases}$$

and $H = \mathrm{diag}(h_1, h_2, \ldots, h_{mn})$ is an $(mn) \times (mn)$ diagonal matrix with elements $h_i = \{m^2 n\sum_{r=1}^{b}\tau_r - m^2 n\sum_{r=0}^{a-1}\tau_r - mn[i-1-(a-1)m]\tau_a\}/(2s_b)$ for $(a-1)m < i \le am$

with $a = 1, \ldots, b$; $h_i = 0$ for $(a-1)m < i \le am$ with $a = b+1, \ldots, n-b$; and $h_i = \{m^2 n \sum_{r=1}^{b} \tau_r - m^2 n \sum_{r=0}^{n-a} \tau_r - mn(am-i)\tau_{n+1-a}\}/(2s_b)$ for $(a-1)m < i \le am$ with $a = n-b+1, \ldots, n$. Then

$$\frac{1}{2s_b}\boldsymbol{\varepsilon}^T D\boldsymbol{\varepsilon} = \frac{1}{mn}\boldsymbol{\varepsilon}^T C\boldsymbol{\varepsilon} - \frac{1}{mn}\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}. \tag{3.25}$$

For the symmetric matrix C, let $c_{ij} = c_{i-j}$ with $c_0 = m^2 n \sum_{r=1}^{b} \tau_r / s_b$; $c_{i-j} = c_{j-i} = -mn\tau_a/(2s_b)$ for $(a-1)m < |i-j| \le am$ with $a = 1, \ldots, b$; and $c_{i-j} = c_{j-i} = 0$ for $|i-j| > bm$. By Lemma 5, for any $b = n^{\vartheta}$ with $0 < \vartheta < 1$,

$$\sum_{-\infty}^{\infty} c_k^2 = c_0^2 + 2\sum_{k=1}^{bm} c_k^2 = \frac{m^4 n^2}{s_b^2}\left(\sum_{r=1}^{b} \tau_a\right)^2 + \frac{m^3 n^2}{2s_b^2}\sum_{a=1}^{b} \tau_a^2 = O(1) < \infty.$$

Now given that $\mathrm{E}(\varepsilon^{4+2\delta})$ is finite for some δ in $(0,1)$, by Lemma 7 we have

$$\sqrt{mn}\left(\frac{1}{mn}\boldsymbol{\varepsilon}^T C\boldsymbol{\varepsilon} - c_0\sigma^2\right) \xrightarrow{\mathcal{D}} N(0, \sigma_c^2), \quad \text{as } n \to \infty,$$

where $\sigma_c^2 = (\gamma_4 - 3)\sigma^4 c_0^2 + 2\sigma^4 \sum_{i=1}^{mn}\sum_{j=1}^{mn} c_{i-j}^2/(mn)$. For the second term in (3.25), by Lemma 5 it is easy to verify that for any $b = n^{\vartheta}$ with $0 < \vartheta < 1/2$, $\mathrm{E}(\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}/mn)^2 = O(b^2/n^2)$ and further $\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}/(mn) = o_p(n^{-1/2})$.

By (i)-(iii) and Slutsky's theorem, we have for any $b = n^{\vartheta}$ with $0 < \vartheta < 1/2$,

$$\frac{\sqrt{mn}(\hat{\sigma}_3^2 - c_0\sigma^2)}{\sigma_c} \xrightarrow{\mathcal{D}} N(0,1), \quad \text{as } n \to \infty. \tag{3.26}$$

Note that $c_0 = m^2 n \sum_{r=1}^{b} \tau_r / s_b = 1 + O(b/n)$. This leads to $\sqrt{mn}(c_0 - 1) = o(1)$ for any $b = n^{\vartheta}$ with $0 \le \vartheta < 1/2$. In addition, it is easy to verify that

$$\sigma_c^2 = \frac{m^4 n^2 (\gamma_4 - 1)\sigma^4}{s_b^2}\left(\sum_{r=1}^{b} \tau_r\right)^2 + \frac{m^3 n^2 \sigma^4}{s_b^2}\sum_{r=1}^{b} \tau_r^2 = (\gamma_4 - 1)\sigma^4 + o(1).$$

This leads to $(\gamma_4 - 1)\sigma^4/\sigma_c^2 \to 1$ as $n \to \infty$. Finally, by (3.26) and Slutsky's theorem,

$$\begin{aligned} \frac{\sqrt{mn}(\hat{\sigma}_3^2 - \sigma^2)}{\sqrt{(\gamma_4 - 1)\sigma^4}} &= \frac{\sigma_c}{\sqrt{(\gamma_4 - 1)\sigma^4}}\left\{\frac{\sqrt{mn}(\hat{\sigma}_3^2 - c_0\sigma^2)}{\sigma_c} + \frac{\sqrt{mn}(c_0 - 1)\sigma^2}{\sigma_c}\right\} \\ &\xrightarrow{\mathcal{D}} N(0,1), \quad \text{as } n \to \infty, \end{aligned}$$

for any $b = n^{\vartheta}$ with $0 < \vartheta < 1/2$.

Proof of Lemma 4:

For any $1 \leq r < p \leq b$, by Section 3.7.3 it is easy to verify that

$$\mathrm{Cov}(\hat{\sigma}^2_{\mathrm{Rt}}(r), \hat{\sigma}^2_{\mathrm{Rt}}(p)) = \frac{1}{(n-r)(n-p)} \sum_{i=r+1}^{n} \sum_{l=p+1}^{n} \mathrm{Cov}(\zeta_i(r), \zeta_l(p)) + o(\frac{1}{n}), \quad (3.27)$$

where $\zeta_i(r)$ is defined in (3.22).

Let $Q = \sum_{i=r+1}^{n} \sum_{l=p+1}^{n} \mathrm{Cov}(\zeta_i(r), \zeta_l(p))$. When $r \geq 2$ and $p \geq r+2$, the term Q can be calculated as follows,

$$\begin{aligned}
Q &= \sum_{i=r+p+1}^{n} \mathrm{Cov}(\zeta_i(r), \zeta_{i-r}(p)) + \sum_{i=r+p}^{n} \mathrm{Cov}(\zeta_i(r), \zeta_{i-r+1}(p)) + \sum_{i=p+1}^{n} \mathrm{Cov}(\zeta_i(r), \zeta_i(p)) \\
&+ \sum_{i=r+2}^{n-p+r+1} \mathrm{Cov}(\zeta_i(r), \zeta_{i+p-r-1}(p)) + \sum_{i=r+1}^{n-p+r} \mathrm{Cov}(\zeta_i(r), \zeta_{i+p-r}(p)) \\
&+ \sum_{i=r+1}^{n-p+r-1} \mathrm{Cov}(\zeta_i(r), \zeta_{i+p-r+1}(p)) + \sum_{i=r+1}^{n-p+1} \mathrm{Cov}(\zeta_i(r), \zeta_{i+p-1}(p)) + \sum_{i=r+1}^{n-p} \mathrm{Cov}(\zeta_i(r), \zeta_{i+p}(p)) \\
&= (\gamma_4 - 1)\sigma^4 \left[\sum_{i=r+p+1}^{n} \frac{m+1}{8m^2} + \sum_{i=r+p}^{n} \frac{m-1}{8m^2} + \sum_{i=p+1}^{n} \frac{1}{4m} + \sum_{i=r+2}^{n-p+r+1} \frac{m^2-1}{24m^3} \right. \\
&\qquad \left. + \sum_{i=r+1}^{n-p+r} \frac{2m^2+1}{12m^3} + \sum_{i=r+1}^{n-p+r-1} \frac{m^2-1}{24m^3} + \sum_{i=r+1}^{n-p+1} \frac{m-1}{8m^2} + \sum_{i=r+1}^{n-p} \frac{m+1}{8m^2} \right] \\
&= \frac{1}{2m}(2n-2p-r)(\gamma_4 - 1)\sigma^4 + O(1). \qquad (3.28)
\end{aligned}$$

Similarly, we can verify that $Q = (2n-2p-r)/(2m) + O(1)$ holds for $r \geq 2$ and/or $p = r+1$. We omit their derivations here for saving space. Plugging (3.28) into (3.27) leads to

$$\begin{aligned}
\mathrm{Cov}(\hat{\sigma}^2_{\mathrm{Rt}}(r), \hat{\sigma}^2_{\mathrm{Rt}}(p)) &= \frac{2n-2p-r}{2m(n-r)(n-p)}(\gamma_4 - 1)\sigma^4 + o(\frac{1}{n}) \\
&= \frac{1}{mn}(\gamma_4 - 1)\sigma^4 + o(\frac{1}{n}).
\end{aligned}$$

Finally, we note that $\mathrm{Var}(\hat{\sigma}^2_{\mathrm{Rt}}(p)) = [\gamma_4 - 1 + 1/m]\sigma^4/(mn) + o(1/n)$ for any $1 \leq p \leq b$ is an immediate result from Theorem 3.

Chapter 4

Pairwise Method for Variance Estimation

4.1 Introduction

The aforementioned methods are all derived under the general assumption that the regression function $g(x)$ is a smooth function. In practical applications, however, such a smoothness assumption may be too restrictive if the data include some change points, e.g., the Nile river discharge data (Cobb; 1978), the stock market return data (Wang; 1995), the sea-level pressure data (Qiu and Yandell; 1998), the infants growth data (Müller and Stadtmüller; 1999), and the annual temperature data (Gijbels and Goderniaux; 2004). In this chapter, we consider the estimation of the residual variance when jumps appear in the mean function.

Consider a nonparametric regression model with jump discontinuities,

$$y_i = g(x_i) + h(x_i) + \varepsilon_i, \qquad 1 \le i \le n, \tag{4.1}$$

where y_i are observations, g is a continuous function, h is a step function, and ε_i are i.i.d. random errors with zero mean and variance σ^2. To be specific, we write the step function h as

$$h(x) = \sum_{j=1}^{p} c_j I(x > t_j),$$

where p is the number of jumps, $I(\cdot)$ is the identify function with value 1 when $x > t_j$ and value 0 otherwise, and c_j are the magnitudes of jumps at the jump points $t_j \in (0, 1)$, respectively. Note that $g + h$ is the mean function.

Model (4.1) has wide applications in statistical process control (Qiu and Hawkins; 2001), piecewise linear regression (Hinkley; 1969; Brown et al.; 1975; Kim and Siegmund; 1989), image processing (McDonald and Owen; 1986; Hall and Titterington; 1992; Qiu; 2005), and other related areas. There is an abundant literature for analyzing model (4.1) including the detection and estimation of the number, positions, and magnitudes of jump points (Müller; 1992; Wu and Chu; 1993a,b; Eubank and Speckman; 1994; Loader; 1996).

Needless to say, an accurate estimate of σ^2 is very important in regression models with jump discontinuities. Usually, one applies a two-step procedure to estimate σ^2 in such models. The first step is to estimate the positions of change points and then divide the mean function into several continuous sections accordingly. The second step is to estimate the residual variance within each individual section and then use them to make a final estimate of σ^2. Note that one may apply the residual-based methods (Hall and Marron; 1990) or apply the difference-based methods (Müller; 1992; Wu and Chu; 1993a) to estimate the residual variance within each individual section.

Apart from the above, Müller and Stadtmüller (1999) proposed a single-step method for estimating σ^2 in model (4.1). Consider the equally spaced design where $x_i = i/n$, $i = 1, \ldots, n$. Let

$$z_k = \sum_{i=1}^{n-L} (y_{i+k} - y_i)^2 / [2(n - L)],$$

where $k = 1, \ldots, L$ with $L = L(n) \geq 1$. Under certain conditions on the mean function and the bandwidth L, Müller and Stadtmüller showed that

$$E(z_k) \approx \sigma^2 + \gamma l_k + \delta l_k^2, \tag{4.2}$$

where $l_k = k/(n - L)$, $\gamma = \sum_{j=1}^{p-1} (c_{j+1} - c_j)^2/2$ is the amount of discontinuity in the data, and $\delta = \int_0^1 [g'(x)]^2 dx/2 + \sum_{j=1}^{p-1} g'(t_{j+1})(c_{j+1} - c_j)$ is the measurement of

the interaction between continuous and discontinuous parts. By (4.2), they fitted a quadratic regression that regresses z_k on l_k and then estimate the residual variance as the intercept. Specifically, they estimated σ^2 by

$$\hat{\sigma}^2_{\text{MS}} = \frac{3\sum_{k=1}^{L}(3L^2+3L+2-6(2L+1)k+10k^2)z_k}{2L(L-1)(L-2)}. \tag{4.3}$$

This method does not require an estimate of the positions of change points and is popular in practice.

Note that z_k only uses the first $n-L$ pairs of observations for performing the quadratic regression. Ignoring the last $L-k$ terms can make z_k a less efficient representation for σ^2, especially when $L-k$ is large. In addition, Müller and Stadtmüller (1999) required that $\min_{1\le i\le p-1}(t_{i+1}-t_i) \ge 2L/N$ for the possibility of change-points separation. In the special case when $\gamma = 0$, i.e., when $h(x) = 0$, Tong et al. (2013) have demonstrated that the least squares estimator in Tong and Wang (2005) provides a smaller mean squared error (MSE) than $\hat{\sigma}^2_{\text{MS}}$. In addition, the equally spaced design condition in Müller and Stadtmüller (1999) is somewhat strong and has limited the practical use of $\hat{\sigma}^2_{\text{MS}}$.

In this chapter, we propose a pairwise regression method for estimating σ^2 in model (4.1). Specifically, we regress the squared difference between observations on the squared distance between design points, and then estimate the residual variance as the intercept. Our method generalizes the existing methods from the following perspectives: (i) it does not require to estimate the positions of change points compared to the two-step estimators in the literature; (ii) it does not require to estimate the discontinuity parameter γ compared to the single-step estimator in Müller and Stadtmüller (1999); and (iii) it also applies to the settings where the design points are unequally spaced.

The remainder of the chapter is organized as follows. In Section 4.2.1, we review the difference-based methods in estimating the residual variance in continuous nonparametric regression. In Section 4.2.2, we propose a pairwise regression method that extends the least squares estimator in Tong and Wang (2005) to unequally spaced designs. In Section 4.2.3, we further extend the proposed pairwise regression method

to adaptively estimate the residual variance in nonparametric regression with jump discontinuities. In Section 4.3, we conduct extensive simulation studies to evaluate the finite-sample performance of the proposed method with some existing competitors. We then apply the proposed method to a real data example in Section 4.4 and make some discussions in Section 4.5.

4.2 Main Results

4.2.1 Difference-based Estimators

In the special case when $h(x) = 0$, model (4.1) reduces to

$$y_i = g(x_i) + \varepsilon_i, \qquad 1 \le i \le n. \tag{4.4}$$

Under model (4.4), there are many difference-based methods in the literature for estimating σ^2. Assume that $0 \le x_1 \le \cdots \le x_n \le 1$. von Neumann (1941) and Rice (1984) proposed a first order difference-based estimator,

$$\hat{\sigma}^2_{\mathrm{R}} = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2.$$

Gasser et al. (1986) and Hall et al. (1990) extended the idea and proposed some higher order difference-based estimators. In addition, Müller et al. (2003), Tong et al. (2008) and Du and Schick (2009) proposed covariate-matched U-statistic estimators for the residual variance.

Apart form them, Tong and Wang (2005) and Tong et al. (2013) proposed a variation of the difference-based estimator in nonparametric regression. Let $x_i = i/n$ and $s_k = \sum_{i=k+1}^{n} (y_i - y_{i-k})^2/[2(n-k)]$. Suppose that g has a bounded first derivative. Tong and Wang (2005) showed that for any fixed $m = o(n)$,

$$E(s_k) \approx \sigma^2 + d_k J, \qquad k = 1, \ldots, m, \tag{4.5}$$

where $d_k = k^2/n^2$ and $J = \int_0^1 [g'(x)]^2 dx/2$. By (4.5), they regressed s_k on d_k and then estimated the residual variance as the intercept. Specifically, their least squares

estimator is given as

$$\hat{\sigma}^2_{\text{TW}} = \sum_{k=1}^{m} w_k s_k - \hat{\beta}\bar{d}_w, \tag{4.6}$$

where $N_1 = mn - m(m+1)/2$, $w_k = (n-k)/N_1$, $\bar{d}_w = \sum_{k=1}^{m} w_k s_k$, and $\hat{\beta} = \sum_{k=1}^{m} w_k s_k (d_k - \bar{d}_w) / \sum_{k=1}^{m} w_k (d_k - \bar{d}_w)^2$.

Recall that $\hat{\sigma}^2_{\text{TW}}$ is developed under model (4.4) with a continuous mean function. When $h(x) \neq 0$, $\hat{\sigma}^2_{\text{TW}}$ may not perform well in model (4.1). To illustrate this, we consider the following regression model with a single jump at $t = 0.5$,

$$y_i = g(x_i) + cI(x_i > 0.5) + \varepsilon_i, \qquad c > 0. \tag{4.7}$$

Assume that J and c are both finite values. We have

$$\begin{aligned} E(s_k) &= \sigma^2 + \frac{1}{2(n-k)} \sum_{i=k+1}^{n} \{[g(x_i) + cI(x_i > 0.5)] \\ &\quad -[g(x_{i-k}) + cI(x_{i-k} > 0.5)]\}^2 \\ &= \sigma^2 + \left[d_k J + o(\frac{k^2}{n^2}) \right] + \left[\frac{k}{n} c^2 + o(\frac{k}{n}) \right]. \end{aligned}$$

Note that the bias owing to the jump, $(k/n)c^2$, dominates the bias owing to the continuous function, $d_k J = (k/n)^2 J$. This implies that $\hat{\sigma}^2_{\text{TW}}$ may suffer a severe bias for estimating σ^2, especially when c is large.

For a visualization of the bias pattern along with the c value, consider $g(x) = 5x(1-x)$ and $h(x) = cI(x > 0.5)$ with $0 < c < 20$. We let $n = 100$, $m = 10$ and $\sigma^2 = 1$ throughout the simulations. The estimated variance against the c value is plotted in Figure 4.1. We observe that $\hat{\sigma}^2_{\text{TW}}$ increases rapidly as c increases. As a consequence, $\hat{\sigma}^2_{\text{TW}}$ does not provide a satisfactory performance in this example.

4.2.2 Pairwise Regression

Recall that the least squares estimator in Tong and Wang (2005) only applies to the equally spaced design. This has largely restricted the usage of their method in practice. In this section, we introduce a pairwise regression method for estimating the residual variance that extends the least squares estimator from the equally spaced design to unequally spaced designs.

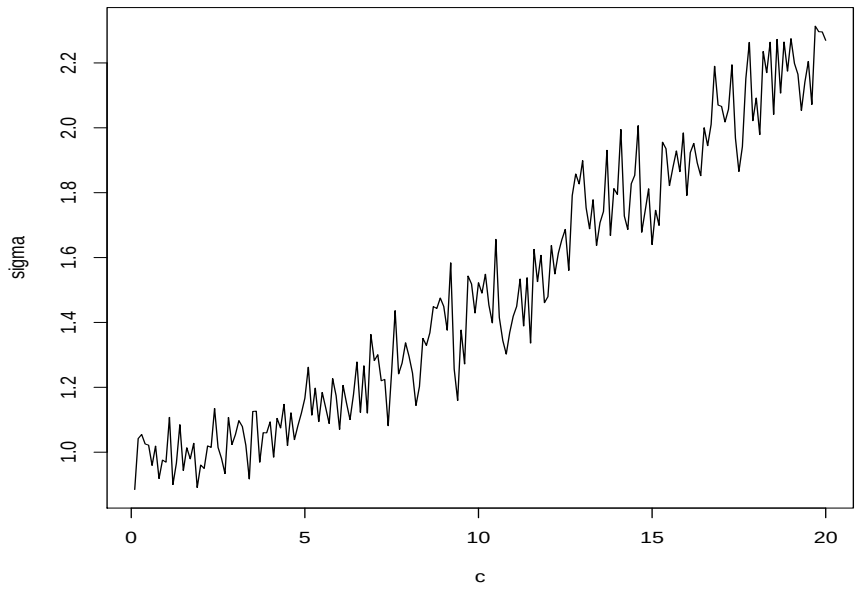

Figure 4.1: The estimated $\hat{\sigma}^2$ corresponding to different c values.

Let $s_{ij} = (y_j - y_i)^2/2$ be the half squared differences and $d_{ij} = (x_j - x_i)^2$ be the corresponding squared distances for any $1 \leq i < j \leq n$. Let $d = o(1)$ be the bandwidth. We collect all d_{ij} values that satisfy $d_{ij} \leq d$.

For ease of notation, let $A = \{(i,j) : d_{ij} \leq d,\ 1 \leq i < j \leq n\}$ and $N = \#(A)$ be the total number of pairs in A. Correspondingly, we collect the s_{ij} values for all $(i,j) \in A$.

Note that $E(s_{ij}) = \sigma^2 + (g(x_j) - g(x_i))^2/2$. When g is a linear function with slop ψ, we have $s_{ij} = \sigma^2 + d_{ij}\psi^2/2$. In view of this, for the paired data $\{(d_{ij}, s_{ij}) : (i,j) \in A\}$ with $d = o(1)$, we fit a simple linear regression model that regresses s_{ij} directly on d_{ij}

$$s_{ij} = \alpha + d_{ij}\beta + \eta_{ij}. \tag{4.8}$$

We then use the ordinary least-squares method to estimate σ^2 using the fitted intercept. This leads to

$$\hat{\sigma}^2 = \hat{\alpha} = \frac{\sum_A (S_2 - S_1 d_{ij}) s_{ij}}{N S_2 - S_1^2} \tag{4.9}$$

where $S_1 = \sum_A d_{ij}$ and $S_2 = \sum_A d_{ij}^2$. We refer to (4.9) as a pairwise regression estimator.

Let $c_{ij} = S_2 - S_1 d_{ij}$ and $y = (y_1, \ldots, y_n)^T$. The estimator (4.9) has a quadratic form $\hat{\sigma}^2 = y^T My/\mathrm{tr}(M)$, where M is an $n \times n$ symmetric matrix with upper triangular elements

$$m_{ij} = \begin{cases} \sum_{(i,j)\in A_k} c_{ij}/2 & 1 \le i = j = k \le n \\ -c_{ij}/2 & (i,j) \in A \\ 0 & \text{otherwise} \end{cases}$$

where $A_k = \{(i,j) : i = k \text{ or } j = k,\ (i,j) \in A\}$ with $k = 1, 2, \ldots, n$.

In what follows we draw some connection between the pairwise regression estimator $\hat{\sigma}^2$ and the least squares estimator $\hat{\sigma}^2_{\mathrm{TW}}$. Let $x_i = i/n$ and $d = m^2/n^2$. Then $N = N_1 = mn - m(m+1)/2$ and $d_{ij} = d_{j-i}$. Also, it is easy to verify that $S_1 = N\bar{d}_w$, $S_2 = N\sum_{k=1}^m w_k d_k^2$, $\sum_A s_{ij} = N\sum_{k=1}^m w_k s_k$, and $\sum_A d_{ij} s_{ij} = N\sum_{k=1}^m w_k d_k s_k$. With the above equalities, we have

$$\begin{aligned} \hat{\sigma}^2 &= \frac{S_2 \sum_A s_{ij} - S_1 \sum_A d_{ij} s_{ij}}{N S_2 - S_1^2} \\ &= \frac{\sum_{k=1}^m w_k d_k^2 \sum_{k=1}^m w_k s_k - \bar{d}_w \sum_{k=1}^m w_k d_k s_k}{\sum_{k=1}^m w_k d_k^2 - \bar{d}_w^2} \\ &= \sum_{k=1}^m w_k s_k - \bar{d}_w \hat{\beta} \\ &= \hat{\sigma}^2_{\mathrm{TW}}. \end{aligned}$$

This shows that when the design points are equally spaced, $\hat{\sigma}^2$ and $\hat{\sigma}^2_{\mathrm{TW}}$ are equivalent to each other. From this point of view, we conclude that the pairwise regression estimator (4.9) generalized the least squares estimator $\hat{\sigma}^2_{\mathrm{TW}}$ from the equally spaced design to a general design.

4.2.3 Adaptive Pairwise Regression

As mentioned, most existing difference-based estimators were developed under model (4.4). In this section, we show that the pairwise regression method in Section 4.2.2 can be readily extended to model (4.1) with jump discontinuities.

To apply the pairwise regression to models with jump discontinuities, we revisit the simple regression model presented in (4.7). Let $O = \{(i,j) : 0.5 \in (x_i, x_j]\}$ be the pairs of design points that cross the jump point. By (4.9), we have

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{\sum_{A\setminus O}(S_2 - S_1 d_{ij})E(s_{ij})}{NS_2 - S_1^2} + \frac{\sum_O (S_2 - S_1 d_{ij})E(s_{ij})}{NS_2 - S_1^2} \\ &= \frac{\sum_{A\setminus O}(S_2 - S_1 d_{ij})(\sigma^2 + O(m^2/n^2))}{NS_2 - S_1^2} + \frac{\sum_O (S_2 - S_1 d_{ij})(\sigma^2 + c^2/2 + O(m/n))}{NS_2 - S_1^2} \\ &= \frac{\sum_A (S_2 - S_1 d_{ij})}{NS_2 - S_1^2}\sigma^2 + \frac{\sum_O (S_2 - S_1 d_{ij})}{NS_2 - S_1^2}\left(\frac{c^2}{2} + O(\frac{m}{n})\right) \\ &= \sigma^2 + \frac{\sum_O (S_2 - S_1 d_{ij})}{NS_2 - S_1^2}\frac{c^2}{2} + O(\frac{m^2}{n^2}), \end{aligned} \tag{4.10}$$

where

$$\frac{\sum_A (S_2 - S_1 d_{ij})}{NS_2 - S_1^2} = 1 \qquad \text{and} \qquad \frac{\sum_O (S_2 - S_1 d_{ij})}{NS_2 - S_1^2} = O(\frac{m}{n}).$$

By (4.10), to obtain a good estimate of σ^2, it is clear that the pairs in O should be excluded from the regression to eliminate the bias. Otherwise, given that the quantity c is large, the extra bias introduced by the jump can be very severe.

In what follows, we examine how excluding the pairs in O takes effect on the MSE of the estimator. We will also suggest ways to exclude certain pairs of data from the pairwise regression. Let $z_{ij} = y_j - y_i$ for any $1 \le i < j \le n$. For $d = o(1)$, we have $E(z_{ij}) \to 0$ for $(i,j) \in O$ and $E(z_{ij}) \to c$ for $(i,j) \in A\setminus O$. Whereas for any $(i,j) \in A$, $\text{var}(z_{ij}) = 2\sigma^2$. To visualize the discrepancy between the two groups of z_{ij}, we consider $c = 0$, 2 and 5 for the example in Section 4.2.1. All other settings are kept the same as before except that now $\sigma = 0.5$.

We plot the histograms of the simulated z_{ij} values in the first column of Figure 4.2, respectively. When the mean function is continuous (i.e., $c = 0$), the histogram is unimodal and almost symmetric around zero. When c increases, the histogram tends to be right-skewed and eventually separates to two disjoint sections, one consisting of the pairs without jump and the other consisting of the pairs with jump. To eliminate the impact of the jump on the variance estimation, we can treat the extremely large $|z_{ij}|$ values, or correspondingly the extremely large s_{ij} values, as outliers and exclude them in the pairwise regression.

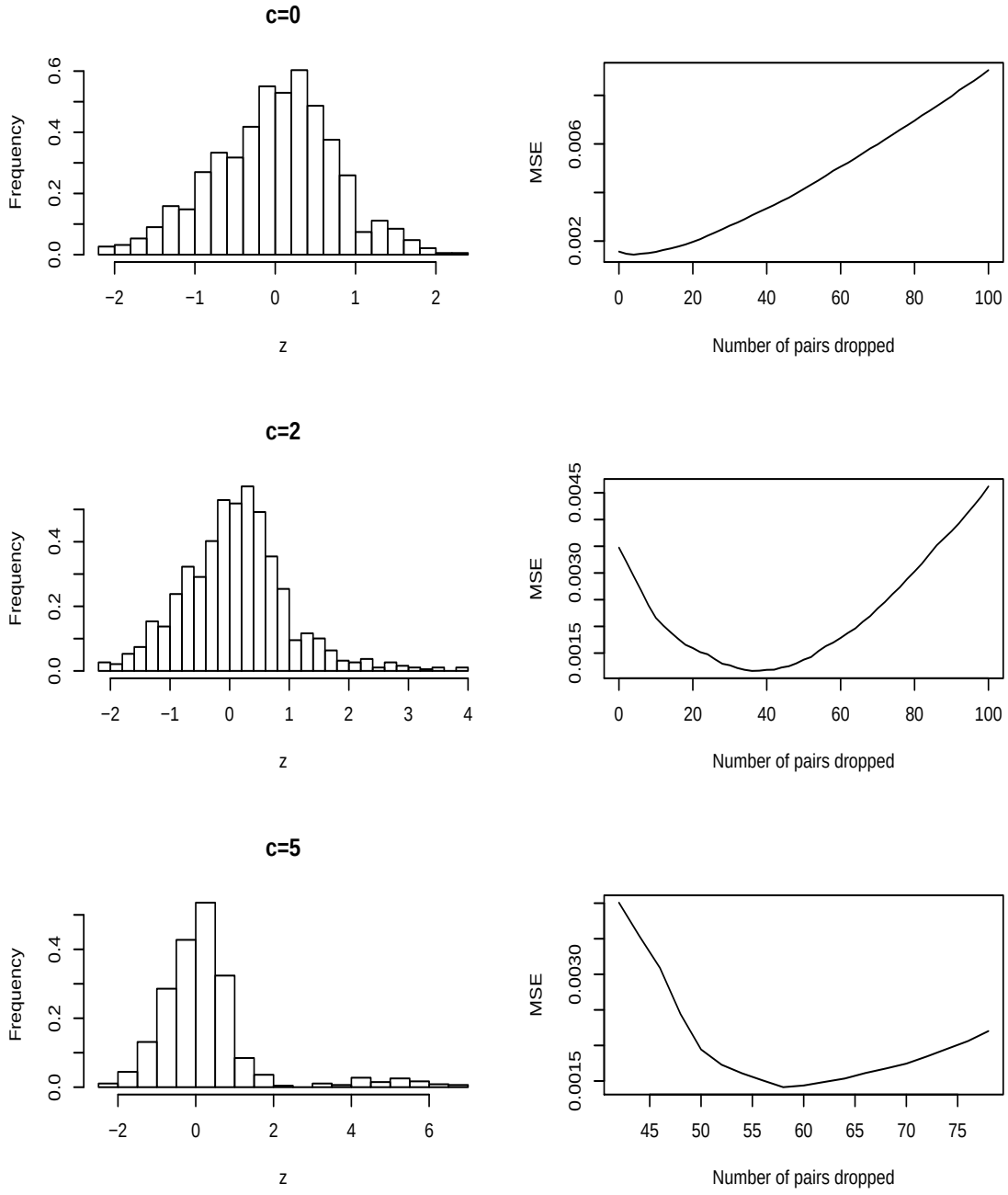

Figure 4.2: The histogram of z_{ij} and the change of MSE against the number of pairs dropped, where $c = 0$, 2 and 5, respectively.

Ideally, none of the z_{ij} values should be detected as outliers when the mean function is continuous. When c is non zero, to reduce the bias or essentially to gain a small MSE we may wish to drop the pairs that across the jump point. As an illustration, we also plot in Figure 4.2 the simulated MSE against the number of pairs dropped for the three c values, respectively. It suggests to drop few pairs for $c = 0$, drop around 40 pairs for $c = 2$, and drop around 55 pairs for $c = 5$, for estimating the residual variance with a minimum MSE. Finally, it is interesting to point out that for an equally spaced design with $m = 10$, there is a total of $m(m + 1)/2 = 55$ pairs across the jump point.

In what follows, we suggest two practical rules that identify certain z_{ij} values as outliers and then exclude them from the pairwise regression. The resulting methods are referred to as adaptive pairwise regression estimators.

Box Plot Method

The first method uses the box plot to detect certain z_{ij} values as outliers. Let $\mathrm{Q_L}(\{z_{ij}\})$ and $\mathrm{Q_U}(\{z_{ij}\})$ denote the lower quartile and the upper quartile of the observed z_{ij} values within the bandwidth, respectively. Follow the form of Sim et al. (1994), we define $\mathrm{LB} = \mathrm{Q_L}(\{z_{ij}\}) - C \cdot \mathrm{IQR}$ and $\mathrm{UB} = \mathrm{Q_U}(\{z_{ij}\}) + C \cdot \mathrm{IQR}$, where $\mathrm{IQR} = \mathrm{Q_U}(\{z_{ij}\}) - \mathrm{Q_L}(\{z_{ij}\})$ is the interquartile range and C is an adjustment factor. Here, we assign a value of 2 or 3 to C. We then identify z_{ij} as an outlier if $z_{ij} \in (-\infty, \mathrm{LB})$ or $z_{ij} \in (\mathrm{UB}, \infty)$. We refer to the estimator by the box plot method as $\hat{\sigma}^2_{\mathrm{box}}$.

Cross-Validation Method

Note that the bandwidth d is also critical to the variance estimation. Our second method uses a V-fold cross-validation (CV) approach to simultaneously choose the bandwidth d and the adjustment factor C. Specifically, we first split the whole data set into V disjoint subsamples, S_1, ..., S_V as in Tong and Wang (2005). Second, for given d and C, we estimate σ^2 by $\hat{\sigma}^2_v(d, C)$ based on the subsample $\cup_{i \neq v} S_i$ and the pairs with $d_{ij} \leq d$ and $z_{ij} \in [\mathrm{LB}(C), \mathrm{UB}(C)]$. Finally, we choose the optimal tuning

parameters d and C that minimize

$$\mathrm{CV}(d,C)=\sum_{v=1}^{V}\left[\hat{\sigma}^2(d,C)-\hat{\sigma}_v^2(d,C)\right]^2,$$

where $\hat{\sigma}^2(d,C)$ is the estimate of σ^2 based on the whole data set with pairs $d_{ij}\le d$ and $z_{ij}\in[\mathrm{LB}(C),\mathrm{UB}(C)]$. We refer to the estimator by the CV method as $\hat{\sigma}^2_{\mathrm{CV}}$.

4.3 Simulation Studies

In this section, we conduct extensive simulation studies to evaluate the finite-sample performance of the proposed estimators and compare them with some existing competitors.

4.3.1 Equidistant Design

The first study assumes an equally spaced design. Specifically, let $x_i=i/n$ with $i=1,\ldots,n$. We consider the following four estimators for comparison: $\hat{\sigma}^2_{\mathrm{box}}$, $\hat{\sigma}^2_{\mathrm{CV}}$, $\hat{\sigma}^2_{\mathrm{MS}}$ and $\hat{\sigma}^2_{\mathrm{TW}}$. We consider a total of 4 mean functions with combinations g_i+h_j from the following functions:

$$g_1(x)=5x(1-x),$$
$$g_2(x)=5\sin(2x),$$

and

$$h_1(x)=4I(x>\sqrt{2}/2),$$
$$h_2(x)=0.$$

For each mean function, we consider $n=30$, 100 and 500, ranging from small to large sample sizes respectively, and $\sigma=0.2$, 0.5, 1, 2 and 5, ranging from small to large variances respectively. Finally, for given n and σ, we simulate the random errors ε_i independently from $N(0,\sigma^2)$.

For each simulation setting, we generate observations and compute the estimators $\hat{\sigma}^2_{\mathrm{TW}}(m)$, $\hat{\sigma}^2_{\mathrm{MS}}(L)$, $\hat{\sigma}^2_{\mathrm{box}}(d,C)$ and $\hat{\sigma}^2_{\mathrm{cv}}$. Note that the bandwidth L in Müller and

Stadtmüller (1999) is not very sensitive to the estimation of σ^2. We consider both $L_s = m_s = n^{1/2}$ and $L_t = m_t = n^{1/3}$ as in Tong and Wang (2005). This leads to the corresponding d values as $d_s = (m_s/n)^2$ and $d_t = (m_t/n)^2$. Then together with $C = 2$ and 3, we have 4 different estimates for $\hat{\sigma}^2_{\text{box}}$. Recall that the CV estimator, $\hat{\sigma}^2_{\text{cv}}$, aims to figure out the best combination between d and C. We consider leave-one-out CV for $n = 30$, and 10-fold CV for $n = 100$ and $n = 500$, throughout the simulations.

We repeat the process 1000 times and compute the following relative MSEs, MSE/MSE$_{opt}$, for each method. Here, MSE$_{opt} = n^{-1}(\gamma_4 - 1)\sigma^4$ is specified as the optimal efficiency bound of all root-n consistent estimators of σ^2, and $\gamma_4 = E(\varepsilon^4)/\sigma^4$. For normal errors, we have $\gamma_4 = 3$ and MSE$_{opt} = 2\sigma^4/n$. We observe that negative estimates indicated by Tong and Wang (2005) and Müller and Stadtmüller (1999) do appear in certain simulations, though very rarely. We replace the negative estimates with zero when calculating the relative MSEs.

4.3.2 Non-equidistant Design

This section carries out simulation studies for unequally spaced designs. We generate design points from the beta distribution Beta$(3, 3)$. This is a bell shaped distribution on [0,1] with a mode at 0.5. Also for simplicity, we consider only two mean functions $g_2 + h_i$. All other settings are kept the same as those in Section 4.3.1.

Finally, recall that Müller and Stadtmüller (1999) and Tong and Wang (2005) do not apply to unequally spaced designs. We thus omit both the estimators but add in the pairwise regression estimator $\hat{\sigma}^2$ in (4.9) for comparison. Then correspondingly, we compute the relative MSEs for $\hat{\sigma}^2$, $\hat{\sigma}^2_{\text{box}}(d, C)$, and $\hat{\sigma}^2_{\text{cv}}$, respectively.

4.3.3 Simulation Results

Tables 4.5 and 4.2 list the relative MSEs for the mean functions with jump points, respectively, under the equidistant design. In general, we observe that MSE$(\hat{\sigma}^2_{\text{CV}}) \doteq$ MSE$(\hat{\sigma}^2_{\text{box}}) <$ MSE$(\hat{\sigma}^2_{\text{MS}}) <$ MSE$(\hat{\sigma}^2_{\text{TW}})$ for small and moderate σ values, and MSE$(\hat{\sigma}^2_{\text{CV}}) \doteq$ MSE$(\hat{\sigma}^2_{\text{box}}) \doteq$ MSE$(\hat{\sigma}^2_{\text{TW}}) <$ MSE$(\hat{\sigma}^2_{\text{MS}})$ for large σ values. These results show that the

proposed adaptive estimators outperform the existing estimators in the presence of jump discontinuities. We also observe that the comparative performance of $\hat{\sigma}^2_{\text{box}}(d_t, 2)$, $\hat{\sigma}^2_{\text{box}}(d_t, 3)$, $\hat{\sigma}^2_{\text{box}}(d_s, 2)$ and $\hat{\sigma}^2_{\text{box}}(d_s, 3)$ depends on the smoothness and continuity of the mean function, the sample size and the signal-to-noise ratio. As reported in Tong and Wang (2005), $\hat{\sigma}^2_{\text{box}}(d_s, \cdot)$ may not perform well when the sample size is small. As a compromise, $\hat{\sigma}^2_{\text{CV}}$ performs well in most settings.

In contrast, we list in Tables 4.3 and 4.4 the relative MSEs for the continuous mean functions $f_3(x)$ through $f_4(x)$, under the equidistant design. We observe that $\hat{\sigma}^2_{\text{box}}$, $\hat{\sigma}^2_{\text{CV}}$ and $\hat{\sigma}^2_{\text{TW}}$ perform very similar under various settings. More specifically, we observe that for a continuous mean function, very few z_{ij} values were detected from simulations as outliers. As a consequence, both $\hat{\sigma}^2_{\text{box}}(d_t, 2)$ and $\hat{\sigma}^2_{\text{box}}(d_t, 3)$ perform essentially the same as $\hat{\sigma}^2_{\text{TW}}(m_t)$, and both $\hat{\sigma}^2_{\text{box}}(d_s, 2)$ and $\hat{\sigma}^2_{\text{box}}(d_s, 3)$ perform essentially the same as $\hat{\sigma}^2_{\text{TW}}(m_s)$. Apart from them, $\hat{\sigma}^2_{\text{MS}}$ does not provide a comparable performance. This coincides the observation in Tong et al. (2013) that $\hat{\sigma}^2_{\text{MS}}$ is worse than $\hat{\sigma}^2_{\text{TW}}$ when the mean function is continuous.

Finally, we list in Tables 4.5 and 4.6 the relative MSEs for the settings with non-equidistant designs. Similarly as above, we observe that $\hat{\sigma}^2_{\text{box}}$ perform better than $\hat{\sigma}^2$ in the presence of jump discontinuities, and their performance are similar when the mean function is continuous. Meanwhile, $\hat{\sigma}^2_{\text{CV}}$ performs very well in most settings, especially when the sample size is small.

4.4 Real Application

For illustration, we apply the proposed methods to a real data example. The data were reported in Cobb (1978) on the annual volume of discharge in the Nile River from 1895 to 1934. In Figure 4.3, we find several observations with large variation and we suspect that the mean function might contain jump discontinuities. For this data with $n = 40$ observations, we choose $L_t = m_t = \lfloor n^{1/3} \rfloor = 3$ and $L_s = m_s = \lfloor n^{1/2} \rfloor = 6$ for $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\sigma}^2_{\text{TW}}$, respectively. Here, $\lfloor a \rfloor$ denotes the largest integer smaller than or equal to a. For the proposed methods, correspondingly we choose $d_t =$

$(m_t/n)^2 = 0.075^2$ and $d_s = (m_s/n)^2 = 0.15^2$. The estimated residual variances are as follows: $\hat{\sigma}^2_{\text{MS}}(L_t) = 126.9$, $\hat{\sigma}^2_{\text{MS}}(L_s) = 47.7$; $\hat{\sigma}^2_{\text{TW}}(m_t) = 119.9$ and $\hat{\sigma}^2_{\text{TW}}(m_s) = 144.8$; $\hat{\sigma}^2_{\text{box}}(d_t, 2) = 126.1$, $\hat{\sigma}^2_{\text{box}}(d_t, 3) = 119.9$, $\hat{\sigma}^2_{\text{box}}(d_s, 2) = 137.2$, $\hat{\sigma}^2_{\text{box}}(d_s, 3) = 144.8$ and $\hat{\sigma}^2_{\text{CV}} = 119.9$. We note that for a standard with $C = 3$, no outliers were identified so that $\hat{\sigma}^2_{\text{box}}(d_t, 3) = \hat{\sigma}^2_{\text{TW}}(m_t) = 119.9$ and $\hat{\sigma}^2_{\text{box}}(d_s, 3) = \hat{\sigma}^2_{\text{TW}}(m_s) = 144.8$. In addition, the cross validation method suggests to take $C = 3$ with a bandwidth at d_t and that results in the variance estimate as 119.9. Recall that the suggested value of σ^2 is 125 in Cobb (1978). We conclude that our pairwise regression method performs at least as well as the least squares estimator σ^2_{TW}. Nevertheless, the estimator σ^2_{MS} is very sensitive to the choice of the bandwidth and so is less reliable.

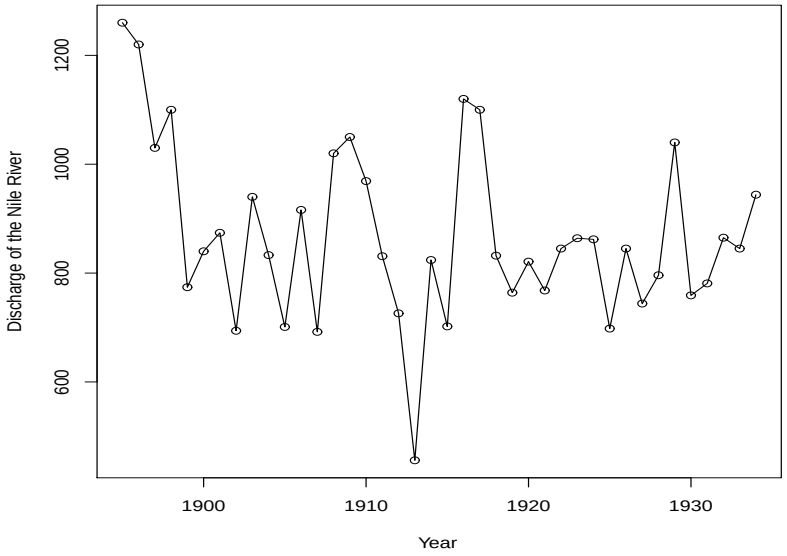

Figure 4.3: The Nile discharge data from 1895 to 1934.

4.5 Discussion

The proposed method can be readily extended to higher-dimensional regression models. Consider, for instance, the following bivariate nonparametric regression

model with jump discontinuities,

$$y_i = g^*(x_{1i}, x_{2i}) + h^*(x_{1i}, x_{2i}) + \varepsilon_i, \qquad 1 \le i \le n,$$

We can define d_{ij} by $d_{ij} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2}$ or by $d_{ij} = |x_{1i} - x_{1j}| + |x_{1i} - x_{1j}|$, and then proceed the estimation similarly as in Sections 4.2.2 and 4.2.3. Further research is necessary to investigate the practical rules for the corresponding adaptive method as well as to evaluate its finite-sample performance. Further, recall that the proposed method in this chapter is restricted to a constant residual variance assumption. As this may not be realistic in applications, it should be of interest to propose new pairwise regression methods for estimating the variance function in regression models with jump discontinuities.

Table 4.1: Relative MSEs of various estimators for the mean function $f_1(x) = g_1(x) + h_1(x)$, under equidistant design.

		$\hat{\sigma}^2_{\text{MS}}$		$\hat{\sigma}^2_{\text{TW}}$		$\hat{\sigma}^2_{\text{box}}$				$\hat{\sigma}^2_{\text{CV}}$
n	σ	L_t	L_s	m_t	m_s	$(d_t, 2)$	$(d_t, 3)$	$(d_s, 2)$	$(d_s, 3)$	
30	0.2	65.5	34.6	513	1368	1.64	1.60	1.67	1.79	1.62
	0.5	20.8	9.58	16.8	39.0	2.41	7.78	6.12	24.4	2.82
	1	13.2	4.94	2.87	4.15	2.48	2.83	3.40	4.13	2.87
	2	11.0	3.61	1.66	1.54	1.63	1.66	1.49	1.53	1.47
	5	10.3	3.23	1.49	1.26	1.51	1.50	1.25	1.26	1.30
100	0.2	14.3	12.0	308	939	1.37	1.38	1.22	1.21	1.25
	0.5	5.38	3.76	9.75	26.0	1.41	2.80	1.39	5.22	1.33
	1	4.05	2.52	2.02	2.92	1.71	1.98	1.99	2.82	1.59
	2	3.73	2.19	1.43	1.34	1.37	1.44	1.28	1.34	1.26
	5	3.64	2.10	1.36	1.20	1.35	1.36	1.18	1.20	1.19
500	0.2	3.85	3.92	130	777	1.24	1.22	1.11	1.11	1.17
	0.5	2.36	1.79	4.46	20.8	1.25	1.43	1.15	1.92	1.20
	1	2.15	1.47	1.41	2.29	1.25	1.38	1.35	2.13	1.32
	2	2.10	1.38	1.23	1.17	1.22	1.23	1.11	1.17	1.13
	5	2.08	1.36	1.22	1.10	1.24	1.22	1.13	1.11	1.16

Table 4.2: Relative MSEs of various estimators for the mean function $f_2(x) = g_2(x) + h_1(x)$, under equidistant design.

		$\hat{\sigma}^2_{\text{MS}}$		$\hat{\sigma}^2_{\text{TW}}$		$\hat{\sigma}^2_{\text{box}}$				$\hat{\sigma}^2_{\text{CV}}$
n	σ	L_t	L_s	m_t	m_s	$(d_t, 2)$	$(d_t, 3)$	$(d_s, 2)$	$(d_s, 3)$	
30	0.2	61.4	22.7	506	1297	1.70	1.95	76.1	1427	1.69
	0.5	20.6	9.06	16.6	37.2	3.91	11.7	24.7	37.0	4.51
	1	13.2	4.90	2.86	4.05	2.60	2.83	3.81	4.05	2.71
	2	11.0	3.61	1.66	1.53	1.64	1.66	1.50	1.53	1.50
	5	10.3	3.23	1.49	1.26	1.51	1.50	1.26	1.26	1.30
100	0.2	14.3	12.0	306	927	1.39	1.38	1.28	1.27	1.35
	0.5	5.39	3.78	9.72	25.7	1.45	3.34	1.81	9.95	1.41
	1	4.05	2.52	2.01	2.90	1.72	1.98	2.15	2.86	1.60
	2	3.73	2.20	1.43	1.34	1.40	1.44	1.29	1.34	1.26
	5	3.64	2.10	1.36	1.20	1.35	1.36	1.18	1.20	1.19
500	0.2	3.85	3.93	130	775	1.23	1.22	1.11	1.12	1.17
	0.5	2.36	1.79	4.46	20.7	1.25	1.44	1.16	2.33	1.23
	1	2.15	1.47	1.41	2.29	1.25	1.38	1.38	2.15	1.33
	2	2.10	1.38	1.23	1.17	1.22	1.23	1.11	1.17	1.13
	5	2.08	1.36	1.22	1.11	1.24	1.22	1.13	1.11	1.16

Table 4.3: Relative MSEs of various estimators for the mean function $f_3(x) = g_1(x) + h_2(x)$, under equidistant design.

		$\hat{\sigma}^2_{\text{MS}}$		$\hat{\sigma}^2_{\text{TW}}$		$\hat{\sigma}^2_{\text{box}}$				$\hat{\sigma}^2_{\text{CV}}$
n	σ	L_t	L_s	m_t	m_s	$(d_t, 2)$	$(d_t, 3)$	$(d_s, 2)$	$(d_s, 3)$	
30	0.2	10.2	3.54	1.53	1.64	1.56	1.53	1.72	1.64	1.58
	0.5	10.1	3.21	1.49	1.27	1.52	1.49	1.26	1.27	1.39
	1	10.1	3.17	1.48	1.25	1.51	1.49	1.24	1.25	1.40
	2	10.1	3.16	1.48	1.24	1.51	1.49	1.25	1.25	1.24
	5	10.1	3.16	1.48	1.24	1.51	1.49	1.25	1.25	1.29
100	0.2	3.65	2.10	1.35	1.21	1.34	1.36	1.21	1.21	1.25
	0.5	3.64	2.08	1.35	1.19	1.35	1.36	1.18	1.19	1.18
	1	3.63	2.08	1.35	1.19	1.35	1.36	1.18	1.19	1.17
	2	3.63	2.08	1.35	1.19	1.35	1.36	1.18	1.19	1.19
	5	3.63	2.08	1.35	1.19	1.35	1.36	1.18	1.19	1.18
500	0.2	2.08	1.35	1.22	1.11	1.24	1.22	1.12	1.11	1.17
	0.5	2.08	1.35	1.22	1.11	1.25	1.22	1.14	1.11	1.16
	1	2.08	1.35	1.22	1.12	1.25	1.22	1.14	1.12	1.16
	2	2.08	1.35	1.22	1.12	1.25	1.22	1.15	1.12	1.10
	5	2.08	1.35	1.22	1.12	1.25	1.22	1.15	1.12	1.15

Table 4.4: Relative MSEs of various estimators for the mean function $f_4(x) = g_2(x) + h_2(x)$, under equidistant design.

		$\hat{\sigma}^2_{\text{MS}}$		$\hat{\sigma}^2_{\text{TW}}$		$\hat{\sigma}^2_{\text{box}}$				$\hat{\sigma}^2_{\text{CV}}$
n	σ	L_t	L_s	m_t	m_s	$(d_t, 2)$	$(d_t, 3)$	$(d_s, 2)$	$(d_s, 3)$	
30	0.2	10.4	8.34	1.61	2.26	1.68	1.61	2.48	2.26	1.70
	0.5	10.1	3.48	1.50	1.32	1.54	1.50	1.34	1.32	1.44
	1	10.1	3.21	1.49	1.26	1.51	1.50	1.24	1.26	1.42
	2	10.1	3.17	1.48	1.25	1.51	1.49	1.24	1.25	1.24
	5	10.1	3.16	1.48	1.24	1.50	1.49	1.24	1.25	1.28
100	0.2	3.70	2.30	1.35	1.24	1.36	1.36	1.29	1.24	1.34
	0.5	3.65	2.09	1.35	1.20	1.34	1.35	1.19	1.20	1.19
	1	3.64	2.08	1.35	1.19	1.34	1.35	1.18	1.19	1.18
	2	3.64	2.08	1.35	1.19	1.34	1.36	1.18	1.19	1.19
	5	3.63	2.08	1.35	1.19	1.35	1.36	1.18	1.19	1.18
500	0.2	2.08	1.36	1.22	1.11	1.24	1.22	1.10	1.11	1.16
	0.5	2.08	1.35	1.22	1.11	1.25	1.22	1.13	1.11	1.16
	1	2.08	1.35	1.22	1.11	1.25	1.22	1.14	1.12	1.16
	2	2.08	1.35	1.22	1.12	1.25	1.22	1.14	1.12	1.09
	5	2.08	1.35	1.22	1.12	1.25	1.22	1.15	1.12	1.15

Table 4.5: Relative MSEs of various estimators for the mean function $f_2(x) = g_2(x) + h_1(x)$, under non-equidistant design.

		$\hat{\sigma}^2$		$\hat{\sigma}^2_{\text{box}}$				$\hat{\sigma}^2_{\text{CV}}$
n	σ	d_t	d_s	$(d_t, 2)$	$(d_t, 3)$	$(d_s, 2)$	$(d_s, 3)$	
30	0.2	1174	2875	34.7	34.7	659	1647	50.2
	0.5	35.2	81.6	7.41	21.7	49.1	76.4	8.94
	1	4.28	7.45	3.74	4.17	6.89	7.40	4.16
	2	1.85	1.92	1.81	1.85	1.88	1.92	1.84
	5	1.56	1.35	1.57	1.57	1.35	1.35	1.42
100	0.2	654	1726	1.60	1.61	1.57	1.59	1.57
	0.5	18.8	46.2	1.64	4.31	2.30	16.7	1.72
	1	2.63	4.23	1.99	2.54	2.97	4.12	2.00
	2	1.54	1.50	1.50	1.54	1.45	1.50	1.43
	5	1.45	1.31	1.45	1.45	1.30	1.31	1.36
500	0.2	252	1233	1.43	1.44	1.33	1.33	1.40
	0.5	7.90	33.1	1.43	1.80	1.36	3.48	1.41
	1	1.86	3.34	1.52	1.82	1.94	3.17	1.58
	2	1.46	1.44	1.44	1.46	1.36	1.43	1.44
	5	1.43	1.30	1.43	1.43	1.30	1.30	1.38

Table 4.6: Relative MSEs of various estimators for the mean function $f_4(x) = g_2(x) + h_2(x)$, under non-equidistant design.

		$\hat{\sigma}^2$		$\hat{\sigma}^2_{\text{box}}$				$\hat{\sigma}^2_{\text{cv}}$
n	σ	d_t	d_s	$(d_t, 2)$	$(d_t, 3)$	$(d_s, 2)$	$(d_s, 3)$	
30	0.2	2.29	3.55	2.21	2.28	3.39	3.54	2.22
	0.5	1.63	1.51	1.62	1.63	1.49	1.51	1.52
	1	1.56	1.35	1.57	1.56	1.34	1.35	1.40
	2	1.54	1.31	1.55	1.54	1.30	1.31	1.38
	5	1.53	1.30	1.55	1.54	1.30	1.30	1.29
100	0.2	1.58	1.50	1.56	1.58	1.47	1.50	1.53
	0.5	1.48	1.33	1.47	1.48	1.33	1.33	1.38
	1	1.46	1.32	1.45	1.46	1.32	1.32	1.38
	2	1.46	1.32	1.44	1.46	1.32	1.32	1.38
	5	1.45	1.32	1.44	1.45	1.33	1.32	1.38
500	0.2	1.43	1.32	1.43	1.43	1.30	1.32	1.37
	0.5	1.43	1.30	1.43	1.43	1.30	1.30	1.38
	1	1.43	1.30	1.43	1.43	1.30	1.30	1.38
	2	1.43	1.30	1.43	1.43	1.31	1.30	1.38
	5	1.43	1.29	1.43	1.43	1.31	1.29	1.38

Chapter 5

Testing Discontinuities in Nonparametric Regression

5.1 Introduction

It is known that the estimates of the mean function with or without the smoothness assumption are not only quantitatively but also qualitatively different (Müller and Stadtmüller; 1999). Also, we have illustrated in the Chapters 2 and 4, that the variance estimation for a smooth mean function is quite different from that for a mean function with jump discontinuities. A method derived under incorrect assumption about the smoothness of the mean function, may generate a significant loss of efficiency or provide even invalid estimators. This results in a fundamental model selection problem and it is often needed to detect whether there are jump discontinuities in the mean function before fitting the model.

Consider the following nonparametric regression model

$$y_i = g(x_i) + h(x_i) + \varepsilon_i, \qquad 0 \leq i \leq n, \tag{5.1}$$

where y_i are observations, x_i are design points and ε_i are i.i.d. random errors with mean zero and variance σ^2. In model (5.1), $f(x) = g(x) + h(x)$ is the mean function. We assume that $g(x)$ is a smooth function and $h(x)$ is a step function with form

$$h(x) = \sum_{j=1}^{p} \psi_j I(x \geq t_j), \qquad 0 < t_1 < \cdots < t_p < 1,$$

where p is the number of jumps, ψ_j are the jump magnitudes taking at positions t_j, and $I(x \geq t)$ is the indicator function with value 1 if $x \geq t$ and value 0 otherwise.

A number of papers in literature focus on testing the smoothness for the mean function of nonparametric regression including, Wu and Chu (1993a); Qiu and Yandell (1998); Gijbels and Goderniaux (2004),Bowman et al. (2006); Qiu (2007); Neumeyer and Van Keilegom (2009). A common approach is to estimate the left and right limits of the mean function and then use their difference to construct appropriate test statistics. As a common practice, a large absolute value of the difference indicates that the mean function may contain jump discontinuities (Müller; 1992; Grégoire and Hamrouni; 2002; Qiu; 2005; Cheng and Raimondo; 2008; Joo and Qiu; 2009; Chu et al.; 2012).

Apart from the traditional methods, Müller and Stadtmüller (1999) proposed a difference-based procedure for testing jump discontinuities the mean function $f(x)$ in model (5.1). Consider an equivalent design with $x_i = i/n$ for $i = 1, ..., n$, and define $\gamma = \sum_{j=1}^{p} \psi_j^2$ as the total amount of discontinuity in the data. When $\gamma = 0$, we have $\psi_j = 0$ for all j and so the mean function is smooth; otherwise, the regression model may contain jump discontinuities. By this, the model selection problem is equivalent to testing

$$H_0 : \gamma = 0 \qquad \text{versus} \qquad H_1 : \gamma > 0. \tag{5.2}$$

Let $L = o(n) \geq 1$ and

$$Z_k = \frac{1}{2(n-L)} \sum_{i=1}^{n-L} (y_{i+k} - y_i)^2, \qquad 1 \leq k \leq L.$$

Noting that $E(Z_k) \approx \sigma^2 + l_k \gamma$ where $l_k = k/2(n-L)$, the authors regressed Z_k on l_k through the linear model

$$Z_k = \sigma^2 + l_k \gamma + \xi_k, \qquad 1 \leq k \leq L. \tag{5.3}$$

The intercept σ^2 and the slope γ were then estimated by least squares. We denote them by $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\gamma}_{\text{MS}}$. Now for testing the hypotheses (5.2), they constructed a test statistic as

$$T_{\text{MS}} = \frac{\sqrt{m}\hat{\gamma}_{\text{MS}}}{\sqrt{12(\tilde{\mu}_4 - \tilde{\sigma}^4)/5}}, \tag{5.4}$$

where $\tilde{\mu}_4$ and $\tilde{\sigma}^4$ are consistent estimators for $\mu_4 = \mathrm{E}(\varepsilon^4)$ and σ^4. For the special case of normal errors, $\mu_4 - \sigma^4 = 2\sigma^4$ and so $\tilde{\mu}_4 - \tilde{\sigma}^4$ can be replaced by $2\hat{\sigma}^4_{\mathrm{MS}}$. Under some regularity conditions, the null distribution of T_{MS} follows asymptotically a standard normal distribution.

Note that Z_k uses only the first $n - L$ pairs of differences out of a total of $n - k$ pairs in $\{y_{1+k} - y_1, \ldots, y_n - y_{n-k}\}$. By ignoring the last $L - k$ pairs, we will show in Sections 5.3 and 5.4 that their estimator $\hat{\sigma}^2_{\mathrm{MS}}$ is a less efficient estimator for σ^2, especially when n is small or $L - k$ is large. As a consequence, the test statistic (5.4) is either less powerful or is even invalid for testing jump discontinuities, that is, the type I error of the test may not be well controlled. For this point, we can refer to the simulation study in Section 5.3, where the simulated type I error of T_{MS} will be as large as 0.237 at the significance level of 0.05 when $n = 30$. Note that a similar phenomenon was also observed in Tong et al. (2013) where a smooth regression model is assumed.

In this chapter, we propose to fully use the pairs of differences and develop new estimators for σ^2 and γ. A new test procedure will also be constructed and we show that the proposed test provides a better performance than T_{MS}. Specifically, we organize the rest of this chapter as follows. In Section 5.2, we propose a new estimation method for σ^2 and γ. In Section 5.3, we study the theoretical properties of the new estimators and then propose a new procedure for testing whether the mean function contains jump discontinuities. In Section 5.4, we evaluate and compare the proposed estimators and the proposed test procedure with some existing methods. We then provide the proofs of the theorems in Section 5.5.

5.2 Main Results

To make full use of the available pairs of differences, we define

$$s_k = \frac{1}{2(n-k)} \sum_{i=1}^{n-k} (y_{i+k} - y_i)^2, \qquad k = 1, ..., m.$$

The same as in Müller and Stadtmüller (1988) and Tong and Wang (2005), we refer to them as the lag-k Rice estimators. For the sake of fairness, we let $m = L$ and also

assume the same conditions as in Müller and Stadtmüller (1999). In particular, the following condition on the locations of jump points is assumed:

$$\min_{1\le j\le p+1}(t_j - t_{j-1}) \ge 2m/n, \tag{5.5}$$

where $t_0 = 0$ and $t_{p+1} = 1$. This condition ensures that the different change points are not located too close to each other so that they can be separated by the proposed method.

For the equidistant design, it is easy to verify that

$$\begin{aligned} E(s_k) &= \sigma^2 + \frac{1}{2(n-k)}\sum_{i=1}^{n-k}\Big[g(x_{i+k}) - g(x_i) + h(x_{i+k}) - h(x_i)\Big]^2 \\ &= \sigma^2 + \frac{k}{2(n-k)}\gamma + \frac{k^2}{2n^2}\delta + o(\frac{k^2}{n^2}), \end{aligned} \tag{5.6}$$

where $\gamma = \sum_{j=1}^{p}\psi_j^2$ as in Section 5.1 and $\delta = \int_0^1 g'(t)^2dt + 2\int_0^1 g'(t)dh(t)$. Let $\theta = (\theta_1, \theta_2) = (\sigma^2, \gamma/2)$ and $d_k = k/(n-k)$. Model (5.6) can be represented as

$$E(s_k) = \theta_1 + d_k\theta_2 + \eta_k,$$

where $\eta_k = k^2\delta/2n^2 + o(k^2/n^2)$. Treating η_k as a negligible term, we have a simple linear regression model with d_k being the independent variable and s_k the response variables. Unlike the Z_k values where exactly $n - L$ pairs are involved, we note that the lag-k Rice estimators s_k involve different number of pairs and this makes the computation much more challenging. To assign a same weight to each pair, we assign weights $w_k = (n-k)/N$ for $k = 1, \ldots, m$ to each response s_k accordingly, where $N = (2n - m - 1)m/2$ is the total number of pairs involved in the regression.

Note also that the responses s_k are correlated with each other. With some straightforward computation, the asymptotic covariance matrix of $\mathbf{s} = (s_1, ..., s_m)^T$ is given by $\Sigma = (\Sigma_{ij})_{m\times m}$, where

$$\Sigma_{i,j} = \begin{cases} n^{-1}(\mu_4 - \sigma^4) + o(n^{-1}) & \text{if } i \ne j, \\ n^{-1}\mu_4 + o(n^{-1}) & \text{if } i = j. \end{cases}$$

we then apply the generalized least squares method to estimate the parameters.

Specifically by McElroy (1967), we have the estimators of θ_1 and θ_2 as

$$\hat{\theta}_{\text{new},1} = \sum_{k=1}^{m} w_k s_k - \bar{d}_w \hat{\theta}_{\text{new},2},$$
$$\hat{\theta}_{\text{new},2} = (\sum_{k=1}^{m} w_k d_k^2 - \bar{d}_w^2)^{-1} \sum_{k=1}^{m} w_k (d_k - \bar{d}_w) s_k,$$

where $\bar{d}_w = \sum_{k=1}^{m} w_k d_k$. Then correspondingly, the new estimators of σ^2 and γ are $\hat{\sigma}^2_{\text{new}} = \hat{\theta}_{\text{new},1}$ and $\hat{\gamma}_{\text{new}} = 2\hat{\theta}_{\text{new},2}$. They are the best linear unbiased estimators of σ^2 and γ, respectively (Kariya and Kurata; 2004).

5.3 Asymptotic Properties

In this section, we investigate the theoretical properties of the proposed estimators. For ease of notation, let

$$a_k = 1 - \bar{d}_w \frac{(d_k - \bar{d}_w)}{\sum_{k=1}^{m} w_k d_k^2 - \bar{d}_w^2} \quad \text{and} \quad b_k = \frac{d_k - \bar{d}_w}{\sum_{k=1}^{m} w_k d_k^2 - \bar{d}_w^2}.$$

Then for $\lambda, \rho \in \mathrm{R}$, we define

$$\hat{\theta}_{\text{new}}(\lambda, \rho) = \lambda \hat{\theta}_{\text{new},1} + \rho \hat{\theta}_{\text{new},2} = \sum_{k=1}^{m} (\lambda a_k + \rho b_k) w_k s_k.$$

Let also $y = (y_1, \ldots, y_n)^T$, $c_0 = 0$ and $c_k = (\lambda a_k + \rho b_k)/2N$ for $k = 1, \ldots, m$. We note that $\hat{\theta}_{\text{new}}(\lambda, \rho)$ can be represented as a quadratic form $\hat{\theta}_{\text{new}}(\lambda, \rho) = Y^T D Y$, where D is an $n \times n$ matrix with elements

$$d_{ij} = \begin{cases} \sum_{k=1}^{m} c_k + \sum_{k=0}^{\min\{i-1, n-i, m\}} c_k & 1 \le i = j \le n \\ -c_{|i-j|} & 0 < |i-j| = k \le m \\ 0 & \text{otherwise.} \end{cases}$$

We can derive its variance as

$$\begin{aligned} \text{var}[\hat{\theta}_{\text{new}}(\lambda, \rho)] &= (\text{var}(\varepsilon^2) - 2\sigma^4)\text{tr}(\text{diag}(D)^2) + 2\sigma^4 \text{tr}(D^2) \\ &\quad + 4\sigma^2 f^T D^2 f + 4\sigma^3 \mu_3 (f^T D \text{diag}(D) \mathbf{1}), \end{aligned}$$

where $f = (f(x_1), \ldots, f(x_n))^T$ and $\text{diag}(D)$ is the diagonal matrix of D. This form makes it convenient to calculate the theoretical results. The following theorem gives the mean squared errors (MSE) for the proposed estimators and also the estimators in Müller and Stadtmüller (1999) for comparison.

Theorem 8. *Assume that condition (5.5) holds, $m \to \infty$ and $L = m = n^r$ for $1/2 < r < 2/3$. Then for any mean function $f = g + h$ with the first derivative of g being bounded, we have*

$$\begin{aligned}
\mathrm{MSE}(\hat{\sigma}^2_{\mathrm{new}}) &= \frac{1}{n}\mathrm{var}(\varepsilon^2) + \frac{m}{15n^2}\mathrm{var}(\varepsilon^2) + \frac{4m}{15n^2}\sigma^2\gamma + o(\frac{m}{n^2}),\\
\mathrm{MSE}(\hat{\sigma}^2_{\mathrm{MS}}) &= \frac{1}{n}\mathrm{var}(\varepsilon^2) + \frac{16m}{15n^2}\mathrm{var}(\varepsilon^2) + \frac{4m}{15n^2}\sigma^2\gamma + o(\frac{m}{n^2}),\\
\mathrm{MSE}(\hat{\gamma}_{\mathrm{new}}) &= \frac{12}{5m}\mathrm{var}(\varepsilon^2) + \frac{48}{5m}\sigma^2\gamma + o(\frac{1}{m}),\\
\mathrm{MSE}(\hat{\gamma}_{\mathrm{MS}}) &= \frac{12}{5m}\mathrm{var}(\varepsilon^2) + \frac{48}{5m}\sigma^2\gamma + o(\frac{1}{m}).
\end{aligned}$$

The proof of Theorem 8 is given in Section 5.5.1. Theorem 8 indicates that $\hat{\gamma}_{\mathrm{new}}$ and $\hat{\gamma}_{\mathrm{MS}}$ are asymptotically equivalent. Both $\hat{\sigma}^2_{\mathrm{new}}$ and $\hat{\sigma}^2_{\mathrm{MS}}$ attain the root-n convergence rate. Compared to the MS estimator, the proposed new estimator $\hat{\gamma}_{\mathrm{new}}$ has a smaller second order term in the MSE expression. This improvement results from the information of residual variance provided by $m(m-1)/2$ extra pairs in our model. The finite-sample performance of the new estimators is presented in Section 5.4.

Theorem 9. *Assume that $E(\varepsilon^3) = 0$, $m \to \infty$ and $m = n^r$ for $1/2 < r < 2/3$. We have the following asymptotic normality for the proposed estimators,*

$$\begin{pmatrix} \sqrt{n}(\hat{\sigma}^2_{\mathrm{new}} - \sigma^2) \\ \sqrt{m}(\hat{\gamma}_{\mathrm{new}} - \gamma) \end{pmatrix} \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, (\mu_4 - \sigma^4) \begin{pmatrix} 1 & 0 \\ 0 & 12/5 \end{pmatrix} \right) \qquad \text{as } n \to \infty.$$

where $\xrightarrow{D}$ denotes convergence in distribution.

The proof of Theorem 9 is given in Section 5.5.2. Theorem 9 can be used to test whether or not the mean function is smooth. Specifically, to test $H_0 : \gamma = 0$ versus $H_1 : \gamma > 0$, we construct the test statistic as

$$T_{\mathrm{new}} = \frac{\sqrt{m}\hat{\gamma}_{\mathrm{new}}}{\sqrt{12(\tilde{\mu}_4 - \tilde{\sigma}^4)/5}},$$

where $\tilde{\mu}_4$ and $\tilde{\sigma}^2$ are consistent estimates of μ_4 and σ^2, respectively. If we treat η_k as normal errors, then $\mu_4 - \sigma^4 = 2\sigma^4$ and we can replace $\tilde{\mu}_4 - \tilde{\sigma}^4$ by $2\hat{\sigma}^4_{\mathrm{new}}$. Under H_0, the test statistic T_{new} follows a standard normal distribution. Then at the significance level of α, we reject H_0 if the observed T_{new} value is larger than z_α, where z_α is the upper αth percentile of the standard normal distribution.

5.4 Simulation Studies

In this section, we conduct Monte Carlo simulations to assess the finite-sample performance of the proposed estimators. For a fair comparison, we follow the same simulation settings as in Müller and Stadtmüller (1999). Specifically, we consider respectively three smooth functions:

$$g_1(x) = 0,$$
$$g_2(x) = x,$$
$$g_3(x) = 4x(1-x),$$

and three jump functions:

$$h_1(x) = 0,$$
$$h_2(x) = I(x \geq 0.5),$$
$$h_3(x) = I(x \geq 0.25) - 1.5I(x \geq 0.5).$$

In total, we have 9 combinations for the mean function. Note that $g_i + h_1$ for $i = 1, 2, 3$ are smooth functions with no jump points.

For the sample size, we consider $n = 30$, 100 and 500 that correspond to small, moderate and large sample sizes, respectively. The random errors are generated independently from the normal distribution $N(0, 0.5^2)$. For the bandwidth $\hat{m}$, we again follow the method proposed in Müller and Stadtmüller (1999). Specifically, it is

$$\hat{m} = \underset{m}{\operatorname{argmin}} \left\{ \frac{1}{2m_0+1} \sum_{i=m-m_0}^{m+m_0} \hat{\gamma}^2(i) - \left[\frac{1}{2m_0+1} \sum_{i=m-m_0}^{m+m_0} \hat{\gamma}(i) \right]^2 \right\},$$

where $m_0 = \max\{[n/50], 2\}$ with $[x]$ being the greatest integer smaller than or equal to x. This criterion suggests to choose the bandwidth that minimizes an approximation for the variance of estimator in the area of $[m - m_0, m + m_0]$.

For each simulated data set, we calculate the estimates of γ and σ^2 for both the new method and the MS method. We then repeat the procedure 1000 times for each simulation setting and report the MSE of $\hat{\gamma}$ and the relative mean squared errors

Table 5.1: Simulation results for MSE of $\hat{\gamma}_{\text{new}}$, $\hat{\gamma}_{\text{MS}}$ and relative MSE of $\hat{\sigma}^2_{\text{new}}$, $\hat{\sigma}^2_{\text{MS}}$.

n	$h(x)$	$g(x)$	MSE($\hat{\gamma}_{\text{new}}$)	MSE($\hat{\gamma}_{\text{MS}}$)	rMSE($\hat{\sigma}^2_{\text{new}}$)	rMSE($\hat{\sigma}^2_{\text{MS}}$)
30	1	1	0.12	0.18	1.20	1.86
	1	2	0.20	0.27	1.23	1.94
	1	3	0.69	0.74	1.32	2.07
	2	1	0.44	0.53	1.48	2.45
	2	2	1.25	1.41	1.70	2.92
	2	3	1.21	1.29	1.65	2.82
	3	1	2.45	2.61	2.65	4.02
	3	2	2.72	2.89	2.69	3.99
	3	3	4.67	5.10	2.67	3.89
100	1	1	0.027	0.025	1.07	1.48
	1	2	0.068	0.076	1.10	1.54
	1	3	0.45	0.42	1.17	1.75
	2	1	0.13	0.15	1.25	1.76
	2	2	0.62	0.66	1.47	2.16
	2	3	0.58	0.57	1.37	2.04
	3	1	0.56	0.58	1.60	2.24
	3	2	0.60	0.61	1.61	2.19
	3	3	2.95	2.97	2.38	3.69
500	1	1	0.0047	0.0061	1.06	1.37
	1	2	0.029	0.030	1.09	1.37
	1	3	0.32	0.30	1.27	1.77
	2	1	0.040	0.043	1.19	1.51
	2	2	0.27	0.28	1.43	1.83
	2	3	0.36	0.34	1.42	1.99
	3	1	0.12	0.13	1.47	1.92
	3	2	0.13	0.13	1.48	1.94
	3	3	1.47	1.46	2.54	3.60

(rMSE) of $\hat{\sigma}^2$ (i.e., $(n/2\sigma^4)$MSE) in Table 5.1, respectively. From the results in Table 5.1, we observe that the performance of $\hat{\gamma}_{\text{new}}$ and $\hat{\gamma}_{\text{MS}}$ is very similar to each other. In addition, it is evident that $\hat{\sigma}^2_{\text{new}}$ outperforms $\hat{\sigma}^2_{\text{MS}}$ significantly in all the settings. Combining the theoretical results in Theorem 8, we conclude that the additional pairs introduced into the regression does improve the overall performance of the estimators, in both theory and simulations.

Finally, we conduct a simple simulation study to compare the performance of the test statistics T_{new} and T_{MS} for testing $H_0 : \gamma = 0$ versus $H_1 : \gamma > 0$. For the smooth function, we consider only $g = 0$ for simplicity. Whereas for the step function, we consider $h(x) = \psi I(x \geq 0.5)$ with the ψ value ranging from 0 to 1.6. The sample size n is set to be 30, 100, 200 and 500, and the random errors are generated independently from $N(0, 0.25)$. We set the significance level at $\alpha = 0.05$ and repeat the test procedure 1000 times for each setting. The simulation results are reported in Figure 5.4. Note that $\gamma = 0$ in $h(x)$ represents the null hypothesis is true and any non-zero γ value indicates a false null hypothesis. From Figure 5.4, we observe that the simulated type I errors of T_{MS} exceed the nominal level at 0.05 in all three settings. For instance, when $n = 30$, the simulated type I error of T_{MS} is as large as 0.237. Whereas for the new method, the simulated type I error is always smaller than that of T_{MS}; In particular, when $n = 500$, the simulated type I error reduces to about the nominal level. This shows that our proposed test is an asymptotically valid test. Overall, it is evident that the proposed test method provides a more accurate control than the test method in Müller and Stadtmüller (1999). In addition, we note that the simulated power of T_{new} is larger than that of T_{MS} in a wide range of settings, especially when γ is not too small.

5.5 Proofs

In this section, we provide technical proofs for main results.

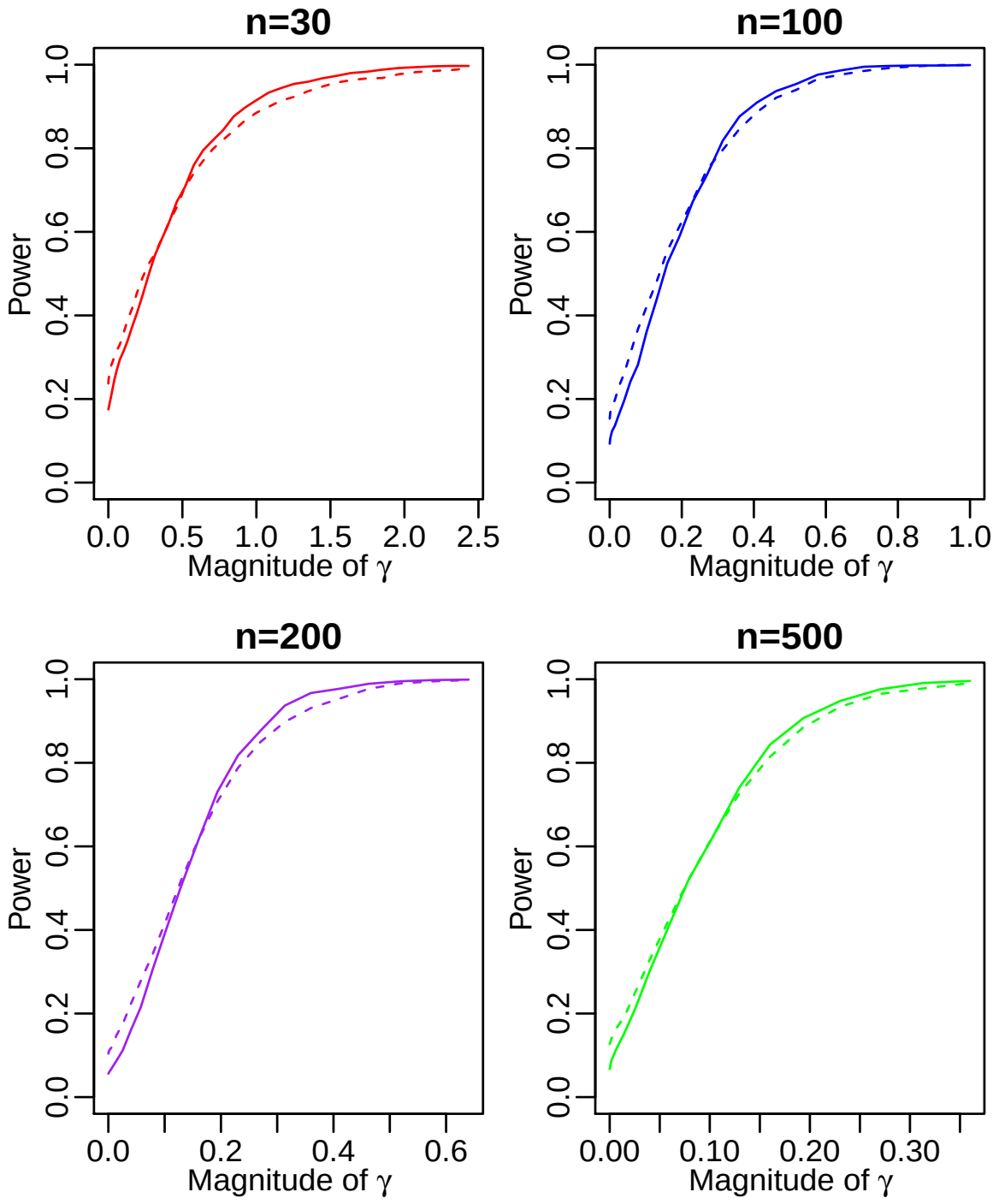

Figure 5.1: Test power of T_{MS} (dashed lines) and T_{new} (solid lines) against the magnitude of γ.

5.5.1 Proof of Theorem 8

We first present two lemmas. Lemma 8 can be derived with some tedious but simple calculation. For Lemma 9, we provide a detailed proof.

Lemma 8. *Assume that $m \to \infty$ and $m/n \to 0$. We have*

(a) $\sum_{k=1}^{m} c_k = \frac{1}{2N}[(m - \frac{m^2}{2n} + o(\frac{m^2}{n}))\lambda + (m + o(m))\rho]$.

(b) $\sum_{k=1}^{l-1} c_k = \frac{1}{2N}[(4l - \frac{3l^2}{m} + o(l) + O(1))\lambda$
$+ (\frac{6nl^2}{m^2} - \frac{6nl}{m} - \frac{12ln}{m^2} + \frac{4l^3}{m^2} + 9l + \frac{6n}{m} - \frac{12l^2}{m} + o(l) + o(\frac{n}{m}))\rho], \quad 1 \le l \le m.$

(c) $\sum_{k=1}^{m} c_k^2 = \frac{1}{4N^2}[(4m - \frac{4m^2}{n} - 12 + o(\frac{m^2}{n}) + o(1))\lambda^2$
$+ (\frac{12n^2}{m} - 24n - \frac{72n^2}{m^2} + o(n) + o(\frac{n^2}{m^2}))\rho^2$
$+ (-12n + 20m + \frac{60n}{m} + o(m) + o(\frac{n}{m}))\lambda\rho]$.

(d) $\sum_{k=l}^{m} kc_k = \frac{1}{2N}[\lambda(O(m^2)) + \rho(O(mn))], \quad 1 \le l \le m.$

(e) $\sum_{k=1}^{l-1} k^2 c_k = \frac{1}{2N}[\lambda(O(l^3)) + \rho(O(l^3 n/m))], \quad 1 \le l \le m.$

(f) $\sum_{k=1}^{m} k^2 c_k = \frac{1}{2N}[\lambda(O(m^3)) + \rho(O(m^2 n))]$.

Lemma 9. *Assume that $m \to \infty$ and $m/n \to 0$. We have*

(I) $\mathrm{tr}(D^2) = \frac{1}{4N^2}\Big[\lambda^2\big(4nm^2 - \frac{56}{15}m^3 + 8nm + o(m^3) + o(nm)\big)$
$+ \rho^2\big(\frac{12}{5}n^2 m + 24\frac{n^3}{m} - 60n^2 - \frac{34}{5}nm^2 - 144\frac{n^3}{m^2} + o(nm^2) + o(n^2) + o(\frac{n^3}{m^2})\big)$
$+ \lambda\rho\big(-\frac{2}{5}nm^2 - 24n^2 + o(nm^2) + o(n^2)\big)\Big]$.

(II) $\mathrm{tr}(\mathrm{diag}(D)^2) = \frac{1}{4N^2}\Big[\lambda^2\big(4nm^2 - \frac{56}{15}m^3 + o(m^3) + o(nm)\big)$
$+ \rho^2\big(\frac{12}{5}n^2 m - \frac{34}{5}nm^2 + o(nm^2) + o(n^2) + o(\frac{n^3}{m^2})\big) + \lambda\rho\big(-\frac{2}{5}nm^2 + o(nm^2)\big)\Big]$.

(III) $f^T D^2 f = \frac{1}{4N^2}\Big[\lambda^2\big(\frac{4}{15}m^3 + o(m^3)\big) + \rho^2\big(\frac{12}{5}n^2 m + 24n^2 - \frac{14}{5}nm^2 + o(n^2) + o(nm^2)\big)$
$+ \lambda\rho\big(-\frac{2}{5}nm^2 + o(nm^2)\big)\Big]$.

(IV) $f^T D\mathrm{diag}(D)\boldsymbol{1} = O(\frac{m^2}{n^3}) + O(\frac{1}{n^2})$.

Proof of Lemma 9. We first prove (I). Note that

$$\begin{aligned}\mathrm{tr}(D^2) &= (n-2m)\big[(2\sum_{k=1}^{m} c_k)^2 + 2\sum_{k=1}^{m} c_k^2\big] \\ &\quad +2\sum_{l=1}^{m}\big[(\sum_{k=1}^{m} c_k + \sum_{k=1}^{l-1} c_k)^2 + \sum_{k=1}^{m} c_k^2 + \sum_{k-1}^{l-1} c_k^2\big]. \qquad (5.7)\end{aligned}$$

By Lemma 8 (a) and (c), we have

$$\begin{aligned}(\sum_{k=1}^{m} c_k)^2 &= \frac{1}{4N^2}\big[\lambda^2(m^2 - \frac{m^3}{n} + o(\frac{m^3}{n})) + \rho^2(m^2 + o(m^2)) + \lambda\rho(2m^2 + o(m^2))\big],\\ \sum_{k=1}^{m} c_k^2 &= \frac{1}{4N^2}\big[\lambda^2(4m + o(m)) + \rho^2(\frac{12n^2}{m} - 24n - \frac{72n^2}{m^2} + o(n) + o(\frac{n^2}{m^2}))\\ &\quad +\lambda\rho(-12n + o(n))\big].\end{aligned}$$

This leads to

$$\begin{aligned}&(n-2m)\big[(2\sum_{k=1}^{m} c_k)^2 + 2\sum_{k=1}^{m} c_k^2\big]\\ =\ &\frac{1}{4N^2}\Big[\lambda^2(4nm^2 - 12m^3 + 8nm + o(m^3) + o(nm))\\ &+\rho^2(\frac{24n^3}{m} - 96n^2 - \frac{144n^3}{m^2} + 4nm^2 + o(n^2) + o(\frac{n^3}{m^2}) + o(nm^2))\\ &+\lambda\rho(8nm^2 - 24n^2 + o(m^2) + o(n))\Big]. \qquad (5.8)\end{aligned}$$

In addition, by Lemma 8 (a), (b) and (c) we have

$$\begin{aligned}\sum_{l=1}^{m}(\sum_{k=1}^{m} c_k + \sum_{k=1}^{l-1} c_k)^2 &= \frac{1}{4N^2}\Big[\lambda^2(\frac{62}{15}m^3 + o(m^3)) + \lambda\rho(-\frac{21}{5}nm^2 + o(nm^2))\\ &\quad +\rho^2(\frac{6}{5}n^2m - \frac{27}{5}nm^2 + o(nm^2) + o(n^2))\Big] \qquad (5.9)\end{aligned}$$

and

$$\sum_{l=1}^{m}(\sum_{k=1}^{m} c_k^2 + \sum_{k=1}^{l-1} c_k^2) = \frac{1}{4N^2}(\lambda^2 O(m^2) + \rho^2(18n^2 + o(n^2)) + \lambda\rho O(mn)). \qquad (5.10)$$

Now plugging (5.8), (5.9) and (5.10) into (5.7), we have $\mathrm{tr}(D^2)$ as shown in (I).

For (II), we have

$$\begin{aligned}\mathrm{tr}(\mathrm{diag}(D)^2) &= \frac{1}{4N^2}\left[(n-2m)\left(2\sum_{k=1}^{m} c_k\right)^2 + 2\sum_{l=1}^{m}\left(\sum_{k=1}^{m} c_k + \sum_{k=1}^{l-1} c_k\right)^2\right]\\ &= \frac{1}{4N^2}\Big[\lambda^2\big(4nm^2 - \frac{56}{15}m^3 + o(m^3) + o(nm)\big)\\ &\quad +\rho^2\big(\frac{12}{5}n^2m - \frac{34}{5}nm^2 + o(nm^2) + o(n^2) + o(\frac{n^3}{m^2})\big)\\ &\quad +\lambda\rho\big(-\frac{2}{5}nm^2 + o(nm^2)\big)\Big]\end{aligned}$$

Now we prove (III). Let $f_i = f(x_i)$, $g_i = g(x_i)$, $h_i = h(x_i)$, $g_i^{'} = g^{'}(x_i)$ and $g_i^{''} = g^{''}(x_i)$. Noting that D is a symmetric matrix, we have

$$f^T D^2 f = g^T D^T D g + 2 g^T D^T D h + h^T D^T D h = p^T p + 2 p^T q + q^T q,$$

where $p = Dg = (p_1, p_2, \ldots, p_n)^T$ and $q = Dh = (q_1, q_2, \ldots, q_n)^T$. For $i \in [m+1, n-m]$, by Lemma 8 (f) we have

$$\begin{aligned} p_i &= \sum_{k=1}^{m} c_k \left[(g_i - g_{i-k}) - (g_{i+k} - g_i) \right] \\ &= -\frac{2 g_i^{''}}{n^2} \sum_{k=1}^{m} k^2 c_k (1 + o(1)) \\ &= \frac{1}{2N} [\lambda(O(\frac{m^3}{n^2})) + \rho(O(\frac{m^2}{n}))] \end{aligned}$$

and

$$\begin{aligned} q_i &= \sum_{k=1}^{m} c_k \left[(h_i - h_{i-k}) - (h_{i+k} - h_i) \right] \\ &= \sum_{k=1}^{m} c_k \left[\psi_j I(x_{i-k} < t_j \le x_i) - \psi_{j+1} I(x_i < t_j \le x_{i+k}) \right]. \end{aligned}$$

For $i \in [1, m]$, by Lemma 8 (d), (e) and (f) we have

$$\begin{aligned} p_i &= \sum_{k=1}^{i-1} c_k (g_i - g_{i-k}) - \sum_{k=1}^{m} c_k (g_{i+k} - g_i) \\ &= -\frac{g_i^{'}}{n} \sum_{k=i}^{m} k c_k (1 + o(1)) \\ &= \frac{1}{2N} [\lambda(O(\frac{m^2}{n})) + \rho(O(m))]. \end{aligned}$$

Similar, for $i \in [n-m+1, n]$ we have

$$p_i = \frac{1}{2N} [\lambda(O(\frac{m^2}{n})) + \rho(O(m))].$$

With assumption (5.5), it is easy to check that $q_i = 0$ for $i \in [1, m]$ or $[n-m+1, n]$.

Thus we have the following results for p_i and q_i:

$$g^T D^2 g = p^T p = \sum_{i=1}^{n} p_i^2 = \frac{1}{4N^2}[\lambda^2 O(\frac{m^5}{n^2}) + \rho^2 O(m^3) + \lambda\rho O(\frac{m^4}{n})], \quad (5.11)$$

$$h^T D^2 h = q^T q = \sum_{i=1}^{n} q_i^2 = 2\gamma \sum_{l=1}^{m} (\sum_{k=l}^{m} c_k)^2, \quad (5.12)$$

$$g^T D^2 h = p^T q = \sum_{i=1}^{n} p_i q_i = \frac{1}{4N^2}[\lambda^2 O(\frac{m^4}{n}) + \rho^2 O(m^3) + \lambda\rho O(\frac{m^4}{n})]. \quad (5.13)$$

Note also that

$$\begin{aligned} \sum_{l=1}^{m} (\sum_{k=l}^{m} c_k)^2 &= \frac{1}{4N^2}\Big[\lambda^2\big(\frac{2}{15}m^3 + o(m^3)\big) + \lambda\rho\big(-\frac{1}{5}nm^2 + o(nm^2)\big) \\ &\quad +\rho^2\big(\frac{6}{5}n^2 m + 12n^2 - \frac{7}{5}nm^2 + o(n^2) + o(nm^2)\big)\Big]. \end{aligned}$$

Therefore, by (5.11), (5.12) and (5.13), we have (III).

For (IV), it is an immediate result from Lemma 8 (a), (5.11) and (5.12). This finishes the proof of Lemma 9.

Proof of Theorem 8. By Lemma 9, we have the variance of $\hat{\theta}_{\text{new}}(\lambda, \rho)$ as

$$\begin{aligned} \text{var}[\hat{\theta}_{\text{new}}(\lambda,\rho)] &= (\text{var}(\varepsilon^2) - 2\sigma^4)\text{tr}\{\text{diag}(D)^2\} + 2\sigma^4 \text{tr}(D^2) + 4\sigma^2 f^T D^2 f \\ &+ 4\sigma^3 \mu_3 (f^T D \text{diag}(D)\mathbf{1}) \\ &= \lambda^2\Big[\frac{1}{n}\text{var}(\varepsilon^2) + \frac{m}{15n^2}\text{var}(\varepsilon^2) + \frac{4}{nm}\sigma^4 + \frac{4m}{15n^2}\sigma^2\gamma + o(\frac{1}{nm}) + o(\frac{m}{n^2})\Big] \\ &+ \rho^2\Big[(\frac{3}{5m} - \frac{11}{10n})\text{var}(\varepsilon^2) + (\frac{12n}{m^3} - \frac{18}{m^2} - \frac{72n}{m^4})\sigma^4 + (\frac{12}{5m} + \frac{24}{m^2} - \frac{2}{5n})\sigma^2\gamma \\ &\quad +o(\frac{1}{n}) + o(\frac{1}{m^2}) + o(\frac{n}{m^4})\Big] \\ &+ \lambda\rho\Big[-\frac{1}{10n}\text{var}(\varepsilon^2) - \frac{2}{5n}\sigma^2\gamma - \frac{12}{m^2}\sigma^4 + o(\frac{1}{n}) + o(\frac{1}{m^2})\Big]. \end{aligned}$$

Since λ and ρ are arbitrary, $\text{var}(\hat{\theta}_1)$ and $\text{var}(\hat{\theta}_2)$ can be derived directly from the above result. Using the conditions in Theorem 8, we have

$$\text{var}(\hat{\theta}_{\text{new},1}) = \frac{1}{n}\text{var}(\varepsilon^2) + \frac{m}{15n^2}\text{var}(\varepsilon^2) + \frac{4m}{15n^2}\sigma^2\gamma + o(\frac{m}{n^2}), \quad (5.14)$$

$$\text{var}(\hat{\theta}_{\text{new},2}) = \frac{3}{5m}\text{var}(\varepsilon^2) + \frac{12}{5m}\sigma^2\gamma + o(\frac{1}{m}). \quad (5.15)$$

Simple calculation shows that

$$\sum_{k=1}^{m} k^2 a_k w_k = -\frac{1}{6}m^2(1 + o(1)), \ \sum_{k=1}^{m} k^2 b_k w_k = nm(1 + o(1))$$

and

$$\sum_{k=1}^{m} a_k w_k = 1, \ \sum_{k=1}^{m} b_k w_k = 0, \ \sum_{k=1}^{m} \frac{k a_k w_k}{n-k} = 0, \ \sum_{k=1}^{m} \frac{k b_k w_k}{n-k} = 1.$$

This leads to the biases of $\hat{\theta}_{\text{new},1}$ and $\hat{\theta}_{\text{new},2}$ as

$$\begin{aligned} E(\hat{\theta}_{\text{new},1}) &= \sum_{k=1}^{m} a_k w_k E(s_k) = \sum_{k=1}^{m} a_k w_k (\theta_1 + \frac{k}{(n-k)}\theta_2 + \frac{k^2}{2n^2}\delta + o(\frac{k^2}{n^2})) \\ &= \theta_1 + \big[\delta \sum_{k=1}^{m} \frac{k^2 a_k w_k}{2n^2}\big](1 + o(1)) \\ &= \theta_1 - \frac{m^2}{12n^2}\delta + o(\frac{m^2}{n^2}) \end{aligned} \tag{5.16}$$

and

$$\begin{aligned} E(\hat{\theta}_{\text{new},2}) &= \sum_{k=1}^{m} b_k w_k E(s_k) = \sum_{k=1}^{m} b_k w_k (\theta_1 + \frac{k}{(n-k)}\theta_2 + \frac{k^2}{2n^2}\delta + o(\frac{k^2}{n^2})) \\ &= \theta_2 + \big[\delta \sum_{k=1}^{m} \frac{k^2 b_k w_k}{2n^2}\big](1 + o(1)) \\ &= \theta_2 + \frac{m}{2n}\delta + o(\frac{m}{n}). \end{aligned} \tag{5.17}$$

Finally, by (5.14), (5.15), (5.16) and (5.17), we can derive the MSEs of $\hat{\sigma}^2_{\text{new}}$ and $\hat{\gamma}_{\text{new}}$ as in Theorem 1. The derivation of MSEs for $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\gamma}_{\text{MS}}$ is similar and so is omitted.

5.5.2 Proof of Theorem 9

For simplicity, we give the proof for the case $\gamma = 0$ only as in Müller and Stadtmüller (1999). The proof for the case $\gamma \neq 0$ is similar and so is omitted. To derive the asymptotic normality for $(\hat{\theta}_{\text{new},1}, \hat{\theta}_{\text{new},2})^T$, by the Cramér-Wold device it is sufficient to show that for any pair of $(\lambda, \ \rho) \in R^2$, the following X_n is asymptotically normal:

$$X_n = \lambda\sqrt{n}(\hat{\theta}_{\text{new},1} - E(\hat{\theta}_{\text{new},1})) + \rho\sqrt{m}(\hat{\theta}_{\text{new},2} - E(\hat{\theta}_{\text{new},2})).$$

Let $\alpha_k = w_k a_k$ and $\beta_k = w_k b_k$ for $1 \le k \le m$. We rewrite X_n as follows:

$$\begin{aligned} X_n &= \lambda\sqrt{n}(\hat{\theta}_{\text{new},1} - E(\hat{\theta}_{\text{new},1})) + \rho\sqrt{m}(\hat{\theta}_{\text{new},2} - E(\hat{\theta}_{\text{new},2})) \\ &= \lambda\sqrt{n}(\sum_{k=1}^{m} \alpha_k s_k - \theta_1 + O(\frac{m^2}{n^2})) + \rho\sqrt{m}(\sum_{k=1}^{m} \beta_k s_k + O(\frac{m}{n})) \\ &= \sum_{k=1}^{m} (\lambda\sqrt{n}\alpha_k + \rho\sqrt{m}\beta_k) s_k - \lambda\sqrt{n}\sigma^2 + O(m^{3/2}/n) \\ &= \sum_{k=1}^{m} (\lambda\sqrt{n}\alpha_k + \rho\sqrt{m}\beta_k)(s_k - \sigma^2) + O(m^{3/2}/n) \\ &= \sum_{k=1}^{m} \omega_{k,n}\eta_k + O(m^{3/2}/n), \end{aligned}$$

where $\omega_{k,n} = (\lambda\sqrt{n}\alpha_k + \rho\sqrt{m}\beta_k)$ and $\eta_k = s_k - \sigma^2$. For $\omega_{k,n}$, it is easy to show that

$$\sum_{k=1}^{m} \omega_{k,n} \sim \lambda\sqrt{n} \qquad \text{and} \qquad \sum_{k=1}^{m} \omega_{k,n}^2 = O(\frac{n^2}{m^2}).$$

For η_k, we divide it into two parts: $\eta_k = \tilde{\eta}_{k,n} + r_{k,n}$, where

$$\begin{aligned} \tilde{\eta}_{k,n} &= \frac{1}{2(n-k)} \sum_{i=1}^{n-k} [(\varepsilon_{i+k} - \varepsilon_i)^2 - 2\sigma^2], \\ r_{k,n} &= \frac{1}{2(n-k)} \sum_{i=1}^{n-k} [(g_{i+k} - g_i)^2 + 2(g_{i+k} - g_i)(\varepsilon_{i+k} - \varepsilon_i)]. \end{aligned}$$

It is easy to check that

$$E(\sum_{k=1}^{m} \omega_{k,n} r_{k,n}) = O(m^{3/2}/n) \to 0. \tag{5.18}$$

In addition, we have

$$\begin{aligned} \text{cov}(r_{k,n}, r_{l,n}) &= \frac{1}{4(n-k)(n-l)} \sum_{i=1}^{n-k} \sum_{j=1}^{n-l} (g_{i+k} - g_i)(g_{j+l} - g_j) \text{cov}(\varepsilon_{i+k} - \varepsilon_i, \varepsilon_{j+h} - \varepsilon_j) \\ &= O(\frac{m^3}{n^2}) \frac{\sigma^2}{4(n-k)(n-l)} \\ &= O(\frac{m^3}{n^4}). \end{aligned}$$

Therefore,

$$\text{var}(\sum_{k=1}^{m} \omega_{k,n} r_{k,n}) = \sum_{k,l=1}^{m} \omega_{k,n}\omega_{l,n} \text{cov}(r_{k,n}, r_{l,n}) = O(\frac{m^3}{n^4}) \sum_{k,l=1}^{m} |\omega_{k,n}\omega_{l,n}| = o(1). \tag{5.19}$$

Define $\tilde{X}_n = \sum_{k=1}^{m} \omega_{k,n}\tilde{\eta}_{k,n}$. Then, we can show that

$$E(X_n - \tilde{X}_n)^2 = E(\sum_{k=1}^{m} \omega_{k,n} r_{k,n})^2 = \text{var}(\sum_{k=1}^{m} \omega_{k,n} r_{k,n}) + [E(\sum_{k=1}^{m} \omega_{k,n} r_{k,n})]^2 \to 0.$$

This implies that X_n and $\tilde{X}_n$ share the same asymptotic distribution. Hence, now we only focus on the simpler term, i.e. $\tilde{X}_n$. To access the asymptotic normality of $\tilde{X}_n$, we introduce the following statistic

$$\hat{T}_n = \sum_{\mu=1}^{n} E(\tilde{X}_n|\varepsilon_\mu).$$

Then by Lemmas 11 and 12 as described below, we complete the proof of Theorem 2.

Lemma 10. *Let* $c_{v,n} = \sum_{k=1}^{v-1} \omega_{k,n}/(n-k)$. *Assume that* $m \to \infty$ *and* $m/n \to 0$. *Then for* $1 \le v \le m$, *we have*

(i) $c_{m+1,n} = \lambda/\sqrt{n}(1 + o(1))$.

(ii) $\sum_{v=1}^{m} c_{v,n} = \lambda m/\sqrt{n}(1 + o(1)) - \rho\sqrt{m}(1 + o(1))$.

(iii) $c_{v,n} = \lambda\frac{\sqrt{n}(v-1)(4m-3v)}{nm^2}(1 + o(1)) + \rho\frac{6\sqrt{m}(v-1)(v-m)}{m^3}(1 + o(1))$.

(iv) $\sum_{v=1}^{m} c_{v,n}^2 = \frac{6}{5}\rho^2(1 + o(1))$.

(v) $\max_{1\le v\le m}\{|c_{v,n}|\} = O(m^{-1/2})$.

Proof of Lemma 10. This lemma can be readily achieved by Lemma 8 and the following four identities:

$$\begin{aligned}\sum_{k=1}^{v-1} \frac{\alpha_k}{n-k} &= \frac{v-1}{mn}(1 + o(1)) - \frac{m}{2n}\frac{6(v-1)(v-m)}{m^3}(1 + o(1)) \\ &= \frac{(v-1)(4m-3v)}{nm^2}(1 + o(1))\end{aligned}$$

and

$$\begin{aligned}\sum_{k=1}^{v-1} \frac{\beta_k}{n-k} &= \frac{1}{N(\sum_{k=1}^{m} w_k d_k^2 - \bar{d}_w^2)}[\frac{(v-1)^2}{2n}(1 + o(1)) - \frac{m(v-1)}{2n}(1 + o(1))] \\ &= \frac{6(v-1)(v-m)}{m^3}(1 + o(1))\end{aligned}$$

and

$$\begin{aligned}\sum_{v=1}^{m} c_{v,n} &= \lambda\sqrt{n}\sum_{v=1}^{m}\sum_{k=1}^{v-1}\frac{\alpha_k}{n-k} + \rho\sqrt{m}\sum_{v=1}^{m}\sum_{k=1}^{v-1}\frac{\beta_k}{n-k} \\ &= \frac{\lambda\sqrt{n}m^3}{nm^2}(2-1)(1+o(1)) + \frac{6\rho\sqrt{m}m^3}{m^3}(\frac{1}{3}-\frac{1}{2})(1+o(1)) \\ &= \lambda m/\sqrt{n}(1+o(1)) - \rho\sqrt{m}(1+o(1))\end{aligned}$$

and

$$\begin{aligned}\sum_{v=1}^{m} c_{v,n}^2 &= \sum_{v=1}^{m}\Big(\lambda\sqrt{n}\sum_{k=1}^{v-1}\frac{\alpha_k}{n-k} + \rho\sqrt{m}\sum_{k=1}^{v-1}\frac{\beta_k}{n-k}\Big)^2 \\ &= \sum_{v=1}^{m}\Big(\lambda^2 n(\sum_{k=1}^{v-1}\frac{\alpha_k}{n-k})^2 + \rho^2 m(\sum_{k=1}^{v-1}\frac{\beta_k}{n-k})^2 + 2\lambda\rho\sqrt{mn}(\sum_{k=1}^{v-1}\frac{\alpha_k}{n-k})(\sum_{k=1}^{v-1}\frac{\beta_k}{n-k})\Big) \\ &= \frac{6}{5}\rho^2(1+o(1)).\end{aligned}$$

Lemma 11. *For $\hat{T}_n$ and $\tilde{X}_n$, we have*

$$E(\hat{T}_n - \tilde{X}_n)^2 \to 0 \qquad \text{as } n \to \infty.$$

Proof of Lemma 11. Note that $E(\hat{T}_n) = E(\tilde{X}_n) = 0$ and $\text{var}(\hat{T}_n) \sim \text{var}(\tilde{X}_n) = O(1)$. We have

$$\begin{aligned}E(\hat{T}_n - \tilde{X}_n)^2 &= E(\hat{T}_n^2) + E(\tilde{X}_n^2) - 2E(\hat{T}_n\tilde{X}_n) \\ &\sim 2[\text{var}(\hat{T}_n) - E(\hat{T}_n\tilde{X}_n)].\end{aligned}$$

For ease of notation, let $s_\mu^2 = \varepsilon_\mu^2 - \sigma^2$. We then rewrite $\hat{T}_n$ as follows:

$$\begin{aligned}
\hat{T}_n &= \sum_{\mu=1}^{n}\sum_{k=1}^{m}\frac{\omega_{k,n}}{2(n-k)}\sum_{i=1}^{n-k}E[(\varepsilon_{i+k}-\varepsilon_i)^2 - 2\sigma^2|\varepsilon_\mu] \\
&= \sum_{\mu=1}^{n}\sum_{k=1}^{m}\frac{\omega_{k,n}}{2(n-k)}\sum_{i=1}^{n-k}(\delta_{i+k,\mu}+\delta_{i,\mu})s_\mu^2 \\
&= \sum_{\mu=m+1}^{n-m}\Big[\sum_{k=1}^{m}\frac{\omega_{k,n}}{n-k}\Big]s_\mu^2 + \sum_{\mu=1}^{m}\Big[\sum_{k=1}^{\mu-1}\frac{\omega_{k,n}}{n-k} + \sum_{k=\mu}^{m}\frac{\omega_{k,n}}{2(n-k)}\Big]s_\mu^2 \\
&+ \sum_{\mu=n-m+1}^{n}\Big[\sum_{k=1}^{n-\mu}\frac{\omega_{k,n}}{n-k} + \sum_{k=n-\mu+1}^{m}\frac{\omega_{k,n}}{2(n-k)}\Big]s_\mu^2 \\
&= c_{m+1,n}\sum_{\mu=m+1}^{n-m}s_\mu^2 + \frac{1}{2}\sum_{\mu=1}^{m}(c_{m+1,n}+c_{\mu,n})s_\mu^2 + \frac{1}{2}\sum_{\mu=1}^{m}(c_{m+1,n}+c_{\mu,n})s_{n-\mu+1}^2 \\
&= c_{m+1,n}\sum_{\mu=m+1}^{n-m}s_\mu^2 + \frac{1}{2}\sum_{\mu=1}^{m}c_{\mu,n}(s_\mu^2+s_{n-\mu+1}^2) + \frac{c_{m+1,n}}{2}\sum_{\mu=1}^{m}(s_\mu^2+s_{n-\mu+1}^2) \quad (5.20)
\end{aligned}$$

For $\hat{T}_n\tilde{X}_n$, with (5.20) we have

$$\begin{aligned}
E(\hat{T}_n\tilde{X}_n) &= E\Big[\big(c_{m+1,n}\sum_{\mu=m+1}^{n-m}s_\mu^2 + \frac{1}{2}\sum_{\mu=1}^{m}c_{\mu,n}(s_\mu^2+s_{n-\mu+1}^2) + \frac{c_{m+1,n}}{2}\sum_{\mu=1}^{m}(s_\mu^2+s_{n-\mu+1}^2)\big)\tilde{X}_n\Big] \\
&= c_{m+1,n}\sum_{\mu=m+1}^{n-m}E(s_\mu^2\tilde{X}_n) + \frac{1}{2}\sum_{\mu=1}^{m}c_{\mu,n}[E(s_\mu^2\tilde{X}_n)+E(s_{n-\mu+1}^2\tilde{X}_n)] \\
&\quad + \frac{c_{m+1,n}}{2}\sum_{\mu=1}^{m}[E(s_\mu^2\tilde{X}_n)+E(s_{n-\mu+1}^2\tilde{X}_n)].
\end{aligned}$$

Note that

$$\begin{aligned}
E(s_\mu^2\tilde{X}_n) &= E(s_\mu^2\sum_{k=1}^{m}\frac{\omega_{k,n}}{2(n-k)}\sum_{i=1}^{n-k}[(\varepsilon_{i+k}-\varepsilon_i)^2 - 2\sigma^2]) \\
&= \sum_{k=1}^{m}\frac{\omega_{k,n}}{2(n-k)}\sum_{i=1}^{n-k}E[s_\mu^2(\varepsilon_{i+k}-\varepsilon_i)^2] \\
&= (\mu_4-\sigma^4)\sum_{k=1}^{m}\frac{\omega_{k,n}}{2(n-k)}\sum_{i=1}^{n-k}(\delta_{\mu,i}+\delta_{\mu,i+k}).
\end{aligned}$$

For $m+1 \le \mu \le n-m$,

$$E(s_\mu^2\tilde{X}_n) = c_{m+1,n}(\mu_4-\sigma^4).$$

For $1 \le \mu \le m$ or $n-m+1 \le \mu \le n$,

$$E(s_\mu^2 \tilde{X}_n) = \frac{1}{2}(c_{m+1,n} + c_{\mu,n})(\mu_4 - \sigma^4).$$

Hence by Lemma 10,

$$\begin{aligned}
E(\hat{T}_n \tilde{X}_n) &= c_{m+1,n} \sum_{\mu=m+1}^{n-m} E(s_\mu^2 \tilde{X}_n) + \frac{1}{2}\sum_{\mu=1}^{m} c_{\mu,n}[E(s_\mu^2 \tilde{X}_n) + E(s_{n-\mu+1}^2 \tilde{X}_n)] \\
&\quad + \frac{c_{m+1,n}}{2} \sum_{\mu=1}^{m} [E(s_\mu^2 \tilde{X}_n) + E(s_{n-\mu+1}^2 \tilde{X}_n)] \\
&= (\mu_4 - \sigma^4)[(n-2m)c_{m+1,n}^2 + \frac{1}{2}\sum_{\mu=1}^{m} c_{\mu,n}^2 + c_{m+1,n}\sum_{\mu=1}^{m} c_{\mu,n} + \frac{m}{2}c_{m+1,n}^2] \\
&\sim (\mu_4 - \sigma^4)(\lambda^2 + \frac{3}{5}\rho^2) \\
&\sim \mathrm{var}(\hat{T}_n).
\end{aligned}$$

This shows that $E(\hat{T}_n \tilde{X}_n) \sim \mathrm{var}(\hat{T}_n)$ and so proves Lemma 11.

Lemma 12. *Assume that condition (5.5) holds, $m \to \infty$ and $L = m = n^r$ for $1/2 < r < 2/3$. For $\hat{T}_n$, we have*

$$\hat{T}_n \xrightarrow{D} N\big(0, \lambda^2(\mu_4 - \sigma^4) + \frac{3}{5}\rho^2(\mu_4 - \sigma^4)\big) \qquad \text{as } n \to \infty.$$

Proof of Lemma 12. For $\hat{T}_n$, we apply the following notations for ease of presentation

$$\begin{aligned}
\hat{T}_n &= c_{m+1,n} \sum_{\mu=m+1}^{n-m} s_\mu^2 + \frac{1}{2}\sum_{\mu=1}^{m} c_{\mu,n}(s_\mu^2 + s_{n-\mu+1}^2) + \frac{c_{m+1,n}}{2}\sum_{\mu=1}^{m}(s_\mu^2 + s_{n-\mu+1}^2) \\
&= P_{1,n} + P_{2,n} + P_{3,n}.
\end{aligned}$$

For $P_{1,n}$, by the central limit theorem and Lemma 10 (i), we have

$$P_{1,n} \xrightarrow{D} N(0, \lambda^2(\mu_4 - \sigma^4)).$$

For $P_{2,n}$, we have the following fact,

$$\max_{1 \le v \le m}\{|c_{v,n}|\} = O(m^{-1/2}) \to 0.$$

This implies that $\{c_{\mu,n} s_\mu^2\}$ satisfy the Lindeberg's condition (Billingsley; 2012). Thus,

$$\sum_{\mu=1}^{m} c_{\mu,n} s_\mu^2 / ((\mu_4 - \sigma^4)\sum_{v=1}^{m} c_{v,n}^2)^{1/2} \xrightarrow{D} N(0,1).$$

Note that $\sum_{v=1}^{m} c_{v,n}^2 \to 6\rho^2/5$. We have

$$\frac{1}{2}\sum_{\mu=1}^{m} c_{\mu,n} s_{\mu}^2 \xrightarrow{D} N(0, \frac{3}{10}\rho^2(\mu_4 - \sigma^4)).$$

By the same procedure, we have

$$\frac{1}{2}\sum_{\mu=1}^{m} c_{\mu,n} s_{n-\mu+1}^2 \xrightarrow{D} N(0, \frac{3}{10}\rho^2(\mu_4 - \sigma^4)).$$

For $P_{3,n}$, we have

$$\mathrm{Var}(P_{3,n}) = \frac{1}{2} m c_{m+1,n}^2 (\mu_4 - \sigma^4) = O(\frac{m}{n}) \to 0.$$

Finally, by Slutsky's Theorem and the fact that $P_{1,n}$, $P_{2,n}$ and $P_{3,n}$ are independent of each other, we have

$$\hat{T}_n \xrightarrow{D} N\big(0, \lambda^2(\mu_4 - \sigma^4) + \frac{3}{5}\rho^2(\mu_4 - \sigma^4)\big).$$

Chapter 6

Future Work

This chapter includes two topics for future research, among which the optimal-p difference sequences are specifically introduced and the difference-based estimates of variance function are discussed in general.

6.1 Optimal-p Difference Sequence

Difference sequences are widely used for estimating the residual variance in model (2.1). The optimal sequence and the ordinary sequence are the most popular ones. Dette et al. (1998) provided detailed investigation for the estimators based on the two types of sequences. Denote the p-th derivative of the regression function with $g^{(p)}(x)$ and let $g_i^{(p)} = g^{(p)}(x_i)$. Assume $J_p = \int_0^1 [g^{(p)}]^2 dx$ to be bounded, $p = 0, \ldots, 2r$. For a normal error case, the mean squared error of $\hat{\sigma}^2(d_{\text{opt}}(r))$ is approximated with

$$\frac{C^2(r)}{n^4}\left(J_1^2 + \frac{4\sigma^2}{n}J_2\right) + \frac{2 + 1/r}{n}\sigma^4,$$

where, $C(r) = (2r+1)(r+1)/12$. Accordingly, the mean squared error of $\hat{\sigma}^2_{\text{ord}}(r)$ is approximated by

$$\frac{1}{n^{4r}}\binom{2r}{r}^{-2}\left(J_r^2 + \frac{4\sigma^2}{n}J_{2r}\right) + \frac{2\sigma^4}{n}\binom{4r}{2r}\binom{2r}{r}^{-2}.$$

Asymptotically, the bias term is negligible and $\text{MSE}(\hat{\sigma}^2_{\text{opt}}(r))$ is always smaller than $\text{MSE}(\hat{\sigma}^2_{\text{ord}}(r))$. Whereas, for finite samples especially small sample sizes, $\text{MSE}(\hat{\sigma}^2_{\text{ord}}(r))$ can be smaller than $\text{MSE}(\hat{\sigma}^2_{\text{opt}}(r))$. As a rule of thumb, Dette et al. (1998) suggested to use the ordinary sequence if the sample size is small or the signal-to-noise ratio

is large; otherwise, the optimal sequence should be used. It seems that the question has been well solved. However, it deserves more consideration.

First, this rule of thumb can be useless in practice since it counts on some quantities related to the unknown mean function and the unknown residual variance, and also there is no clear threshold for recognizing the number of samples as small or large.

Second, even if we fortunately choose the right one from the optimal sequence and the ordinary sequence, it is not guaranteed that we make the most appropriate choice among all the order-r sequences. Essentially, it is the balance between the variance and the squared bias that motivates the definition of the two types of sequences. However, both sequences go to extremes of the trade-off area, only concentrating either on the variance term or on the squared bias term. It is quite possible that neither of them is real optimal for such a tradeoff.

The main goal of this section is to enlarge the space of existing difference sequence and provide more choices for the sequence selection problem, so as to increase the chance of finding the optimal variance-bias trade-off. To achieve this, we define new types of difference sequences as compromises of the optimal sequence and the ordinary sequence.

6.1.1 Definition of New Sequences

In this section, we provide definition for new kinds of difference sequences. We first treat the optimal and the ordinary sequence as the solutions to two specific optimization problems. Then, we generalize the optimization questions by varying the number of constrains. Eventually, we get a series of new difference sequences, which are the solutions to those generalized optimization questions.

As we discussed in Section 6.1, the trade-off between $\mathrm{var}(\hat{\sigma}^2)$ and $\mathrm{bais}^2(\hat{\sigma}^2)$ is the essential idea of the sequence selection problem. The asymptotic variance of (2.3) was derived as

$$\mathrm{var}(\hat{\sigma}^2) = \frac{1}{n}\left[\mathrm{var}(\varepsilon^2) + 4\sigma^4\delta(d)\right], \tag{6.1}$$

where $\delta(d) = \sum_{k=1}^{r}(\sum_{j=0}^{r-k} d_j d_{j+k})^2$ (Hall et al.; 1990). Also, it is easy to derive $E(\hat{\sigma}^2) = \sigma^2 + B(d)$, where

$$B(d) = \frac{1}{n-r}\sum_{i=1}^{n-r}(\sum_{j=0}^{r} d_j g_{j+i})^2$$

is the bias term.

In this chapter, we follow the assumptions about the mean function g in Dette et al. (1998). When the design points are equidistant, we have

$$\begin{aligned} B(d) &= \frac{1}{n-r}\sum_{i=1}^{n-r}\left[d_0 g_i + d_1\left(\sum_{p=0}^{r} g_i^{(p)}(1/n)^p/p!\right) + \cdots + d_r\left(\sum_{p=0}^{r} g_i^{(p)}(r/n)^p/p!\right) + o(\frac{1}{n^r})\right]^2 \\ &= \frac{1}{n-r}\sum_{i=1}^{n-r}\left[C_0 g_i + C_1 g_i'/n + \cdots + C_r g_i^{(r)}/n^r + o(\frac{1}{n^r})\right]^2, \end{aligned}$$

where we denote $C_p(d)$ with C_p for simplicity of presentation and

$$C_0 = \sum_{j=0}^{r} d_j \text{ and } C_p = \sum_{j=0}^{r} j^p d_j/p!,\ p = 1, ..., r.$$

Focusing only on the variance term, the optimal sequence is defined as the solution to the following optimization problem,

$$\min_{(d_0,\ldots,d_r)\in R^{r+1}} \delta(d), \text{ s.t. } C_0 = 0 \text{ and } \sum_{j=0}^{r} d_j^2 = 1. \tag{6.2}$$

The constrains in (6.2) are only the basic conditions for all the difference sequences, which means no further effort is made to control the bias term. As a result, $\sigma^2(d_{\text{opt}}(r))$ achieves the minimum asymptotic variance, e.i., $n^{-1}[\text{var}(\varepsilon^2) + r^{-1}\sigma^4]$ and the optimal sequence satisfies that $\sum_{j=0}^{r-k} d_j d_{j+k} = -(2r)^{-1},\ k = 1, \ldots, r$. While, the bias term $B(d_{\text{opt}}(r)) = C_1^2 J_1/n^2 + o(1/n^2)$ is of course not well controlled. In particular, refer to Dette et al. (1998) we know that $C_1(d_{\text{opt}}(r)) = (r+1)(2r+1)/12$, that is say $B(d_{\text{opt}}(r))$ increases quickly as along with the order r. Hence, when sample size is small or g is rough, the optimal estimator often suffers serious bias especially for the high order cases.

Corresponding with (6.2), we can also treat ordinary sequence as the solution to an optimization problem with form,

$$\min_{(d_0,\ldots,d_r)\in R^{r+1}} \delta(d), \text{ s.t. } C_0 = 0, \cdots, C_{r-1} = 0 \text{ and } \sum_{j=0}^{r} d_j^2 = 1. \tag{6.3}$$

With the ordinary sequence, the bias term $B(d_{\text{ord}}(r)) = C_r^2 J_r / n^{2r} + o(1/n^{2r})$. Note that the ordinary sequence is actually the unique solution to (6.3) and no more restrictions is admitted. That is say, $\sigma^2(d_{\text{ord}}(r))$ achieves the lowest level for the bias term among all the order-r difference sequences. As the price of that, the asymptotic variance, approximated with $n^{-1}[\text{var}(\varepsilon^2) + \sqrt{2\pi r}\sigma^4]$, increases as the order rises. Consequently, ordinary estimators of order $r \geq 4$ are rarely recommended in practice (Dette et al.; 1998).

According to (6.2) and (6.3), we find that different kinds of sequences are corresponding with different constrains. So, it is natural to generate new sequences through variation of the restrictions. Specifically, we define a series of new difference sequences as the solutions to the following optimization problems,

$$\min_{(d_0,\ldots,d_r)\in R^{r+1}} \delta(d), \text{ s.t. } C_0 = 0, \cdots, C_p = 0 \text{ and } \sum_{j=0}^{r} d_j^2 = 1, \tag{6.4}$$

where $0 \leq p \leq r-1$. We denote the solution sequences with $d_{(p,r)}$, named *optimal-p difference sequence* of order r.

For $\sigma^2(d_{(p,r)})$, we can easily obtain its asymptotic variance and asymptotic bias as follows,

$$\begin{aligned} \text{var}(\sigma^2(d_{(p,r)})) &= \frac{1}{n}\left[\text{var}(\varepsilon^2) + 4\sigma^4\delta(d_{(p,r)})\right], \\ \text{Bias}(\hat{\sigma}^2(d_{(p,r)})) &= B(d_{(p,r)}) = \frac{1}{n^{2(p+1)}} C_{p+1}^2 J_{p+1} + o(\frac{1}{n^{2(p+1)}}). \end{aligned}$$

Then, we can approximate the mean squared errors of $\hat{\sigma}^2(d_{(p,r)})$ with

$$\frac{1}{n}\left[\text{var}(\varepsilon^2) + 4\sigma^4\delta(d_{(p,r)})\right] + \frac{1}{n^{4(p+1)}} C_{p+1}^4 J_{p+1}^2. \tag{6.5}$$

Thus, we have generalized the existing difference sequence with the optimal-p sequences.

Theorem 10. *For equidistant design, the estimator $\hat{\sigma}^2(d_{(p,r)})$ is an unbiased estimator for the residual variance of Model (2.1) when $g(x)$ is a polynomial with order up to p.*

Theorem (10) can be easily derived from the result (6.5).

6.1.2 Properties of Optimal-p Sequence

Obviously, the newly proposed difference sequences include existing ones, e.i., $d_{(0,r)} = d_{\text{opt}}(r)$ and $d_{(r-1,r)} = d_{\text{ord}}(r)$. And they greatly enlarge the space of existing difference sequence by extend the two endpoints ($p = 0$ or $p = r - 1$) to the whole line ($0 \leq p \leq r - 1$), which will certainly increases our chance of finding the most appropriate sequence.

When $r \geq 3$ and $0 < p < r - 1$, $d_{(p,r)}$ can be regarded as a compromise of the optimal sequence and the ordinary sequence. First, $B(d_{(p,r)})$ is apparently between $B(d_{(0,r)})$ and $B(d_{(r-1,r)})$. Second, a larger number of constrains means that we search for the minimum value of $\delta(d)$ on a smaller space, and hence $\delta(d_{(0,r)}) \leq \delta(d_{(p,r)}) \leq \delta(d_{(r-1,r)})$. To show this relationship more clearly, we provide an example in Figure (6.1). The figure illustrates that when $r = 3$, $d_{(1,r)}$ begins to be separated with the ordinary sequence. Afterwards, unlike the raise of $\delta(d_{(r-1,r)})$, $\delta(d_{(1,r)})$ decreases along with the order r and gradually gets closer to $\delta(d_{(0,r)})$. $\delta(d_{(2,r)})$ also performs similarly. So, we may expect $\sigma^2(d_{(1,r)})$ or $\sigma^2(d_{(2,r)})$ process satisfying variance which is comparable with that of $\sigma^2(d_{(0,r)})$.

In addition to extending the family of difference sequences, the newly proposed sequences are also plausible for their robustness to the raise of the order. Originally, a large order r is not recommended by either the optimal sequence or the ordinary sequence for the reasons we mentioned in discussion of (6.2) and (6.3). While, the new sequences manage to overcome this restriction. On one hand, as we illustrated in Figure (6.1) the variance term of $d_{(p,r)}$ starts to go down with r at the point $r = p+1$. On the other hand, the bias term is reduced to be negligible and have not significant effect on the mean squared error when r is not too large, even it increases quickly as r raises.

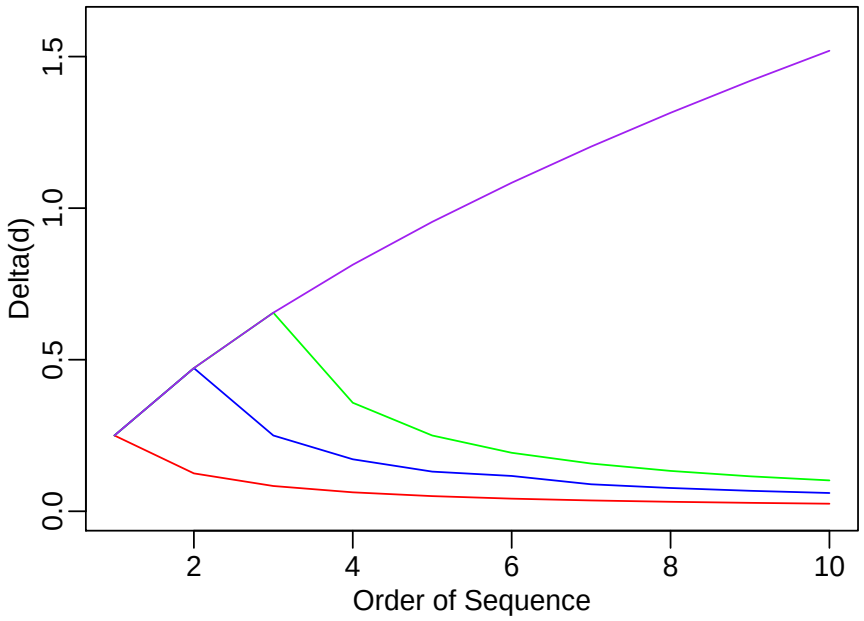

Figure 6.1: The trend of $\delta(d)$ for different kinds of sequences. The red line: $d_{(0,r)}$; the blue line: $d_{(1,r)}$; the green line: $d_{(2,r)}$; the purple line: $d_{(r-1,r)}$. $r = 1, \ldots, 10$.

We provide an example to compare the behaviors of several kinds of sequences in Figure (6.2). The settings are as follows. $n = 100$, $x = i/n$ are equally spaced, $g(x) = 3\sin(w\pi x)$, $w = 0, 1, 2, 4$ and ε come from the standard normal distribution. We illustrate the trend of mean squared errors for $\hat{\sigma}^2(d_{(0,r)})$, $\hat{\sigma}^2(d_{(1,r)})$, $\hat{\sigma}^2(d_{(2,r)})$ and $\hat{\sigma}^2(d_{(r-1,r)})$, $r = 1, \ldots, 10$. The mean squared errors are directly calculated with the approximation form (6.5).

For $w = 0$, we find the same pattern with Figure (6.1), and as the mean function getting oscillating, $\hat{\sigma}^2(d_{(0,r)})$ suffers more and more serious bias especially for high order cases. On the contrast, the other three estimators keep a robust performance across variation settings. Among the robust ones, $\text{mse}[\hat{\sigma}^2(d_{(1,r)})]$ and $\text{mse}[\hat{\sigma}^2(d_{(2,r)})]$ are always smaller than $\text{mse}[\hat{\sigma}^2(d_{(r-1,r)})]$. Besides the robustness, $\hat{\sigma}^2(d_{(1,r)})$ and $\hat{\sigma}^2(d_{(2,r)})$ usually provide the smallest mean squared errors as well. Hence, the new sequences are more reliable than existing ones for estimating the residual variance in Model (2.1). Further, we will show that the optimal choice is often found among the newly defined sequences.

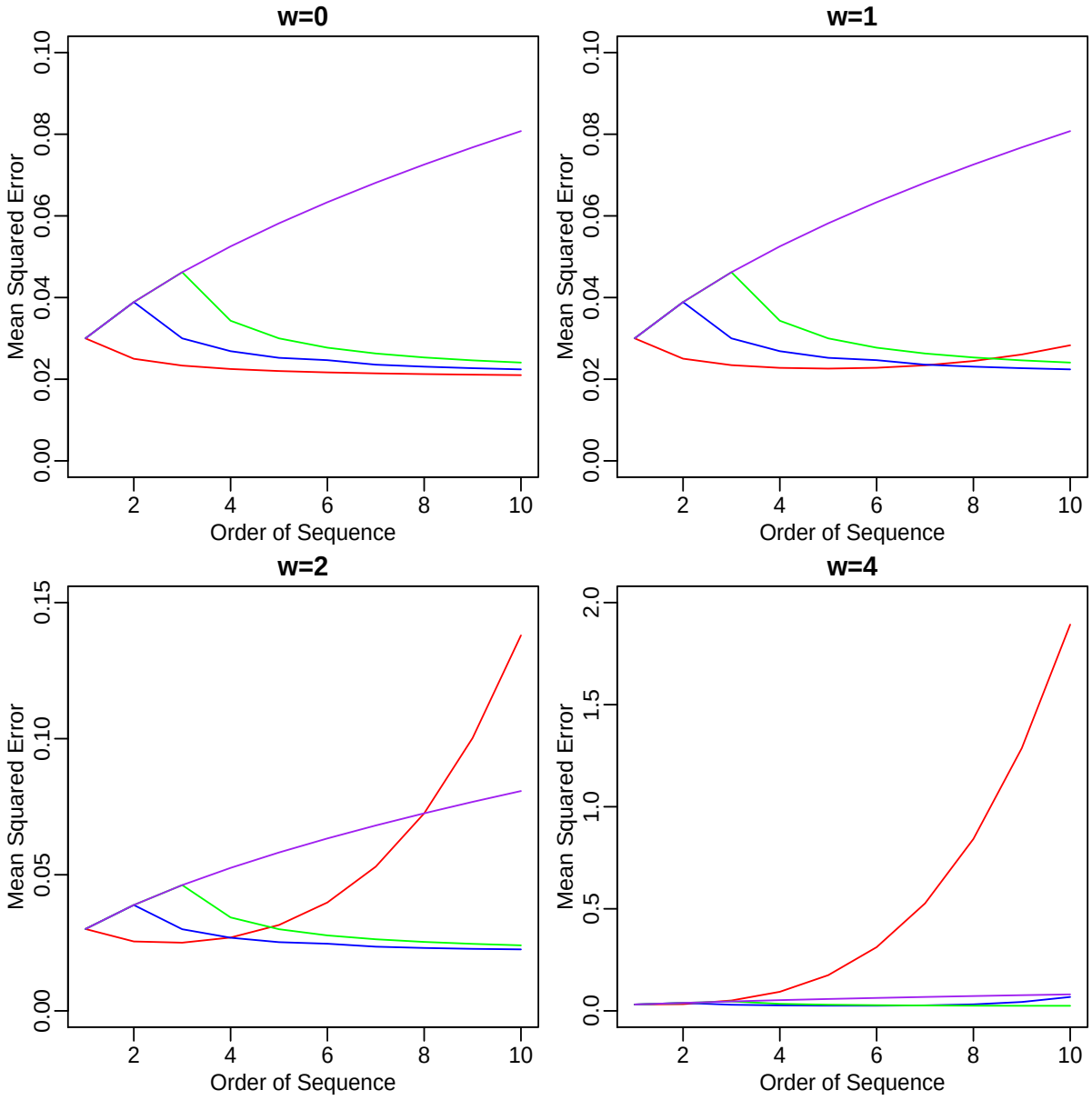

Figure 6.2: The mean squared errors of different kinds of estimators. The red line: $\hat{\sigma}^2(d_{(0,r)})$; the blue line: $\hat{\sigma}^2(d_{(1,r)})$; the green line: $\hat{\sigma}^2(d_{(2,r)})$; the purple line: $\hat{\sigma}^2(d_{(r-1,r)})$. $r = 1, \ldots, 10$.

6.2 Difference-based Variance Function Estimation

In many cases, the variance may change across the design points which results in a heteroscedastic model. The estimation of variance function in nonparametric regression models is also quite important in many contexts. On one hand, the variance

function estimates are needed to compute the weighted least squared estimates of the mean function and also to construct the confidence bands for the mean function. On the other hand, the variance function is of its own interest in many situations. One may assume the variance function as some specific parametric forms, such as a polynomial of the predictors, a power of the mean function, or assume the variance function as a totally nonparametric form.

The estimation of variance function has been widely studied in the literature. Among them, both residual-based methods and difference-based methods are included. For residual-based methods, Hall and Carroll (1989) applied kernel methods to estimate the variance function based on the squared residuals from a rate optimal estimator of the mean function, Ruppert et al. (1997) and Fan and Lin (1998) estimated the variance function by using local polynomial smoothing of the squared residuals from an "optimal" estimator of the mean function, Yuan and Wahba (2004) and Liu et al. (2007) investigated the smoothing spline methods, among others. For difference-based methods, Müller and Stadtmüller (1987) and Brown and Levine (2007) considered difference based kernel estimators of the variance function, Cai et al. (2008) applied a wavelet thresholding approach to adaptive variance function estimation, among others.

For homogenous cases, we have shown that the choice of difference sequence is quite critical for the performance of the variance estimates. Hence, it is natural to investigate the effect of the difference sequence on the estimates of the variance function which has not been studied in the literature. One future direction is to compare different estimates of variance function constructed with various kinds of sequence so as to select suitable ones for different scenarios.

Bibliography

Benko, M., Härdle, W. and Kneip, A. (2009). Common functional principal components, *The Annals of Statistics* **37**: 1–34.

Billingsley, P. (2012). *Probability and Measure, 3rd ed*, Wiley.

Bock, M., Bowman, A. W. and Ismail, B. (2007). Estimation and inference for error variance in bivariate nonparametric regression, *Statistics and Computing* **17**: 39–47.

Bowman, A., Jones, M. and Gijbels, I. (1998). Testing monotonicity of regression, *Journal of Computational and Graphical Statistics* **7**: 489–500.

Bowman, A. W., Bock, M. and Ismail, B. (2006). Detecting discontinuities in nonparametric regression curves and surfaces, *Statistics and Computing* **16**: 377–390.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Springer.

Brown, L. D. and Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method, *The Annals of Statistics* **35**: 2219–2232.

Brown, R. L., Durbin, J. and Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time, *Journal of the Royal Statistical Society B* **37**: 149–192.

Buckley, M. J. and Eagleson, G. K. (1989). A graphical method for estimating the residual variance in nonparametric regression, *Biometrika* **76**: 203–210.

Buckley, M. J., Eagleson, G. K. and Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression, *Biometrika* **75**: 189–199.

Cai, T. T., Levine, M. and Wang, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design, *Journal of Multivariate Analysis* **100**: 126–136.

Cai, T. T., Wang, L. et al. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression, *The Annals of Statistics* **36**: 2025–2054.

Cheng, M. Y., Peng, L. and Wu, J. S. (2007). Reducing variance in univariate smoothing, *The Annals of Statistics* **35**: 522–542.

Cheng, M. Y. and Raimondo, M. (2008). Kernel methods for optimal change-points estimation in derivatives, *Journal of Computational and Graphical Statistics* **17**: 56–75.

Chitty, L. S., Campbell, S. and Altman, D. G. (1993). Measurement of the fetal mandible - feasibility and construction of a centile chart, *Prenatal Diagnosis* **13**: 749–756.

Chu, C. K., Siao, J. S., Wang, L. C. and Deng, W. S. (2012). Estimation of 2D jump location curve and 3D jump location surface in nonparametric regression, *Statistics and Computing* **22**: 17–31.

Cobb, G. W. (1978). The problem of the nile: Conditional solution to a changepoint problem, *Biometrika* **65**: 243–251.

Dette, H. and Hetzler, B. (2009). A simple test for the parametric form of the variance function in nonparametric regression, *Annals of the Institute of Statistical Mathematics* **61**: 861–886.

Dette, H. and Munk, A. (1998). Testing heteroscedasticity in nonparametric regression, *Journal of the Royal Statistical Society, Series B* **60**: 693–708.

Dette, H., Munk, A. and Wagner, T. (1998). Estimating the variance in nonparametric regression - what is a reasonable choice?, *Journal of the Royal Statistical Society, Series B* **60**: 751–764.

Du, J. and Schick, A. (2009). A covariate-matched estimator of the error variance in nonparametric regression, *Journal of Nonparametric Statistics* **21**: 263–285.

Eagleson, G. K. (1989). Curve estimation – whatever happened to the variance?, *Bulletin of the International Statistical Institute* **53**: 535–551.

Eggermont, P. and LaRiccia, V. (2000). Maximum likelihood estimation of smooth monotone and unimodal densities, *The Annals of Statistics* **28**: 922–947.

Einmahl, J. H. and Van Keilegom, I. (2008). Tests for independence in nonparametric regression, *Statistica Sinica* **18**: 601–615.

Eubank, R. L. and Speckman, P. L. (1994). Nonparametric estimation of functions with jump discontinuities, *IMS Lecture Notes: Change-Point Problems* **23**: 130–144.

Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques, *Journal of the American Statistical Association* **85**: 387–392.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, Chapman and Hall.

Fan, J. and Lin, S. (1998). Test of significance when data are curves. Unpublished Manuscript.

Gasser, T., Kneip, A. and Kohler, W. (1991). A flexible and fast method for automatic smoothing, *Journal of the American Statistical Association* **86**: 643–52.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression, *Biometrika* **73**: 625–633.

Ghosal, S., Sen, A. and Van Der Vaart, A. W. (2000). Testing monotonicity of regression, *The Annals of Statistics* **28**: 1054–1082.

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

Gijbels, I., Prosdocimi, I. and Claeskens, G. (2010). Nonparametric estimation of mean and dispersion functions in extended generalized linear models, *Test* **19**: 580–608.

Grégoire, G. and Hamrouni, Z. (2002). Two non-parametric tests for change-point problems, *Journal of Nonparametric Statistics* **14**: 87–112.

Hall, P. and Carroll, R. J. (1989). Variance functin estimation in regression: the effect of estimating the mean, *Journal of the Royal Statistical Society, Series B* **51**: 3–14.

Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions, *The Annals of Statistics* **28**: 20–39.

Hall, P., Kay, J. and Titterington, D. (1991). On estimation of noise variance in two-dimensional signal processing, *Advances in Applied Probability* **23**: 476–495.

Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression, *Biometrika* **77**: 521–528.

Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression, *Biometrika* **77**: 415–419.

Hall, P. and Titterington, D. M. (1992). Edge-preserving and peak-preserving smoothing, *Technometrics* **34**: 429–440.

Härdle, W. and Kneip, A. (1999). Testing a regression model when we have smooth alternatives in mind, *Scandinavian Journal of Statistics* **26**: 221–238.

Härdle, W. and Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression, *Journal of Econometrics* **81**: 223–242.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression, *Biomatrika* **56**: 495–504.

Joo, J. H. and Qiu, P. (2009). Jump detection in a regression curve and its derivative, *Technometrics* **51**: 289–305.

Kariya, T. and Kurata, H. (2004). *Generalized Least Squares*, Wiley.

Kim, H. J. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression, *Biometrika* **76**: 409–423.

Kulasekera, K. B. and Gallagher, C. (2002). Variance estimation in nonparametric multiple regression, *Communications in Statistics - Theory and Methods* **31**: 1373–1383.

Levine, M. (2006). Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach, *Computational Statistics & Data Analysis* **50**: 3405–3431.

Liitiäinen, E., Corona, F. and Lendasse, A. (2010). Residual variance estimation using a nearest neighbor statistic, *Journal of Multivariate Analysis* **101**: 811–823.

Liu, Q., Dinu, I., Adewale, A., Potter, J. and Yasui, Y. (2007). Comparative evaluation of gene-set analysis methods, *BMC Bioinformatics* **8**: 431.

Loader, C. R. (1996). Change point estimation using nonparametric regression, *The Annals of Statistics* **24**: 1667–1678.

McDonald, J. and Owen, A. (1986). Smothing with split linear fits, *Technometrics* **28**: 195–208.

McElroy, F. W. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased, *Journal of the American Statistical Association* **62**: 1302–1304.

Müller, H. (1992). Change-points in nonparametric regression analysis, *The Annals of Statistics* **20**: 737–761.

Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis, *The Annals of Statistics* **15**: 610–635.

Müller, H. G. and Stadtmüller, U. (1988). Detecting dependencies in smooth regression models, *Biometrika* **75**: 639–650.

Müller, H. G. and Stadtmüller, U. (1993). On variance function estimation with quadratic forms, *Journal of Statistical Planning and Inference* **35**: 213–231.

Müller, H. and Stadtmüller, U. (1999). Discontinuous versus smooth regression, *The Annals of Statistics* **27**: 299–337.

Müller, U., Schick, A. and Wefelmeyer, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic, *Statistics* **37**: 179–188.

Munk, A., Bissantz, N., Wagner, T. and Freitag, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional, *Journal of the Royal Statistical Society, Series B* **67**: 19–41.

Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory, *The Annals of Statistics* **26**: 2339–2368.

Neumeyer, N. and Van Keilegom, I. (2009). Change-point tests for the error distribution in non-parametric regression, *Scandinavian Journal of Statistics* **36**: 518–541.

Paige, R., Sun, S. and Wang, K. (2009). Variance reduction in smoothing splines, *Scandinavian Journal of Statistics* **36**: 112–126.

Pendakur, K. and Sperlich, S. (2010). Semiparametric estimation of consumer demand systems in real expenditure, *Journal of Applied Econometrics* **25**: 420–457.

Qiu, P. (2005). *Image Processing and Jump Regression Analysis*, Wiley.

Qiu, P. (2007). Jump surface estimation, edge detection, and image restoration, *Journal of the American Statistical Association* **102**: 745–756.

Qiu, P. and Hawkins, D. (2001). A rank-based multivariate cusum procedure, *Technometrics* **43**: 120–132.

Qiu, P. and Yandell, B. (1998). A local polynomial jump detection algorithm in nonparametric regression, *Technometrics* **40**: 141–152.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*, Springer.

Rice, J. A. (1984). Bandwidth choice for nonparametric regression, *The Annals of Statistics* **12**: 1215–1230.

Rose, C. and Smith, M. D. (2002). *Mathematical Statistics with Mathematica*, Springer.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling, *Applied Statistics* **43**: 429–453.

Ruppert, D., Wand, M. P., Holst, U. and HöSJER, O. (1997). Local polynomial variance-function estimation, *Technometrics* **39**: 262–273.

Seifert, B., Gasser, T. and Wolf, A. (1993). Nonparametric estimation of residual variance revisited, *Biometrika* **80**: 373–383.

Shen, H. and Brown, L. D. (2006). Non-parametric modelling of time-varying customer service times at a bank call centre, *Applied Stochastic Models in Business and Industry* **22**: 297–311.

Sim, C. H., Gan, F. F. and Chang, T. C. (1994). Outlier labeling with boxplot procedures, *Journal of the American Statistical Association,* **100**: 642–652.

Tong, T., Liu, A. and Wang, Y. (2008). Relative errors of difference-based variance estimators in nonparametric regression, *Communications in Statistics - Theory and Methods* **37**: 2890–2902.

Tong, T., Ma, Y. and Wang, Y. (2013). Optimal variance estimation without estimating the mean function, *Bernoulli* **19**: 1839–1854.

Tong, T. and Wang, Y. (2005). Estimating residual variance in nonparametric regression using least squares, *Biometrika* **92**: 821–830.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics With S, 4th ed*, Springer.

von Neumann, J. (1941). Distribution of the ratio of the mean squared successive difference to the variance, *The Annals of Mathematical Statistics* **12**: 367–395.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

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

Wang, L., Brown, L. D. and Cai, T. (2011). A difference based approach to the semiparametric partial linear model, *Electronic Journal of Statistics* **5**: 619–641.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets, *Biomatrika* **82**: 385–397.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*, Chapman and Hall, New York.

Whittle, P. (1964). On the convergence to normality of quadratic forms in independent variables, *Theory of Probability and Its Applications* **9**: 103–108.

Wu, J. S. and Chu, C. K. (1993a). Kernel type estimators of jump points and values of a regression function, *The Annals of Statistics* **21**: 1545–1566.

Wu, J. S. and Chu, C. K. (1993b). Nonparametric function estimation and bandwidth selection for discontinuous regression functions, *Statistica Sinica* **3**: 557–576.

Xu, Q. and You, J. (2007). Difference-based estimation for error variances in repeated measurement regression models, *Statistics and Probability letter* **77**: 811–816.

Yatchew, A. (1997). An elementary estimator of the partial linear model, *Economics Letters* **57**: 135–143.

Yatchew, A. (1999). An elementary nonparametric differencing test of equality of regression functions, *Economics Letters* **62**: 271–278.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association* **93**: 120–131.

You, J., Zhou, X. and Zhou, Y. (2010). Statistical inference for panel data semiparametric partially linear regression models with heteroscedastic errors, *Journal of Multivariate Analysis* **101**: 1079–1101.

Yuan, M. and Wahba, G. (2004). Doubly penalized likelihood estimator in heteroscedastic regression, *Statistics and Probability Letters* **69**: 11–20.

Curriculum Vitae

Academic qualifications of the thesis author, Mr. DAI Wenlin:

- Received the degree of Bachelor of Science (Statistics) from Beijing Institute of Technology, July 2008.
- Received the degree of Master of Science (Probability and Mathematical Statistics) from Beijing Institute of Technology, July 2010.

August 2014