2017

NONPARAMETRIC COMPOUND ESTIMATION, DERIVATIVE ESTIMATION, AND CHANGE POINT DETECTION

Sisheng Liu
University of Kentucky, sishengliu1989@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0001-7888-1387

Digital Object Identifier: https://doi.org/10.13023/ETD.2017.312

Sisheng Liu, Student

Dr. Richard Charnigo, Major Professor

Dr. Constance Wood, Director of Graduate Studies

NONPARAMETRIC COMPOUND ESTIMATION, DERIVATIVE ESTIMATION, AND CHANGE POINT DETECTION

DISSERTATION

A dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of
Philosophy in the College of Arts and Sciences
at the University of Kentucky

By
Sisheng Liu
Lexington, Kentucky
Director: Richard Charnigo, Professor of Statistics
Lexington, Kentucky
2017

ABSTRACT OF DISSERTATION

NONPARAMETRIC COMPOUND ESTIMATION, DERIVATIVE ESTIMATION, AND CHANGE POINT DETECTION

Firstly, we reviewed some popular nonparameteric regression methods during the past several decades. Then we extended the compound estimation (Charnigo and Srinivasan [2011]) to adapt random design points and heteroskedasticity and proposed a modified Cp criteria for tuning parameter selection. Moreover, we developed a *DCp* criteria for tuning paramter selection problem in general nonparametric derivative estimation. This extends *GCp* criteria in Charnigo, Hall and Srinivasan [2011] with random design points and heteroskedasticity. Next, we proposed a change point detection method via compound estimation for both fixed design and random design case, the adaptation of heteroskedasticity was considered for the method. Finally, we applied our change point detection method to a glucose level data set and identified the meal consumption time for five patients.

KEYWORDS: Nonparametric regression, Compound estimation, Derivative estimation, Tuning parameter selection, Change point detection

Author's signature: Sisheng Liu

Date: July 21, 2017

NONPARAMETRIC COMPOUND ESTIMATION, DERIVATIVE ESTIMATION, AND CHANGE POINT DETECTION

By

Sisheng Liu

Director of Dissertation: Richard Charnigo

Director of Graduate Studies: Constance Wood

Date: July 21, 2017

ACKNOWLEDGMENTS

First of all, I deeply express my gratitude to my advisor, Dr.Richard Charnigo, who has guided me from every aspect of the academic research consistently and back me for future career development. Without his help, I won't be able to write down the dissertation and graduate in time.

Secondly, I would like to thank my RA advisor Dr.Chi Wang for thorough guidance and advice of his own research experience. What he had past to me is not just the knowledge in books, but more importantly the precious experience and the upright attitute towards research.

Moreover, I should thank all other committees, Dr. Cidambi Srinivasan, Dr. Arnold J. Stromberg, Dr.Derek S. Young for their advices with profound statistical background.

Also, I thank all the friends in Department of Statistics for a nice life trip and very friendly study and research environment.

Finitely, many thanks to my parients who have brought me up and support/love me during my whole life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 A review of nonparametric regression

1.1 Purpose and rationales for nonparametric regression

Suppose $\mu(X)$ is the mean response function for the regression model

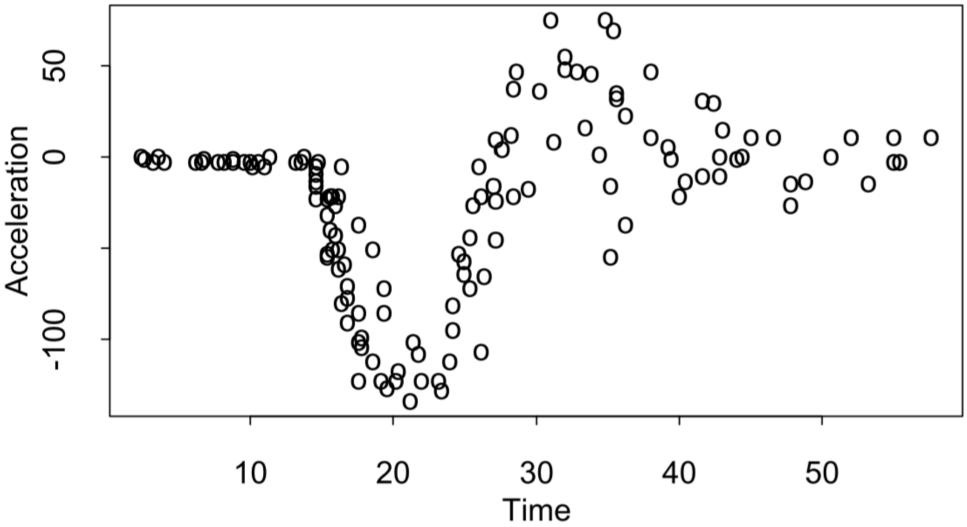
$$Y_i = \mu(X_i) + \epsilon_i \tag{1.1}$$

for $i \in \{1, 2, ..., n\}$ and X_i's belong to a compact interval $\mathcal{X} \subset R$ or compact set $\mathcal{X} \subset R^d$ and ϵ_i's are independent zero-mean random errors with variance bounded above by some constant. For a regression problem, our goal is to predict Y from X given a new sample X. The optimal predictor will be $\mu(X)$.

In parametric regression, we need to pre-specify the determinant form of the mean response function, like linear regression, polynomial regression. However, in our nature, the patterns of data are sometimes very complex, it's hard to make assumption of the parametric form. Thus, we may need to apply the nonparametric regression method to make predictions. Nonparametric regression is constructed based on the information that is purely driven by the data. It does not assume a parametric form of the function in advance. Figure 1.1 displays the measurements of the acceleration of a motorcycle that runs into a solid object. Loader [2006]. There are 133 observations in the motorcycle data. Covariate X here (in milliseconds) is the time that after a simulated impact on motorcycles. Response variable is the head acceleration of a test object. Fan and Gijbels [1996]. The scatter plot of motor cycle data obviously displays nonlinearity and heteroskedasticity. If we fit linear regression and polynomial regression for the data, then the fitted line will be too smooth for the data. Fan and Gijbels [1996]. In this case, nonparametric regression will be a more appropriate method for modeling the relationship between the time and acceleration. Also, we need to use some methods for dealing with the non-constant error variance in the data.

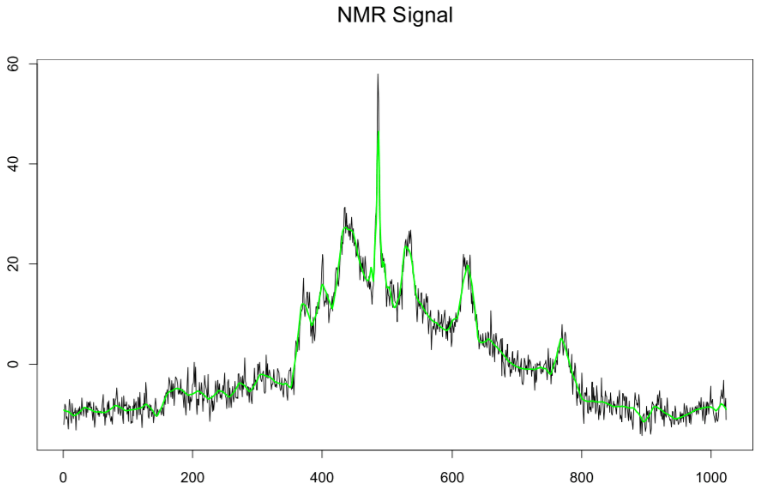
There are several rationales for the nonparametric regression. The first one is trying to get the information from the data points that are close to the target point, such as kernel regression, local polynomial regression, regression trees [Breiman, 1984]. Orthogonal series spline, regression spline and wavelets method try to approximate the true mean function by linear combination of a series of basis functions. Basis functions in a funtional space are used a lot for representing a specific function. Also, there are some other nonparametric methods which are not commonly used, such as sieve estimator, etc.

Figure 1.1: Motor cycle plot. Head acceleration vs. time



Different nonparametric regression methods are developed for various applications. There is no method that can over compete any other nonparametric regression methods. For example, local polynomial regression can deal with the functions that are not very smooth by adaptively choose local bandwidth. However, if the function have too many spikes, wavelets method [Daubechies, 1992], which is used a lot for signal process, may be much better than the local polynomial regression. Wavelets bases is able to represent locally bumby functions in an efficient way. For example, Figure 1.2 is the plot of the NMR Signal data (the black line), it is fitted by the wavelet shrink method via the green line. Friedman et al. [2001]

Figure 1.2: NMR Signal data fit with Wavelets



1.2 Kernel method and Local polynomial regression

Kernel estimator was proposed by Nadaraya [1964] and Watson [1964]. It uses kernel function to control how the nearby data points impact on the prediction point. If a symmetric kernel function $W(x) \geq 0$ satisfy

$$\int W(x)dx = 0 \tag{1.2}$$

$$\int xW(x)dx = 0 \tag{1.3}$$

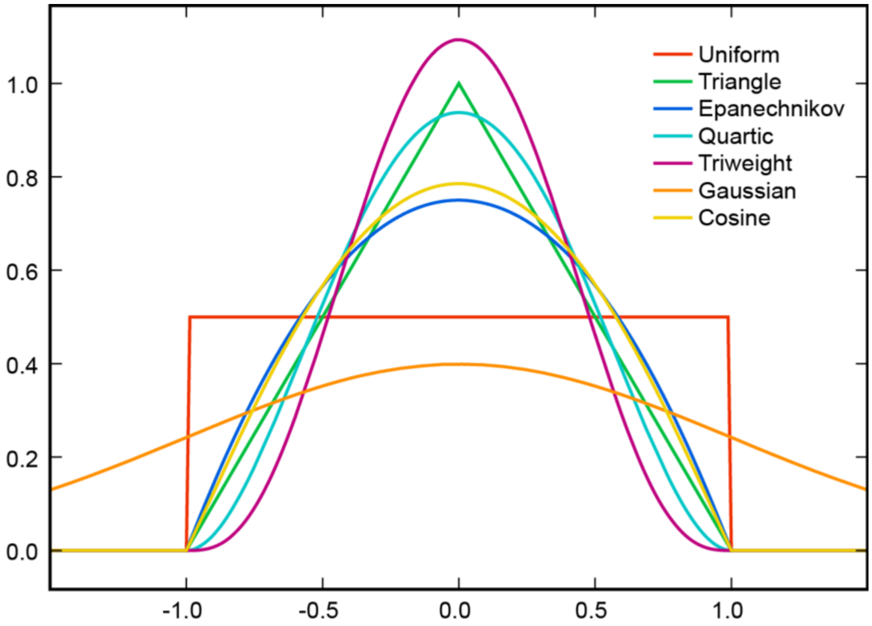
$$\int x^2W(x)dx > 0 \tag{1.4}$$

Then the Nadaraya-Watson kernel estimator is defined as

$$\widehat{\mu(x)} = \frac{\sum_{i=1}^{n} W(|x - X_i|/h)Y_i}{\sum_{i=1}^{n} W(|x - X_i|/h)}. \tag{1.5}$$

Bandwidth h controls the smoothness of the estimation function. if h is large, then many sample points will be used for predicting $\widehat{\mu(x)}$ for a given x. The prediction will borrow "information" from those sample points that have nonzero weights in the formula of Nadaraya-Watson kernel estimator. Bandwidth choice is critical for the kernel regression and the local polynomial regression, it controls the bias-variance trade off for these methods. A too small bandwidth will lead to a wiggly plot with large variance, a large bandwidth will undersmooth the data points and give large bias. The kernel function help us to decide somehow the importance of information from each sampling points. If x_i is in the neighborhood of a prediction point x, then generally, the closer that x to x_i, the larger the value of kernel function. Figure 1.3 shows us different kernel functions. The weight from a given sampling point for the prediction at x is affected by the choice of the function. Gasser and Müller [1979] proposed a modified kernel regression method. It is defined as

Figure 1.3: Kernel functions



$$\mu(x) = h^{-1} \sum_{i=1}^{n} \left[\int_{s_{i-1}}^{s_i} K(\frac{x - X_i}{h}) \right] Y_i \tag{1.6}$$

where $s_i = \frac{x_{i-1}+x_i}{2}$.

The kernel estimator is a linear estimator, which means it can be written as

$$\widehat{\mu(x)} = \sum_{i=1}^{n} l_i(x) Y_i \tag{1.7}$$

Then the kernel derivative estimation is obtain by taking the derivative of $\widehat{\mu(x)}$ respect to x directly. However, the performance of derivative estimation of kernel estimator depends on the smoothness of the kernel function.

It is shown that kernel estimator could be attained by minimizing the weighted squared loss, which is the local linear estimation. Loader [2006]

$$\sum_{i=1}^{n} W(|x - X_i|/h)(Y_i - \widehat{\mu(x)})^2 \tag{1.8}$$

For a close to x, we can write $\mu(a) \approx \beta_0(x) + \beta_1(x)(a - x) + ... + \frac{1}{p!}\beta_p(x)(X_i - x)$, then we minimize

$$\sum_{i=1}^{n} W(|x - X_i|/h)[Y_i - (\beta_0(x) + \beta_1(x)(X_i - x) + ... + \frac{1}{p!}\beta_p(x)(X_i - x))]^2 \tag{1.9}$$

to get $\widehat{\beta_0(x)}$ and $\widehat{\beta_1(x)}$.

Then the local polynomial estimator is

$$\widehat{\mu(x)} = \widehat{\beta_0(x)}. \tag{1.10}$$

The local polynomial regression has a good property that it has the same convergence rate in both interior points and the boundary. (This was shown by [Fan and Gijbels, 1992]). However, the local polynomial derivative estimator is not self-consistent. Namely, the derivative of the local polynomial estimator is not equal to the estimator of the local polynomial derivative. We will define the self-consistency formly later. In Loader [2006], he shows that the local derivative estimator is actually the local slope estimate.

$$\widehat{\mu'(x)} = \widehat{\beta_1(x)} \tag{1.11}$$

This is not the derivative of the fitted curve $\widehat{\mu(x)}$. The exact derivative of the fitted curve is given in the book as (1.12),

$$\widehat{\mu'(x)} = \widehat{\beta_1(x)} + \boldsymbol{e_1}^T(\boldsymbol{X^T W X})^{-1}\boldsymbol{X}^T\boldsymbol{W'}\boldsymbol{\hat{\epsilon}} \tag{1.12}$$

where $\boldsymbol{e_1} = (1, 0, 0..., 0)^T$, $\boldsymbol{X}$ is the design matrix and $\boldsymbol{W}$ is the matrix with kernel functions $K((x - X_i)/h)$, but it is computationally infeasible.

1.3 Other nonparametric regression methods

Orthogonal series estimator/regression spline [Ruppert et al., 2003] assumes that the true underline function can be represented by a linear combination of some orthonormal basis functions $b_j(x), j = 1, 2, 3...$, namely

$$\mu(x) = \sum_{j=1}^{\infty} \beta_j b_j(x). \tag{1.13}$$

Then we need to estimate β_j for the specific series of basis function. Also, a finite number of basis will be used to avoid overfitting of the mean response. The orthogonal series estimator is

$$\widehat{\mu(x)} = \sum_{j=1}^{J} \hat{\beta}_j b_j(x), \tag{1.14}$$

and $\hat{\beta}_j$ is obtained by regress Y on the set of J basis functions. That is to minimize

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i - \sum_{j=1}^{J} \beta_j b_j(x_i)]^2. \tag{1.15}$$

Denote $\boldsymbol{B}_{ij} = b_j(x_i)$, then we obtain the estimation of $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = (\boldsymbol{B^T B})^{\boldsymbol{-1}} \boldsymbol{B^{-1} Y}. \tag{1.16}$$

If we differentiate the estimated mean response function (1.14) with respect to x, then we'll get the derivative estimator. However, some basis functions may not be smooth enough for making the derivative estimation. Also, we need to choose number of basis function and knot and the place of knots. Figure 1.4 is an example of 6 B-spline basis functions with order 3. Regression splines minimize

$$\sum_{i=1}^{n} (Y_i - \mu(X_i))^2 + \lambda \int \mu^{''2} dx \tag{1.17}$$

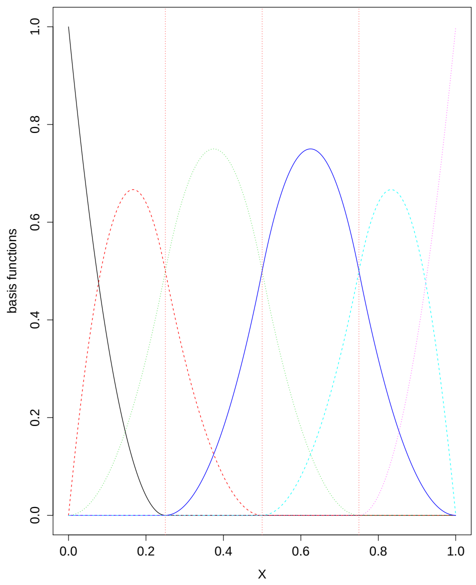


Figure 1.4: B-spline basis functions

to get $\widehat{\mu(x)}$. The regularization part is for controlling the smoothness of the fitted function. Tuning parameter λ is often called smoothing parameter. The interesting part about (1.17) is that the optimization problem has a finite-dimensional solution even if it is minimized over a infinite functional space. The solution of (1.17) is the natural cubic splines with knots at the values of design points $x_i, i = 1, ..., N$. Friedman et al. [2001].

1.4 Compound estimation

A derivative estimator is called self-consistent if

$$\frac{d^r \widehat{\mu(x)}}{dx} = \frac{\widehat{d^r \mu(x)}}{dx} \tag{1.18}$$

Recovering derivatives is important for data analysis in some scenario. For example, like studying human growth data, characterizing nanoparticles from scattering data etc. The local derivative estimator is not self-consistent, and many spline basis functions or kernel functions are not infinitely differentiable. Charnigo and Srinivasan

[2011] developed a compound estimator to simultaneously estimate mean response function and it Jth derivatives if we assume $\mu(x)$ have at least $J+1$ derivatives. The compound estimator is constructed by 2 steps. Firstly, a finite number of pointwise estimators on grid points are obtained by some well-known nonparametric regression methods. Then these pointwise estimators are combined by a weight function to make sure the estimated mean responses are differentiable by at least J times. Compound estimation will be introduced in the next section.

1.5 Modified compound estimation and parameter choice

A modified version of Compound estimation is proposed in chapter 2. In chapter 2, we no longer assume $x's$ are fixed points, but random design points. Also, we incorporate heteroskedasaticity in model (1.1) for the error term ϵ_i by modifying the weight function (2.4) in the compound estimator. A Cp criteria is used for choosing the tuning parameters for the modified compound estimation with heteroskedasaticity. Cp criteria does not require a specific nonparametric regression method. It could be used as long as the fitted model is a linear estimator like (1.7). It is motivated from Tsybakov [2008]. Cp criteria tries to minimize the following quantity by a proxy which satisfy (1.20).

$$E[\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu(X_i)}-\mu(X_i))^2] \tag{1.19}$$

$$E[Cp(\lambda)] = E[\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu_\lambda(X_i)}-\mu(X_i))^2]+C \tag{1.20}$$

1.6 An DCp criterion for tuning parameter selection of first derivative estimation

In chapter 3, we will aim to minimize the quantity (1.21) to choose our tuning parameters instead of using Cp criteria in chapter 2. The idea is from Charnigo, Hall and Srinivasan [2011]. They developed a GCp criteria to minimize (1.21). However,our method is different since they assume the $x's$ are equally spaced and to be fixed design. It is hard to find a appropriate proxy for (1.21) when the samples are from random design.

$$DDIMSE = E[\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu'_\lambda(X_i)} - \mu'(X_i))^2] \tag{1.21}$$

1.7 Estimation of piecewise smooth functions

In chapter 4&5, we'll try to address the problem when the underline mean response function is piecewise smooth. Then we need to detect the change point and fit the model on each side of the jump location. For example, Figure 1.5 and Figure 1.6 show us the scatter plot of observations and the real underline mean response function $\mu(x)$. We let $\mu(x) = I(x \leq 0)(x+1)^2 + I(x > 0)(-x^2 + 2x + 2)$. Then we generate $X \sim Unif(-1, 1)$ and let $Y = \mu(X) + \epsilon$ where $\epsilon \sim N(0, 0.5^2)$. The function is discontinuous with a change point at $x = 0$. Gijbels [2008] gives a brief review of how to use local linear fitting when the mean response curve may have some irregularities. Müller et al. [1999] developed a method for decide whether the unknown function should be modeled as a globally smooth function or a piecewise function with isolated discontinuities.

Figure 1.5: scatter plot of the observations.

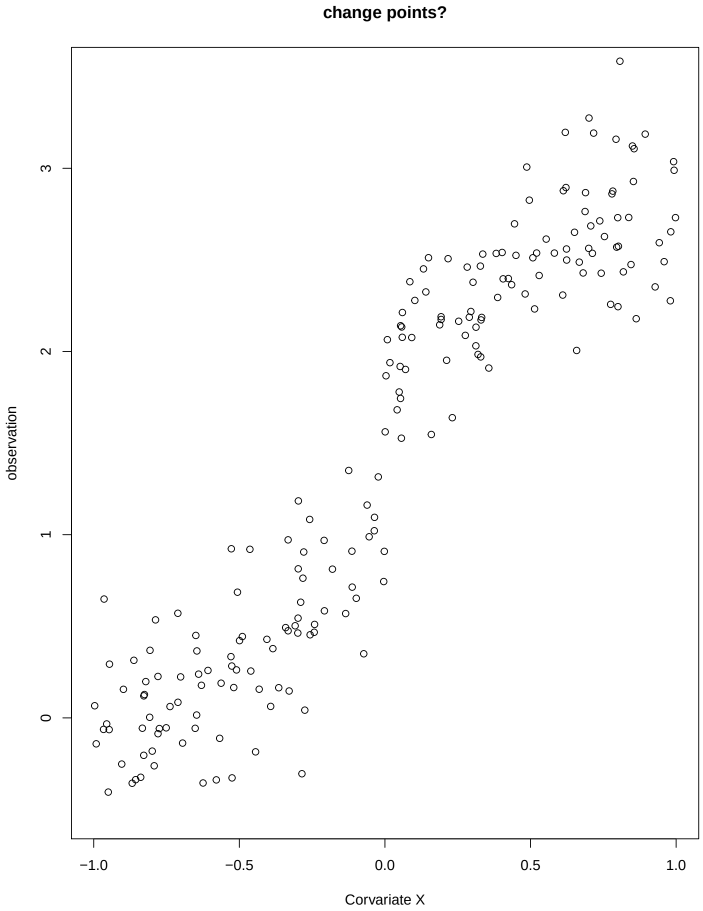
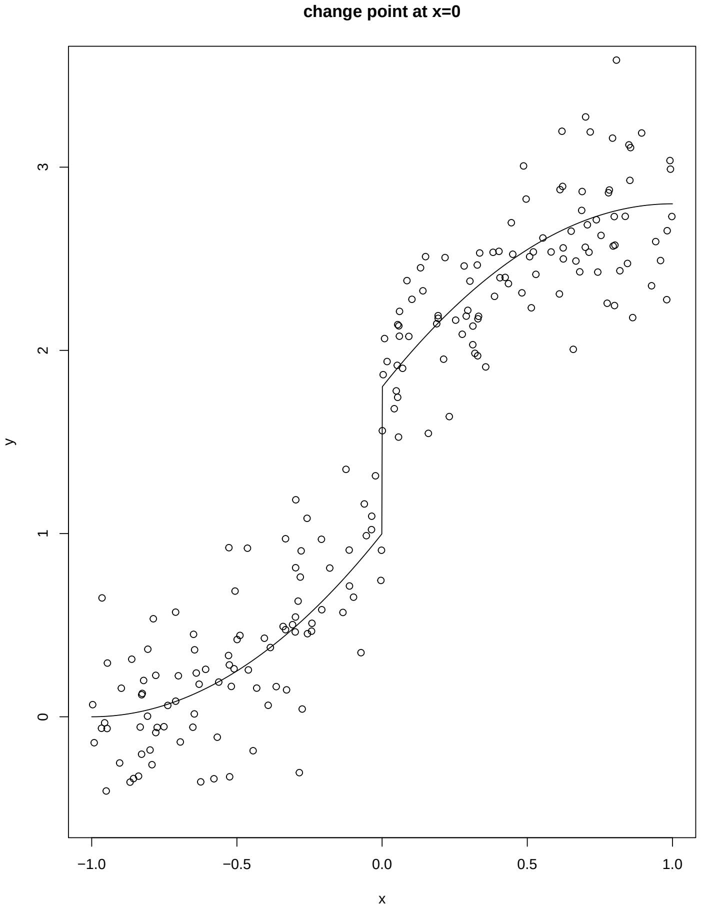


Figure 1.6: Function $\mu(x)$ has a discontinuity $x = 0$.



Chapter 2 Better convolution weights of compound estimation with or without heteroskedasticity

2.1 Review of compound estimator

Suppose $\mu(x)$ is the mean response function for the nonparameric regression model

$$Y_i = \mu(x_i) + \epsilon_i \tag{2.1}$$

for $i \in \{1, 2, ..., n\}$ and x_i's belong to a compact interval $\mathcal{X} \subset R$ and ϵ_i's are independent zero-mean random errors with variance bounded above by some constant. $\mu(x)$ has at least $J + 1$ continuous derivatives.

As in Charnigo and Srinivasan [2011], we assume $\mathcal{X} := [-1, 1]$ without loss of generality. For constructing compound estimator $\mu^\star(x)$ of $\mu(x)$, first we estimate the mean function and its derivative at only a discrete set of points in $[-1, 1]$, denoted by I_n. The estimators are denoted by $\tilde{c}_{j;a}$, where $0 \leq j \leq J$ and a is one of the points in the discrete set I_n. Then a polynomial is defined as

$$\tilde{\mu}_{J;a}(x) := \sum_{j=0}^{J} \tilde{c}_{j;a}(x - a)^j. \tag{2.2}$$

The compound estimator for $\mu(x)$ is then defined as a weighted average of (2.2),

$$\mu^\star(x) := \sum_{a \in I_n} W_{a,n}(x) \tilde{\mu}_{J;a}(x), \tag{2.3}$$

where $W_{a,n}$ is the weight for point $a \in I_n$. In Charnigo and Srinivasan [2011], the weight is

$$W_{a,n}(x) := \frac{exp[-\beta_n(x - a)^2]}{\sum_{c \in I_n} exp[-\beta_n(x - c)^2]}. \tag{2.4}$$

The idea for this weight is to give $\tilde{\mu}_{J;a}(x)$ more weight if a is close to x and less weight if a is away from x. The convolution in (2.3) will ensure the compound estimator is infinitely differentiable and self-consisent. Also,the compound estimation of jth derivative $\mu^{(j)}(x)$ will be

$$\frac{d^j}{dx^j}\mu^\star(x) := \sum_{a\in I_n}\sum_{k=0}^{j}\binom{j}{k}\frac{d^k}{dx^k}\tilde{\mu}_{J;a}(x)\frac{d^{j-k}}{dx^{j-k}}W_{a,n}(x), \tag{2.5}$$

and because differentiation and estimation can be interchanged, we say that the compound estimator is self-consistent.

2.2 Mathematical formulation of optimization problem

For the model (2.1), we assume that the variance of random error ϵ's are bounded above and below by some known positive constant. Furthermore, let

$$var(\epsilon_i) = \sigma_i^2, \tag{2.6}$$

assume the design points are random: $X_1, X_2, ...X_n$ instead of $x_1, ...x_n$ in (2.1). Intuitively, the compound estimator from (2.2) or (2.5) should be adjusted to account for non-constant variance. In the first step, the pointwise estimator $\tilde{c}_{j;a}$ can be some nonparametric estimator with adaptation to non-constant variance. Moreover, the compound estimator can be refined by choosing the weight function(2.4) to adapt to the heteroskedasticity in the sense of minimizing integrated mean square error(IMSE). In statistics, generally we give less weight to the observations which have larger variance, like weighted least square estimator.

We adjust the weight function by

$$W_{a,n}(x) := \frac{exp[-\beta_n(a)(x-a)^2]}{\sum_{c\in I_n} exp[-\beta_n(c)(x-c)^2]} \tag{2.7}$$

and $\beta_n(a)$ is some function of a discrete set of points $a \in I_n$. $\beta_n : I_n \to R^+$. The motivation for this adaptation is somehow intuitive.Firstly, weights for each polynomial can be adjusted by the variance of random error near each a. This is controlled by choosing the function $\beta_n(a)$ such that if the variance of random error near a is large, then we assign less weight to $\tilde{\mu}_{J;a}(x)$. By minimize the integrated MSE, we will be able to get the optimal choice of $\beta_n(a)$. On the other hand, weight function (2.7) inherits the nice property of (2.4). They still vary in x smoothly and the near optimal convergence rate can be achieved under some appropriate assumptions.

The compound estimator requires some mild conditions on the pointease estimators. In Charnigo and Srinivasan [2011], they assume $\tilde{c}_{j;a}$ satisfies

$$\sup_{a \in I_n} |\tilde{c}_{j;a}| \leq C \tag{2.8}$$

and

$$\sup_{a \in I_n} MSE[\tilde{c}_{j;a}] \leq Cn^{-2\alpha_j} \tag{2.9}$$

for some constant $C, \alpha_0, \alpha_1, ...\alpha_J$. Based on Stone [1980], to fulfill (2.9), the distribution of $X_1 ... X_n$ should satisfy that their distribution is absolutely continuous and their densities are bounded away from 0.

2.3 Solution and justification

The following theorem is an improved version of Theorem 1 in Charnigo and Srinivasan [2011] because random covariates are included and the methodology now explicitly adhering heteroskedasticity.

Theorem 2.3.1 *Suppose the model (2.1) hold, and the design points $X_1, ...X_n$ are random. Consider compound estimator from (2.3), (2.5) with the weight function (2.7). Also the pointwise estimators satisfied (2.8) and (2.9). If there exist positive* *numbers* $\delta, \gamma, \phi, \omega_0, ...\omega_J$ *such that*

$$\delta + 2w_j\gamma + \phi < 2\alpha_j \qquad for \qquad 0 \leq j \leq J$$

and

$$0 \leq \alpha_j - \alpha_{j+1} \leq \gamma \qquad for \qquad 0 \leq j \leq J-1$$

then there exist sequence $\beta_n = \{\beta_n(a) : a \in I_n\}$ *and* L_n *such that*

$$\frac{d^j}{dx^j}\mu^\star(x) - \mu^{(j)}(x) = O_p\left(n^{(3j+1)\delta + max\{j-J+1, max_{k\in\{0,1,...j\}}(k-w_{j-k})\}}\right).$$

The proof of the theorem is similar to Charnigo and Srinivasan [2011]. Note, however, that the dimension of β_n increases with the sample size.

Proof:

Let $\beta_{0n}(a)$ be the evaluation set at a function $\beta_{0n} : I_n \to [\frac{1}{M}, M]$. Choose $\beta_n(a)$ to satisfy

$$\beta_n(a) = \beta_{0n}(a)n^{2(\gamma+\delta)}$$

Suppose $[-1, 1]$ was divided to L_n intervals such that the ratio of maximum interval length to the minimum interval length is bounded above and the mid points in each interval form the set I_n. Choose $L_n = O(n^{(\gamma+\delta+\psi)})$.

Set $I_{1n}(x) := \{a \in I_n : |a - x| < n^{-\gamma}\}$, $I_{2n}(x) := \{a \in I_n : |a - x| < \beta_n(a)^{-1/2}\}$, $\beta^\star_{0n}(a) = max\{\beta_{0n}(a)\}$ and $\beta^\star_n = \beta^\star_{0n}n^{2(\gamma+\delta)}$). Then we have

$$\begin{aligned}\sum_{c\in I_n} exp[-\beta_n(c)(x-c)^2] &\geq \sum_{c\in I_{2n}} exp[-1] \\ &\geq C_1 L_n(\beta^\star_n)^{-1/2}\end{aligned}$$

In Chapter 2 and elsewhere, C's denote positive constants. Let $E_{J,a}(x) := \tilde{\mu}_{J,a}(x) - \mu(x)$, $A_{J,a}(x) := \tilde{\mu}_{J,a}(x) - \sum_{j=0}^{J} c_{j,a}(x-a)^j$, $T_{J,a}(x) := A_{J,a}(x) - E_{J,a}(x)$. Then $T_{J,a}(x)\mu(x) - \sum_{j=0}^{J}(x-a)^j$.

By (2.7) and mathematical induction, we have

$$\frac{d^k}{dx^k}W_{a,n}(x) = \frac{\sum_{z_k \in I_n} \dots \sum_{z_1 \in I_n} S_a(x) \prod_{i=1}^{k} S_{z_i}(x)[P_a(x) + P_{z_1}(x) \dots + P_{z_{i-1}}(x) - iP_{z_i}(x)]}{(\sum_{c \in I_n} exp[-\beta_n(c)(x-c)^2])^{k+1}} \tag{2.10}$$

where $S_z(x) = exp[-\beta_n(z)(x-z)^2]$ and $P_z(x) = 2\beta_n(z)(z-x)$.

Similar to the proof in Charnigo and Srinivasan [2011],we consider $I_n(x)$ and $\overline{I}_n(x)$ separately.

Firstly, if $a \in \overline{I}_{1n}(x)$, then by (2.10) and (2.11), there exist C_2 and $C_2^{'}$ such that for every $a \in \overline{I}_{1n}(x)$ and $0 \leq k \leq j \leq J$.

$$\begin{aligned} \left| \frac{d^k}{dx^k} W_{a,n}(x) \right| &\leq \frac{L_n^k C_2 exp^{[-\beta_{0n}(a)n^{2\delta}]}(\beta_n^\star)^k}{[L_n(\beta_n^\star)^{-1/2}]^{k+1}} \\ &= \frac{C_2(\beta_n^\star)^{(3k+1)/2} exp^{[-\beta_{0n}(a)n^{2\delta}]}}{L_n} \end{aligned} \tag{2.11}$$

Also,if $a \in \overline{I}_{1n}(x)$, then

$$\left| \frac{d^{j-k}}{dx^{j-k}} E_{J;a}(x) \right| \leq \left| \frac{d^{j-k}}{dx^{j-k}} A_{J;a}(x) \right| + \left| \frac{d^{j-k}}{dx^{j-k}} T_{J;a}(x) \right| \leq C_3 \tag{2.12}$$

Combining (2.12) and (2.13) we can see

$$\begin{aligned} \left| \sum_{a \in \overline{I}_{1n}(x)} \frac{d^k}{dx^k} W_{a,n}(x) \frac{d^{j-k}}{dx^{j-k}} E_{J;a}(x) \right| &\leq C_3 C_2 (\beta_n^\star)^{(3k+1)/2} exp^{[-\beta_{0n}(a)n^{2\delta}]} \\ &= C_3 C_2 (\beta_{0n}^\star)^{(3k+1)/2} n^{(3k+1)(\gamma+\delta)} exp^{[-\beta_{0n}(a)n^{2\delta}]} \\ &= C_4 n^{(3k+1)(\gamma+\delta)} exp^{[-\beta_{0n}(a)n^{2\delta}]} \\ &= o_p\left(n^{(3j+1)\delta + \max\{j-J+1, \max_{k \in \{0,1,\dots j\}}(k-w_{j-k})\}}\right) \end{aligned} \tag{2.13}$$

Secondly,if $a \in I_{1n}(x)$:

1. If $max\{|z_1 - x|, \dots |z_k - x|\} \geq n^{-\gamma}$,then each summand on top of (2.11) will be dominated by $C_5 n^{2k(\gamma+\delta)} exp^{[-(a)n^{2\delta}]}$ for some constant C_5.

2. If $max\{|z_1 - x|, \dots |z_k - x|\} \leq n^{-\gamma}$,then each summand on top of (2.11) will

be dominated by $C_6' \prod_{i=1}^{k}(\beta_n(z_k)n^{-\gamma}) = C_6 n^{2k\delta+k\gamma}$ for some constant C_6. Also, the number of these summands is $O(L_n^k n^{-k\gamma})$. Thus,

$$\begin{aligned}
\sup_{a\in I_{1n}(x)} \left|\frac{d^k}{dx^k}W_{a,n}(x)\right| &= \frac{O(C_6 n^{2k\delta+k\gamma}L_n^k n^{-k\gamma}) + O(C_5 n^{2k(\gamma+\delta)} exp^{[-(a)n^{2\delta}]} L_n^k)}{L_n^{k+1}(\beta_n^\star)^{-(k+1)/2}} \\
&= \frac{O(n^{2k\delta})}{L_n(\beta_n^\star)^{-(k+1)/2}} \\
&= O\left(\frac{n^{(3k+1)(\gamma+\delta)-2k\gamma}}{L_n}\right) \qquad (2.14)
\end{aligned}$$

Then follow (2.16),(2.19) and (2.20) in Charnigo and Srinivasan [2011], we conclude that

$$\begin{aligned}
&\left|\sum_{a\in I_{1n}(x)} \frac{d^k}{dx^k}W_{a,n}(x)\frac{d^{j-k}}{dx^{j-k}}E_{J;a}(x)\right| \\
&\leq \text{card}(I_{1n}(x))O\left(\frac{d^k}{dx^k}W_{a,n}(x)\right)\left[O_p\left(\frac{d^{j-k}}{dx^{j-k}}A_{J,a}(x)\right) + O\left(\frac{d^{j-k}}{dx^{j-k}}T_{J,a}(x)\right)\right] \\
&\leq L_n n^{-\gamma} O\left(\frac{n^{(3k+1)(\gamma+\delta)-2k\gamma}}{L_n}\right)\left[O_p\left(n^{-w_{j-k}\gamma}\right) + O\left(n^{-\gamma(J+1+k-j)}\right)\right] \\
&= O(n^{3k\delta+\delta+k\gamma})\left[O_p\left(n^{-w_{j-k}\gamma}\right) + O\left(n^{-\gamma(J+1+k-j)}\right)\right] \\
&\leq O_p\left(n^{(3j+1)\delta+\max\{j-J+1,\max_{k\in\{0,1,\ldots j\}}(k-w_{j-k})\}}\right) \qquad (2.15)
\end{aligned}$$

Since finitely many terms of the form $\left|\sum_{a\in I_{1n}(x)} \frac{d^k}{dx^k}W_{a,n}(x)\frac{d^{j-k}}{dx^{j-k}}E_{J;a}(x)\right|$ constitute $\sum_{a\in I_{1n}(x)} \frac{d^j}{dx^{j-k}W_{a,n}(x)E_{J;a}(x)}$, then (2.15) implies Theorem 2.3.1 holds. Then follow the proof of the Corollary 1 in Charnigo and Srinivasan [2011], under (2.8) and (2.9),we have

$$\frac{d^j}{dx^j}\mu^\star(x) - \mu^{(j)}(x) = o_p\left(n^{-(J+1-j)/(2J+3)+\nu}\right) \qquad (2.16)$$

for $0 \leq j \leq J$,$x \in (-1,1)]$ if we choose the $\alpha_0, \ldots\alpha_J$,$\omega_0, \ldots\omega_J$,δ,ϕ appropriately. ν is a arbitrary small number.

Stone [1980] showed that $n^{-(J+1-j)/(2J+3)}$ is the optimal rate for the nonparametric estimator for jth derivative. Therefore, the previous theorem tells us that no matter what function $\beta_{0n}(a)$ is, and with considerable flexibility regarding the weight functions, the near optimal rate can still hold for the compound estimator. However, we want not only favourable asymptotics but also finite sample performance. So our next step is to choose the optimal weight function (2.7), namely, choosing function $\beta_{0n}(a)$. This could in principle be done to minimize the integrated MSE, $E\left[\int(\mu^\star(x)-\mu(x))^2dx\right]$. However the MISE is hard to quantify for compound estimation, which may be based on different nonparametric regression methods for local polynomial estimation. Actually, no method can analytically quantity the MISE exactly. Local regression and Kernel method have an asymptotically formula for quantifying the MISE, but they depend on the unknown quantity $\mu''(x)$. Also, integrated MSE approximately vary across different nonparametric regression estimators and compound estimator is based on them. Therefore, instead we will minimize a discretized version of integrated MSE(Tsybokav 2009 CITE!!!!!!). The discretized version of Integrated MSE is define as (2.18). We then apply a Cp criterion which accounts for heteroskedaticity to approach our goal of minimizing (2.18)

Theorem 2.3.2 *Assume model (2.1) and (2.6) hold, and the estimator of $\mu(x)$ has the form*

$$\mu^\star(x)=\sum_{i=1}^{n}G_{n,i}(X_i,x,\boldsymbol{s})Y_i, \tag{2.17}$$

Where $\boldsymbol{s}$ is the parameters and $X_1,\ldots X_n$ are the random design points.We denote $G_{n,i}(X_i,X_i,\boldsymbol{s})$ as $G_{n,i}(X_i,\boldsymbol{s})$ for simplicity. Define

$$DMISE:=E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mu^\star(X_i)-\mu(X_i)\right)^2\right] \quad \textit{and} \tag{2.18}$$

$$C_p(\boldsymbol{s}):=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu^\star(X_i))^2+\frac{2}{n}\sum_{i=1}^{n}\sigma_i^2G_{n,i}(X_i,\boldsymbol{s}). \tag{2.19}$$

Then

$$E\left(C_p(\boldsymbol{s})\right)=DMISE+\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \tag{2.20}$$

Proof: let

$$Q = \frac{2}{n} E\left(\sum_{i=1}^{n} \mu^{\star}(X_i)\mu(X_i)\right) \tag{2.21}$$

and

$$\hat{Q} = \frac{2}{n}\sum_{i=1}^{n} Y_i \mu^{\star}(X_i) - \frac{2}{n}\sum_{i=1}^{n} \sigma_i^2 G_{n,i}(X_i, \boldsymbol{s}) \tag{2.22}$$

We'll have

$$\begin{aligned}
&\frac{n}{2}E(\hat{Q}|X_1, ...X_n) \\
&= E\left[\sum_{i=1}^{n} Y_i\mu^{\star}(X_i)\middle| X_1, ...X_n\right] - \sum_{i=1}^{n}\mu^{\star}(X_i)\mu(X_i) + \sum_{i=1}^{n}\mu^{\star}(X_i)\mu(X_i) - \sum_{i=1}^{n}\sigma_i^2 G_{n,i}(X_i, \boldsymbol{s}) \\
&= E\left[\sum_{i=1}^{n} \epsilon_i\mu^{\star}(X_i)\middle| X_1, ...X_n\right] + \sum_{i=1}^{n}\mu^{\star}(X_i)\mu(X_i) - \sum_{i=1}^{n}\sigma_i^2 G_{n,i}(X_i, \boldsymbol{s}) \\
&= E\left[\sum_{i=1}^{n}\sum_{k=1}^{n} \epsilon_i\epsilon_k G_{n,k,i}(X_k, X_i, \boldsymbol{s})\middle| X_1, ...X_n\right] + \sum_{i=1}^{n}\mu^{\star}(X_i)\mu(X_i) - \sum_{i=1}^{n}\sigma_i^2 G_{n,i}(X_i, \boldsymbol{s}) \\
&= \sum_{i=1}^{n} G_{n,i}(X_i, \boldsymbol{s}) E\left[\epsilon_i^2\middle| X_1, ...X_n\right] + \sum_{i=1}^{n}\mu^{\star}(X_i)\mu(X_i) - \sum_{i=1}^{n}\sigma_i^2 G_{n,i}(X_i, \boldsymbol{s}) \\
&= \sum_{i=1}^{n}\mu^{\star}(X_i)\mu(X_i)
\end{aligned} \tag{2.23}$$

By (2.23)

$$E(\hat{Q}) = E(Q) \tag{2.24}$$

Therefore by(2.24)

$$
\begin{aligned}
&DMISE \\
&= \frac{1}{n}E\sum_{i=1}^{n}\mu^\star(X_i)^2 + \frac{1}{n}E\sum_{i=1}^{n}\mu(X_i)^2 - \frac{2}{n}E\sum_{i=1}^{n}\mu(X_i)\mu^\star(X_i) \\
&= \frac{1}{n}E\sum_{i=1}^{n}\mu^\star(X_i)^2 + \frac{1}{n}E\sum_{i=1}^{n}\mu(X_i)^2 - \frac{2}{n}E\sum_{i=1}^{n}Y_i\mu^\star(X_i) + \frac{2}{n}E\sum_{i=1}^{n}\sigma_i^2 G_{n,i}(X_i,\boldsymbol{s}) \\
&= \frac{1}{n}E\sum_{i=1}^{n}\mu^\star(X_i)^2 + \frac{1}{n}E\left[\sum_{i=1}^{n}(Y_i-\epsilon_i)^2\right] - \frac{2}{n}E\sum_{i=1}^{n}Y_i\mu^\star(X_i) + \frac{2}{n}E\sum_{i=1}^{n}\sigma_i^2 G_{n,i}(X_i,\boldsymbol{s}) \\
&= \frac{1}{n}E\sum_{i=1}^{n}\mu^\star(X_i)^2 + \frac{1}{n}E\sum_{i=1}^{n}Y_i^2 - \frac{2}{n}E\sum_{i=1}^{n}Y_i\mu^\star(X_i) + \frac{2}{n}E\sum_{i=1}^{n}\sigma_i^2 G_{n,i}(X_i,\boldsymbol{s}) - \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \\
&= \frac{1}{n}E(\mu^\star(X_i)-Y_i)^2 + \frac{2}{n}E\sum_{i=1}^{n}\sigma_i^2 G_{n,i}(X_i,\boldsymbol{s}) - \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \\
&= E(C_p(\boldsymbol{s})) - \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \qquad (2.25)
\end{aligned}
$$

By (2.25), apparently (2.20) holds.

For selecting the function $\beta_{0n}(a)$ of the modified compound estimator $\mu^\star(x)$, the parameters $(\beta_0(a_1),...\beta_0(a_{L_n}))'$ constitute $\boldsymbol{s}$ in Theorem 2.2.2. Then as long as $\tilde{c}_{j,a}$ is a linear estimator

$$\tilde{c}_{j,a} = \sum_{i=1}^{n} l_{j,i}(a;X_i)Y_i, \qquad (2.26)$$

we'll see that

$$G_{n,i}(X_i,\boldsymbol{\beta_n}) = \sum_{a\in I_n} W_{a,n}(X_i,a,\beta_n(a))\sum_{j=0}^{J} l_{j,i}(a;X_i)(X_i-a)^j \qquad (2.27)$$

We can substitute (2.27) into (2.19) directly and combine with (2.7), the sequence $\beta_n(a)$ can be chosen by numerical optimization. If σ_i^2 is unknown for our regression model (2.1), then there are numerous methods for variance estimation in nonparametric regression. We may plug $\hat{\sigma_i}^2$ into (2.24) instead of σ_i^2. Moreover, we could choose $\boldsymbol{s}$ to be a vector including not only the function $\beta_n(a)$, but also some other tuning parameters for the method used in pointwise estimating, such as bandwidth in local regression.

2.4 Simulation study

The simulation study was done in two different parts. First, the Cp criteria (2.19) is examined by comparing it with the quantity:

$$DIMSE_{prac} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu(X_i)} - \mu(X_i))^2, \tag{2.28}$$

where *prac* stands for "practical".(i.e. without expectation). For visualization, we examine the relationship between C_p and $DIMSE_{prac}$ in two different parameter choice problems: (1) Local regression with single bandwidth choice and (2) Compound estimation with previously determined h and a single β_n.

We generate 300 $X's$ from uniform distribution on $(-1, 1)$ and error term ϵ_i is distributed as $N(0, [\frac{1}{(0.8+|X_i|)^2}]^2)$.Figure (2.1) shows how the variance change with the design points $X's$. The mean response function is $\mu(X) = \sin(2\pi X) + \cos(2\pi X) + \log(4/3 + X)$. 10 different candidate bandwidths were evenly spaced from 0.1 to 0.37. We first fit this data with local regression 10 times with different h and then calculate Cp and $DIMSE_{prac}$, also $DIMSE_{prac} + \frac{1}{n}\sum_1^n \sigma_i^2$. The simulation results appears as table (2.1). From Table 2.1, we can see that when $h = 0.3$, Cp will be minimized and $DIMSE_{prac}$ also attains its minimum. Figures to show the curve of Cp and curve of $DIMSE_{prac}$ based on additional local regresion are provided as figure 2.2. Then, h is fixed as 0.3 and we get the 15 pointwise estimators from 15 centering points on $[-0.95, 0.95]$. Again, we pick 10 different values of $\log^{\beta_n}$, and obtain Cp and $DIMSE_{prac}$ from the compound estimation for each $\log^{\beta_n}$. The simulation results are in table (2.2). We actually started with a large range of log^{β_n}, and it turns out the optimal $\log^{beta_n}$ is obviously in a smaller range, namely, 3.5 to 6.5.

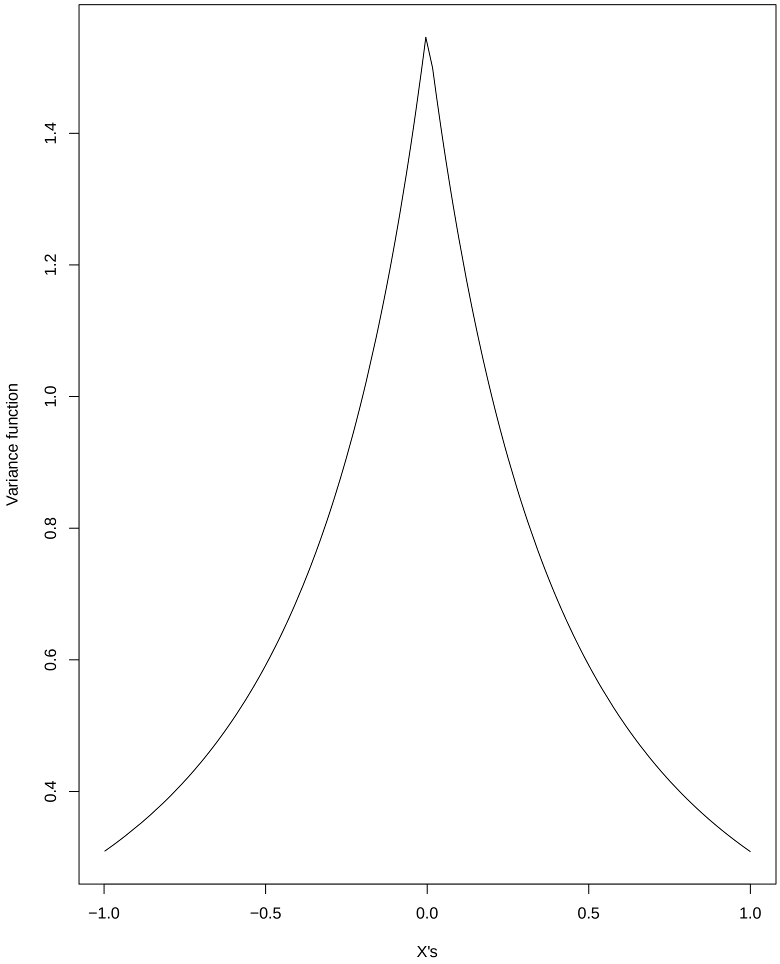
Table 2.1: Results from simulation study from local regression with different h for Cp

h	Cp	$DIMSE_{prac}$	$DIMSE_{prac} + \frac{1}{n}\sum_1^n \sigma_i^2$
0.10	0.6156	0.0372	0.6880
0.15	0.5928	0.0294	0.6802
0.20	0.5765	0.0209	0.6717
0.25	0.5701	0.0180	0.6688
0.30	0.5665	0.0159	0.6667
0.35	0.5689	0.0170	0.6678
0.40	0.5772	0.0227	0.6735
0.45	0.5916	0.0353	0.6860
0.50	0.6126	0.0553	0.7061
0.55	0.6401	0.0825	0.7333

Table 2.2: Results from simulation study from local regression with different h for DIMSE

log^{β_n}	Cp	$DIMSE_{prac}$	$DIMSE_{prac} + \frac{1}{n}\sum_1^n \sigma_i^2$
3.500	0.5781	0.0328	0.6835
3.833	0.5725	0.0267	0.6775
4.167	0.5691	0.0227	0.6734
4.500	0.5674	0.0202	0.6709
4.833	0.5671	0.0187	0.6694
5.167	0.5678	0.0179	0.6686
5.500	0.5690	0.0175	0.6683
5.833	0.5704	0.0175	0.6682
6.167	0.5715	0.0177	0.6684
6.500	0.5722	0.0179	0.6686

Figure 2.1: Variance function



The figure Cp v.s $\log^{\beta_n}$ and $DIMSE_{prac}$ v.s log^{β_n} are provided as figure 2.3. Figure 2.3 shows us that when h is fixed, the Cp informed us that log^{β_n} should be around 4.8 for minimizing the $DIMSE$. Right plot in Figure 2.3 tells us log^{β_n} should be around 5.5. This is due to the variance of both Cp and $DIMSE_{prac}$. However, the $DIMSE_{prac}$ from $\log^{\beta_n} = 4.8$ and $\log^{\beta_n} = 5.5$ do not have too much difference actually.

The Cp still tell us a lot about how to choose β_n's. We can see the curve of $DIMSE_{prac}$ seems flat after $\log^{\beta_n} = 5.5$. This is because the local regression with $h = 0.3$ has

Figure 2.2: Local regression with different bandwidth, Cp v.s. h and DIMSE v.s h

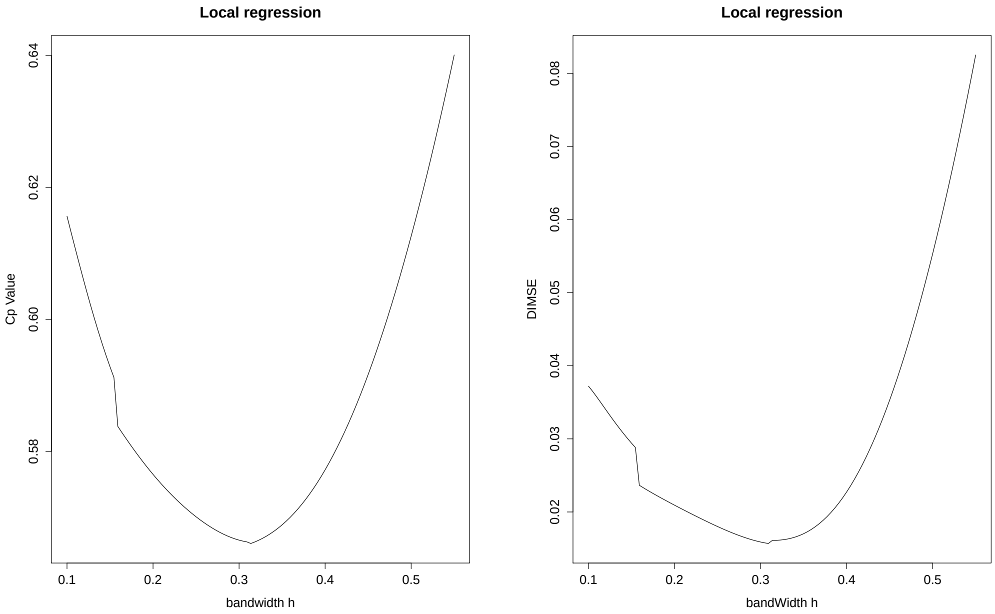
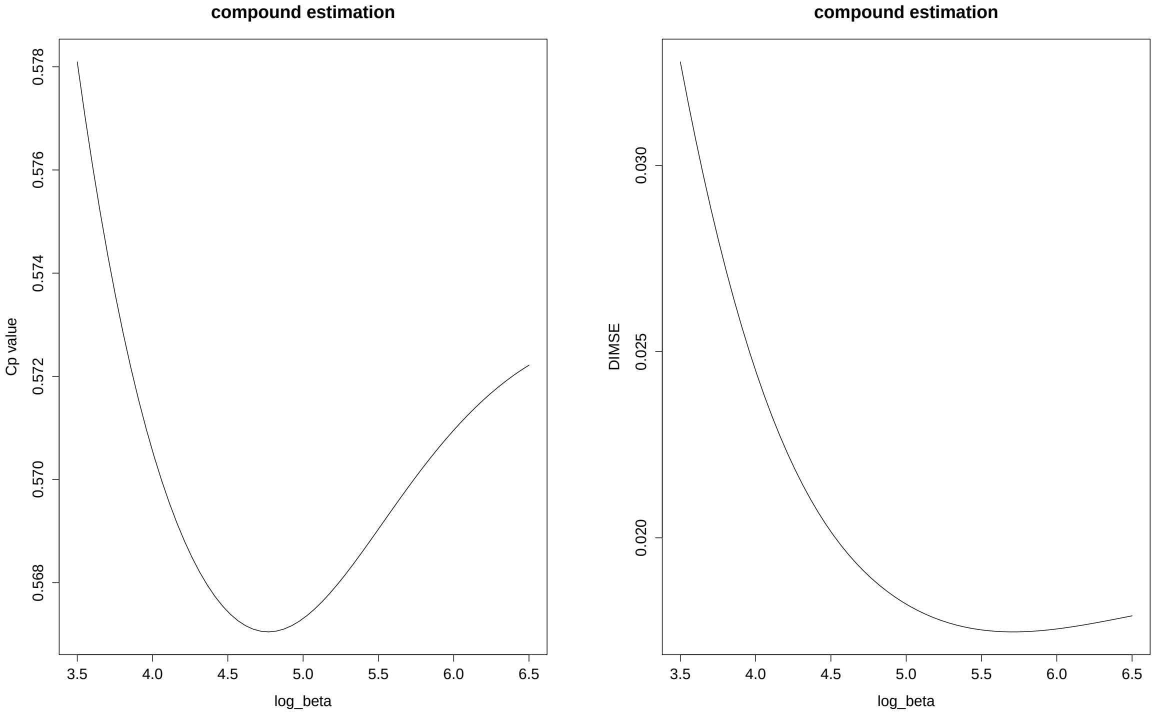


Figure 2.3: Compound estimation with different β's, Cp v.s. log^{β_n} and DIMSE v.s log^{β_n}



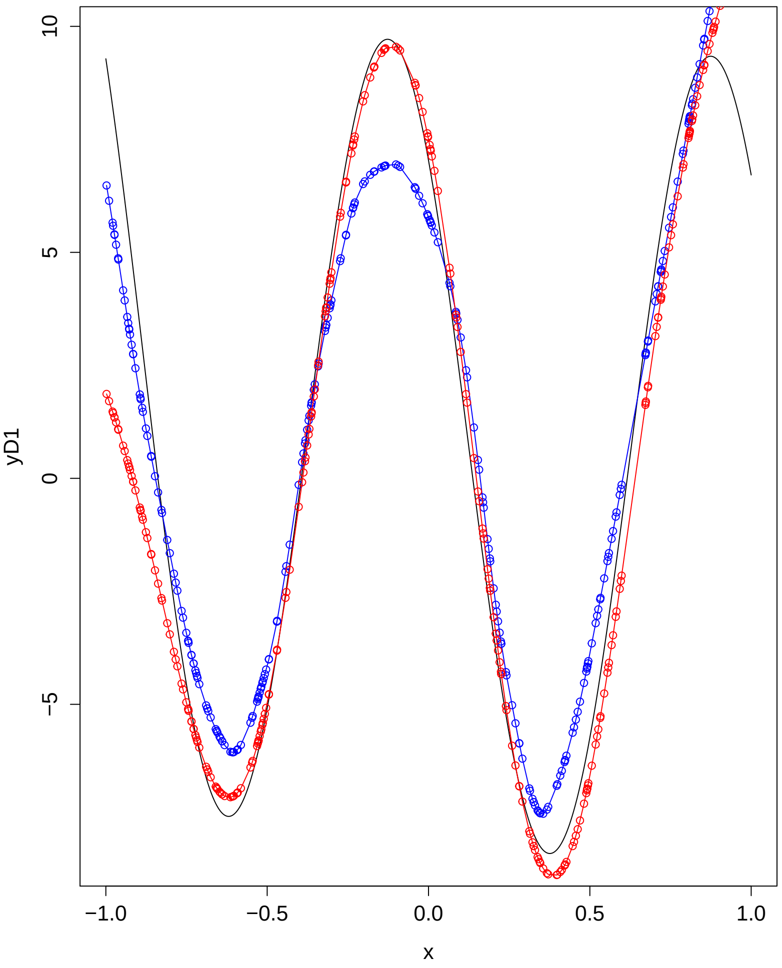
already provided a good estimation of the function. A large value of β_n means the weight function (2.4) at a fixed x for its closest centering point is almost 1. Then the estimate of $\mu(x)$ is almost the local regression estimator $\widehat{\mu(x)}$, witha small adjustment to ensure self-consitency. Thus, for this specific example, local regression may be better. However,in other example, the first derivative estimator of compound estimation may be much better than the local regression if we choose β_n appropriately. This is an advantage of compound estimation over local regression, see figure 2.4. In figure 2.4, red line represents compound estimation, blue line represents local regression estimation, $h = 0.3$ and $\beta_n = 12$. Thus, Cp can help us to choose β_n when looking primarily at the estimation of mean response.

In the second step, we try to do simulation for choosing different β_n in the weight functions to adapt to the heteroskedasticity in (2.1). The Cp criteria is used for choose parameters in four different ways when trying to fit the simulated functions: (1) Choose bandwidth in local regression. (2) Choose bandwidth and single β_n simutaneously. (3) Choose bandwidth and 15 $\beta_n's$ simutaneously. (4)We reparameterize $\beta_n's$ by a quadratic function of a. Let $\log^{\beta_n} = r_0+r_1a+r_2a^2$, then we choose r_0,r_1,r_2,h simultaneously.

The last procedure is for reducing the number of parameters. If there are many parameters in the model, then both the variance of quantity Cp and the variance of quantity $DIMSE_{prac}$ will be inflated. Then minimizing Cp may not give a optimal choice of $\beta_n's$ since what we really want to minimize is the expectation of Cp. By reparameterization, we have less number of tuning parameters, then variance of Cp will then be reduced. In practice, the way of reparameterization should somehow depend on the domain knowledge of the user and the implication of the form of heteroskedasticity from the scatter plot. For example, from the figure (2.5), we could see that the variance in the area around 0 is apparently larger than near the boundary. Thus, we may reparameterize $\log^{\beta_n}$ as a quadratic function of a.

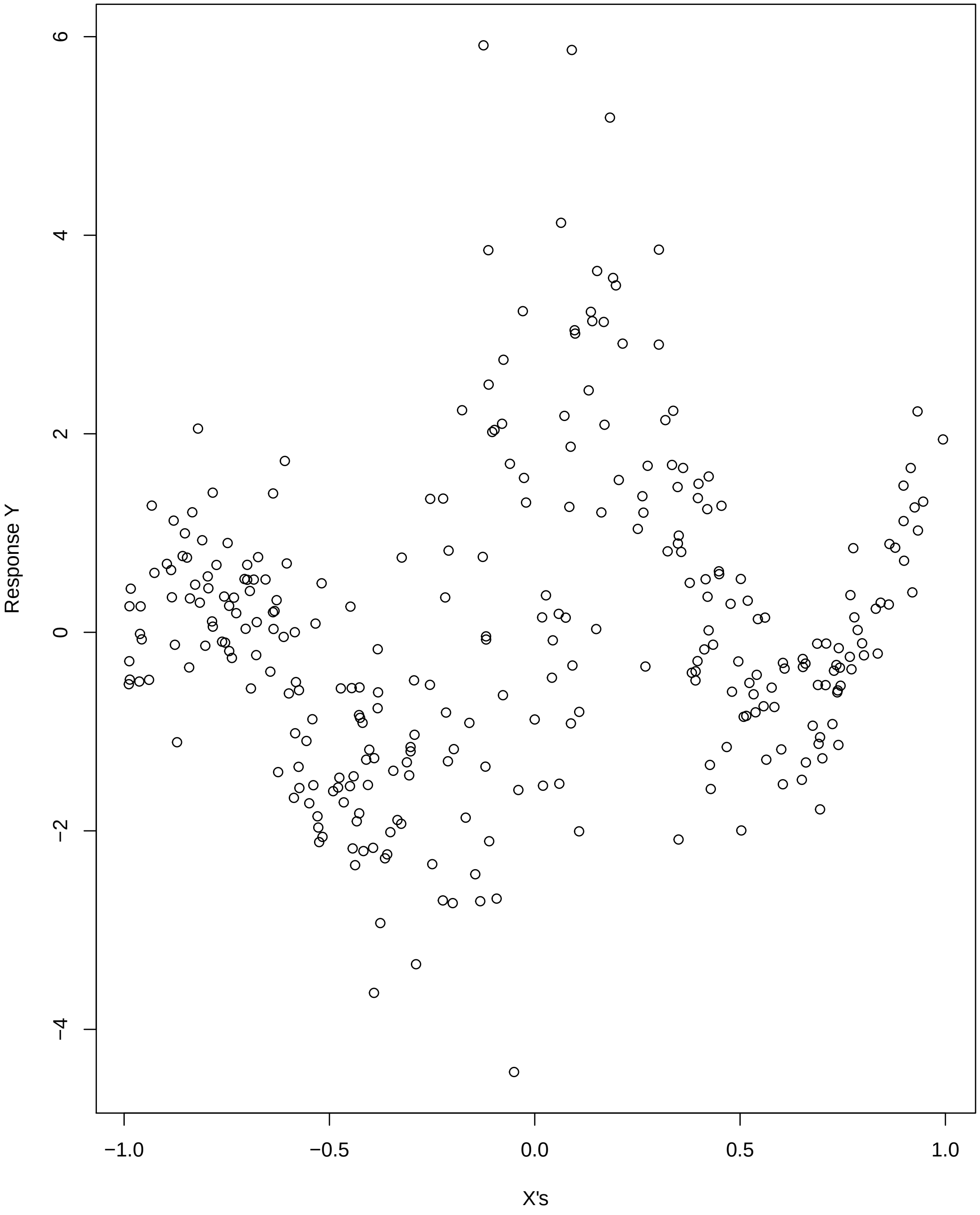
In our simulation study, we let $\mu(x) = \cos(2\pi x) + \sin(2\pi x) + \cos(3\pi x) + \sin(3\pi x)$.

Figure 2.4: plot of compound estimator of first derivative.



Three hundred data points are generated according to a uniform distribution supported on $(-1, 1)$, we pick grid points $a's$ to be equally spaced on [-0.95,0.95], the random error $\epsilon_i \sim N(0, [\frac{1}{(0.6+|X_i|)^2}]^2)$. Then we generate the 300 samples 20 times, in each trial, we use Cp to choose the tuning parameters for 4 different methods, and these methods are applied to the model fitting as mentioned before. For each method, we track 2 quantities: $DIMSE_{prac}$ and $DDIMSE_{prac}$ defined below. Therefore, 8 quantities was recorded for each trial.

Figure 2.5: Example of implication of heteroskedasticity



We define $DDIMSE_{prac}$ is defined as a practical quantity from formula (1.21),

$$DDIMSE_{prac} = \frac{1}{n}\sum_{i=1}^{n}[\widehat{\mu'_{\lambda}(X_i)} - \mu_(X_i)]^2. \tag{2.29}$$

Table 2.3: Look at $DIMSE_{prac}$ for 4 different methods

Trials	$DIMSELR$	$DIMSE2$	$DIMSE3$	$DIMSE4$
1	0.0528	0.0418	0.0623	0.0445
2	0.0464	0.0410	0.0443	0.0399
3	0.0105	0.0169	0.0211	0.0145
4	0.0862	0.0860	0.0955	0.0896
5	0.0713	0.0586	0.0830	0.0612
6	0.0367	0.0273	0.0438	0.0228
7	0.0744	0.0701	0.0699	0.0696
8	0.0371	0.0310	0.0386	0.0275
9	0.0691	0.0565	0.0794	0.0534
10	0.0618	0.0665	0.0764	0.0617
11	0.0403	0.0560	0.0656	0.0533
12	0.0227	0.0126	0.0323	0.0121
13	0.0230	0.0229	0.0281	0.0214
14	0.0316	0.0276	0.0415	0.0235
15	0.0434	0.0402	0.0398	0.0425
16	0.0459	0.0299	0.0419	0.0342
17	0.0500	0.0576	0.0685	0.0583
18	0.111	0.0777	0.0794	0.0775
19	0.0438	0.0372	0.0537	0.0348
20	0.0178	0.0164	0.0208	0.0154
Best Times	3	4	1	12

From table (2.3), we find several interesting points. First, if we compare $DIMSE2$ and $DIMSE4$, it turns out for most of the time, the reparameterization seems to improve the practical $DIMSE$ slightly over just using single β_n, which is less than what we expected. It may be that, in the first step of compound estimation, we get estimation at a discrete set of grid points from local regression that has already

Table 2.4: Look at $DDIMSE_{prac}$ for 4 different methods

Trials	$DIMSELR$	$DDIMSE2$	$DDIMSE3$	$DDIMSE4$
1	15.784	6.319	12.701	6.170
2	16.556	4.055	8.046	4.047
3	8.381	1.727	5.885	1.683
4	8.571	4.134	6.951	4.569
5	11.980	5.121	13.092	5.418
6	9.906	1.602	6.028	1.737
7	15.351	15.592	14.814	15.477
8	11.155	3.467	7.431	3.381
9	15.988	5.2884	8.972	4.845
10	14.494	7.587	12.964	7.119
11	9.381	3.135	8.021	2.851
12	7.333	1.296	8.217	1.566
13	5.791	1.657	3.714	1.617
14	9.975	3.235	11.215	3.965
15	11.764	8.242	7.920	11.007
16	6.845	3.251	7.450	3.837
17	4.656	2.087	7.099	2.544
18	17.864	4.797	11.008	4.811
19	11.619	2.235	7.492	2.123
20	11.205	1.404	3.972	1.325
Best Times	0	8	2	10

addressed the heteroskedasticity by letting the weights(in the local regression) vary proportion to the inverse of the variance. Secondly, the local regression sometimes performs the best. Naturally, β_n's need to be huge for compound estimation to have a good performance as local regression when it does well. However, for the smoothness of derivative estimation, we imposed constraints on β_n values. Moreover, the practical DIMSE from choosing 15 β_n's is not good in general. In practice, using lots of tuning parameters is not a nice idea since the variance of Cp will be inflated a lot. Also, the computational burden will be huge as the sample size increases.

Note that $DDIMSE_{prac}$ in table (2.4) is in general optimized since we choose the parameters by minimizing Cp to enhance mean response estimation but not derivative estimation. However, we still can gain insight. Obviously, the local regression didn't do a good job. The local first derivative estimation is the slope of a local polynomial, which is not self-consistent. This will not perform well when the sample size is relatively small. Compound estimation with single β_n or reparameterization did the best job for most of the times. Even if the $\beta_n's$ we choose is definitely not the optimal tuning parameters, if the user just wants to get the derivative estimation as a by-product instead of going through the estimation by another set of parameters again, compound estimator may be a satisfactory choice.

We also tried many other simulation settings, like $\mu(x) = \sin(2\pi x) + \cos(2\pi x) + \log(4/3+x)$ with 150 data points or 300 data points and $\sigma_i = \frac{1}{0.6+|X_i|}$, $\mu(x) = \sin(2\pi x)$ with 150 data points and $\sigma_i = \frac{1}{1+|X_i|}$ etc. It turns out for most of the time, compound estimator with single β_n or reparameterization works well. The local regression also sometimes did very good job for mean response estimation.

Chapter 3 Tuning parameter selection for first derivative estimation

3.1 Motivation of tuning parameter selection for derivative estimation

In chapter 2, we mentioned that if β_n is chosen appropriately, then the compound estimator will be very good for estimating the first derivative. See Figure (2.3). However, how to choose the β_n value is still a issue for the modified compound estimation. Example in Figure (2.3) tell us local regression is already good enough for the mean function estimation, however, it is not for derivative estimation and we will use compound estimation to approximate the derivatives. Then, how do we choose β_n for the derivative estimation?

If β_n is too large, then the weight function (2.4) or (2.7) will not be smooth enough to get good derivative estimation. Figures (3.1) and (3.2) are extreme examples for compound estimation with a huge value of single β_n. We random generate 300 data points from the uniform distribution supported on $(-1, 1)$. Let $\mu(x) = \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$ and $Y_i = \mu(X_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \frac{1}{(0.6+|X_i|)^2})$. Figure (3.1) shows that Cp choosing h and single β_n with compound estimator did a decent job for the mean response estimation. The red line is the estimated function and the black one is the true mean response. However,Figure (3.2) tell us β_n is apparently too large for derivative estimation. In simulation study, Cp suggests $\beta_n = 331.7$ in this case, which is against our experience that β_n should typically be less than 100. If we want a nice curve for the derivative, Cp may not be a good way for picking tuning parameters. On the other side, if β_n is too small, then the weight function will be too smooth, then we will oversmooth the derivative estimator. Therefore, instead of minimizing Discrete Integrated MSE, we may try to minimize the following quantity:

$$E\left[\sum_{i=1}^{n}\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)} - \mu'(X_i)\right)^2\right] \tag{3.1}$$

Figure 3.1: Mean response estimation by Compound estimation

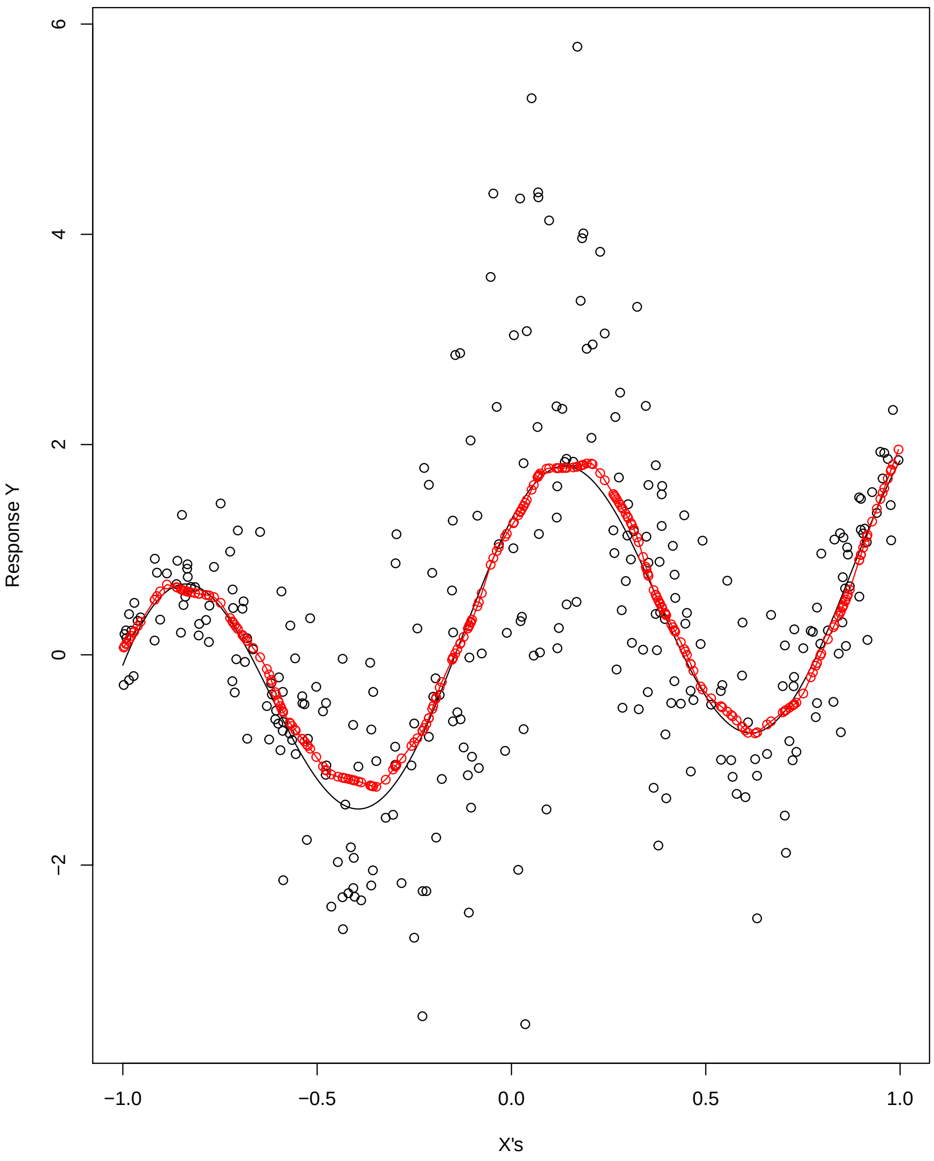
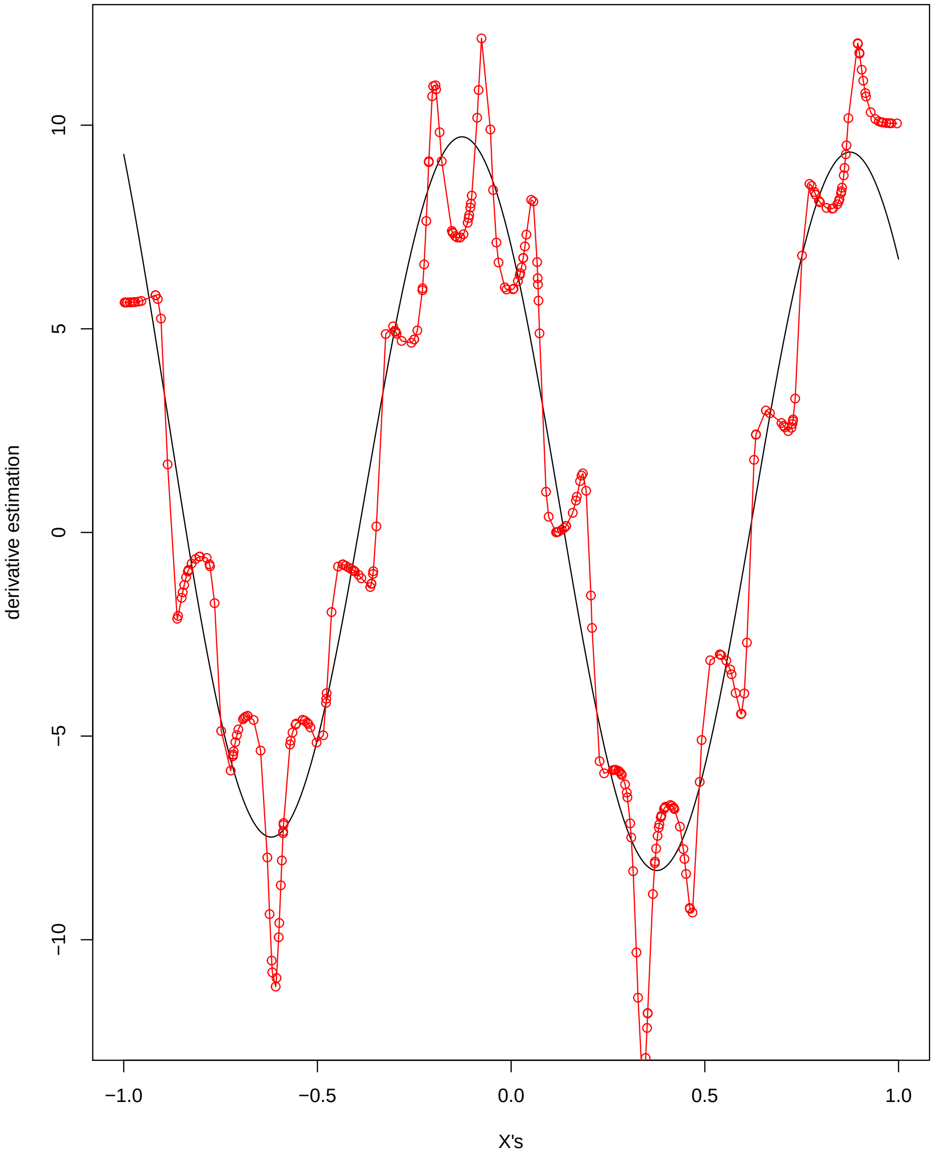


Figure 3.2: Example for bad derivative estimation



as a function of tuning parameter $\boldsymbol{\lambda}$. We name it DDIMSE. (First derivative discrete Integrated MSE).

3.2 An empirical derivative from mean response estimation and its properties

At first, we introduce an empirical first derivative for model in (2.1) and (2.6) with the random design data points $X_1, ..., X_n$. Charnigo et al. [2011] proposed an empirical first derivative for the fixed design, which is different from ours in that we allow random weights. Note: in this chapter the notations will be similar to notations used in Charnigo et al. [2011].

Let k be a positive integer, and $w_{ij}'s$ satisfy that $\sum_{j=1}^{k} w_{ij} = 1$ for $i \in \{1, 2...n\}$, then the empirical first derivative is defined as

$$Y_i^{(1)} = \sum_{j=1}^{k} w_{ij} \frac{Y_{i+j} - Y_{i-j}}{X_{i+j} - X_{i-j}}. \tag{3.2}$$

Since it is a linear combination of $Y_1, ..., Y_n$, we could also write it as

$$Y_i^{(1)} = \sum_{s=1}^{n} c_{is} Y_s, \tag{3.3}$$

where c_{is} is random and depended on $X_1, ..., X_n$. If we expand the RHS of (3.2) and (3.3), it is easy to see that

$$E[Y_i^{(1)} | \boldsymbol{X}] = \sum_{s=1}^{n} c_{is} \mu(X_s) = \sum_{j=1}^{k} w_{ij} \frac{\mu(X_{i+j}) - \mu(X_{i-j})}{X_{i+j} - X_{i-j}}. \tag{3.4}$$

As in Charnigo et al. [2011], expression (3.2) is valid for $k+1 < i < n-k$. If $i \leq k$ or $i \geq n-k+1$, then we define $Y_i^{(1)}$ by replacing k by $k(i)$, where $k(i) := \min\{i-1, n-i\}$.

The purpose of using empirical derivative is to reduce the variance of derivative estimation from ordinary difference quotients. Therefore, we could look for w_{ij} to minimize the conditional variance of (3.2).

Proposition 3.2.1 *Consider model (2.1) with (2.6), we fixed* k *and let* $\eta_{ij} = \frac{(X_{i+j}-X_{i-j})^2}{\sigma_{i+j}^2+\sigma_{i-j}^2}$, *then the conditional variance of* $Y_i^{(1)}$ *will be minimized if* $w_{ij} = \frac{\eta_{ij}}{\sum_{j=1}^{k}\eta_{ij}}$ *for* $k+1 \leq i \leq n-k$.

Proof: Let

$$h_{ij} = \frac{\sigma_{i+j}^2 + \sigma_{i-j}^2}{(X_{i+j} - X_{i-j})^2}. \tag{3.5}$$

Then

$$\begin{aligned}
Var\left[Y_i^{(1)}\middle|\, \boldsymbol{X}\right] &= Var\left[\sum_{j=1}^{k} w_{ij}\frac{Y_{i+j}-Y_{i-j}}{X_{i+j}-X_{i-j}}\middle|\, \boldsymbol{X}\right] \\
&= Var\left[\sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j}-X_{i-j}}\left(Y_{i+j}-Y_{i-j}\right)\middle|\, \boldsymbol{X}\right] \\
&= \sum_{j=1}^{k} \frac{\sigma_{i+j}^2 + \sigma_{i-j}^2}{(X_{i+j} - X_{i-j})^2} w_{ij}^2 \\
&= \sum_{j=1}^{k} h_{ij} w_{ij}^2, \qquad (3.6)
\end{aligned}$$

subject to $\sum_{j=1}^{k} w_{ij} = 1$

We introduce the Lagrange multiplier τ and let

$$L(\boldsymbol{w_i}) = \sum_{j=1}^{k} h_{ij}w_{ij}^2 + \lambda(1 - \sum_{j=1}^{k} w_{ij}). \tag{3.7}$$

This yields

$$\frac{\partial L(\boldsymbol{w_i})}{\partial w_{it}} = 2h_{it}w_{it} - \tau. \tag{3.8}$$

Then set (3.8) to be 0 for each t, leading to

$$w_{it} = \frac{\tau}{2h_{it}}, \tag{3.9}$$

since $\sum_{j=1}^{k} w_{ij} = 1$,

$$\tau = \frac{2}{\sum_{j=1}^{k} 1/h_{ij}}. \tag{3.10}$$

Then we can get

$$w_{ij} = \frac{1/h_{ij}}{\sum_{j=1}^{k} 1/h_{ij}}, \tag{3.11}$$

let $\eta_{ij} = 1/h_{ij}$, then

$$w_{ij} = \frac{\eta_{ij}}{\sum_{j=1}^{k} \eta_{ij}}, \tag{3.12}$$

where $\eta_{ij} = \frac{(X_{i+j}-X_{i-j})^2}{\sigma_{i+j}^2+\sigma_{i-j}^2}$.

Since $w_{ik} = 1 - \sum_{j=1}^{k-1} w_{ij}$, the conditional variance could be written as

$$Var\left[Y_i^{(1)}\middle|\boldsymbol{X}\right] = \sum_{j=1}^{k-1} h_{ij}w_{ij}^2 + h_{ik}(1-\sum_{j=1}^{k-1} w_{ij})^2$$

Let $\boldsymbol{w_i^{(k)}} = [w_{i1},...,w_{i(k-1)}]'$, and $\boldsymbol{H} = \begin{pmatrix} h_{i1}+h_{ik} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & h_{i(k-1)}+h_k \end{pmatrix}$. Then we can calculate the Hessian matrix of $Var\left[Y_i^{(1)}\middle|\boldsymbol{X}\right]$, which is

$$\frac{\boldsymbol{\partial}^2 Var\left[Y_i^{(1)}\middle|\boldsymbol{X}\right]}{\boldsymbol{\partial w_i^{'(k)} \partial w_i^{(k)}}} = \boldsymbol{H}.$$

Since each elements of $\boldsymbol{H}$ is greater or equal to 0, then $Var\left[Y_i^{(1)}\middle|\boldsymbol{X}\right]$ is a convex function with respect to $w_i^{(k)}$. Thus, solution (3.12) minimizes the conditional variance of $Y_i^{(1)}$. ■

3.3 Tuning parameter choice by minimizing DDIMSE

The empirical derivative bias conditional on $\boldsymbol{X}$ is

$$\begin{aligned} b_i &= E[Y_i^{(1)}|\boldsymbol{X}] - \mu'(X_i) \\ &= \sum_{j=1}^{k} w_{ij} \frac{\mu(X_{i+j}) - \mu(X_{i-j})}{X_{i+j} - X_{i-j}} - \mu'(X_i). \end{aligned} \tag{3.13}$$

By (3.4), we get

$$b_i = \sum_{s=1}^{n} c_{is}\mu(X_s) - \mu'(X_i). \tag{3.14}$$

Proposition 3.3.1 *Assume model (2.1) and (2.6) hold, the range of X's is a compact interval instead of* $[-1,1]$*, and the estimator of* $\mu'(x)$ *has the form*

$$\widehat{\mu'_{\boldsymbol{\lambda}}(x)} = \sum_{s=1}^{n} l_{s,\boldsymbol{\lambda}}(x) Y_s, \tag{3.15}$$

where $\boldsymbol{\lambda}$ *is the tuning parameter, we denote* $l_{s,\boldsymbol{\lambda}}(X_i)$ *as* l_{is} *for simplicity. Define*

$$DDIMSE = E\left[\sum_{i=1}^{n}\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)} - \mu'(X_i)\right)^2\right], \tag{3.16}$$

and

$$DCp(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\left(Y_i^{(1)} - \widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)^2 + \sum_{i=1}^{n}\sum_{s=1}^{n} c_{is}(2l_{is} - c_{is})\sigma_s^2, \tag{3.17}$$

then

$$DDIMSE = E\left[DCp(\boldsymbol{\lambda})\right] + E\left[\sum_{i=1}^{n} 2b_i\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)} - \mu'(X_i)\right)\right] - E\left[\sum_{i=1}^{n} b_i^2\right]. \tag{3.18}$$

Proof: Let

$$A_i = E\left[\sum_{t=1}^{n} c_{it}\epsilon_t\left(-\sum_{s=1}^{n} c_{is}\mu(X_s) - \sum_{s=1}^{n} c_{is}\epsilon_s + 2\sum_{s=1}^{n} l_{is}(\mu(X_s)+\epsilon_s)\right)\middle|\, \boldsymbol{X}\right], \tag{3.19}$$

$$B_i = E\left[b_i\left(-Y_i^{(1)} - \mu'(X_i) + 2\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)\middle|\, \boldsymbol{X}\right]. \tag{3.20}$$

Then

$$\begin{aligned}
A_i =& E\left[\sum_{t=1}^{n} c_{it}\epsilon_t \left(-\sum_{s=1}^{n} c_{is}\mu(X_s) - \sum_{s=1}^{n} c_{is}\epsilon_s + 2\sum_{s=1}^{n} l_{is}(\mu(X_s)+\epsilon_s)\right)\middle| \boldsymbol{X}\right] \\
=& E\left[\sum_{t=1}^{n} c_{it}\epsilon_t \left(-\sum_{s=1}^{n} c_{is}\epsilon_s + 2\sum_{s=1}^{n} l_{is}\epsilon_s\right)\middle| \boldsymbol{X}\right] \\
=& E\left[\sum_{s=1}^{n} \left(-c_{is}^2\epsilon_s^2 + 2l_{is}c_{is}\epsilon_s^2\right)\middle| \boldsymbol{X}\right] \\
=& \sum_{s=1}^{n} c_{is}(2l_{is} - c_{is})\sigma_s^2, \qquad (3.21)
\end{aligned}$$

and

$$\begin{aligned}
B_i =& E\left[b_i\left(-Y_i^{(1)} - \mu'(X_i) + 2\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)\middle| \boldsymbol{X}\right] \\
=& E\left[b_i\left(-\sum_{s=1}^{n} c_{is}\mu(X_s) - \sum_{s=1}^{n} c_{is}\epsilon_s - \mu'(X_i) + 2\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)\middle| \boldsymbol{X}\right] \\
=& E\left[b_i\left(-b_i + 2\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)} - 2\mu'(X_i)\right)\middle| \boldsymbol{X}\right] \\
=& E\left[2b_i\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)} - \mu'(X_i)\right)\middle| \boldsymbol{X}\right] - E\left[b_i^2\middle| \boldsymbol{X}\right]. \qquad (3.22)
\end{aligned}$$

Therefore, by (3.21) and (3.22)

$$
\begin{aligned}
&E\left[\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}-\mu'(X_i)\right)^2\middle|\boldsymbol{X}\right]\\
&=E\left[\left(Y_i^{(1)}-\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)^2\middle|\boldsymbol{X}\right]+E\left[\left(Y_i^{(1)}-\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)\left(-Y_i^{(1)}-\mu'(X_i)+2\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)\middle|\boldsymbol{X}\right]\\
&=E\left[\left(Y_i^{(1)}-\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)^2\middle|\boldsymbol{X}\right]\\
&+E\left[\left(\sum_{s=1}^n c_{is}\mu(X_s)+\sum_{s=1}^n c_{is}\epsilon_s-\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)\left(-Y_i^{(1)}-\mu'(X_i)+2\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)\middle|\boldsymbol{X}\right]\\
&=E\left[\left(Y_i^{(1)}-\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)^2\middle|\boldsymbol{X}\right]\\
&+E\left[\sum_{s=1}^n c_{is}\epsilon_s\left(-Y_i^{(1)}-\mu'(X_i)+2\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)\middle|\boldsymbol{X}\right]\\
&+E\left[\left(\sum_{s=1}^n c_{is}\mu(X_s)-\mu'(X_i)\right)\left(-Y_i^{(1)}-\mu'(X_i)+2\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)\middle|\boldsymbol{X}\right]\\
&=E\left[\left(Y_i^{(1)}-\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)^2\middle|\boldsymbol{X}\right]+A_i+B_i. \qquad (3.23)
\end{aligned}
$$

From the law of total expectation, we have

$$
\begin{aligned}
DDIMSE&=E\left[\sum_{i=1}^n\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}-\mu'(X_i)\right)^2\right] \qquad (3.24)\\
&=E_{\boldsymbol{X}}\left\{\sum_{i=1}^n E\left[\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}-\mu'(X_i)\right)^2\middle|\boldsymbol{X}\right]\right\}\\
&=E_{\boldsymbol{X}}\left\{\sum_{i=1}^n E\left[\left(Y_i^{(1)}-\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)^2\middle|\boldsymbol{X}\right]\right\}+E_{\boldsymbol{X}}\left[\sum_{i=1}^n A_i\right]+E_{\boldsymbol{X}}\left[\sum_{i=1}^n B_i\right]\\
&=E\left[\sum_{i=1}^n\left(Y_i^{(1)}-\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)^2\right]+E\left[\sum_{i=1}^n A_i\right]+E\left[\sum_{i=1}^n B_i\right], \qquad (3.25)
\end{aligned}
$$

which leads to (3.18). ■

Define

$$S_n(\boldsymbol{\lambda})=E\left[\sum_{i=1}^n 2b_i\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}-\mu'(X_i)\right)\right]-E\left[\sum_{i=1}^n b_i^2\right]. \qquad (3.26)$$

We could minimize DDIMSE by minimizing the *RHS* of (3.18). In practice, we need to minimize the *RHS* without expectation sign. Moreover, the second part of *RHS* of (3.18)($S_n(\boldsymbol{\lambda})$) depends on unknown mean response, it is impossible to calculate. Nevertheless, $DCp(\boldsymbol{\lambda})$ could be computed by samples, if we could prove that $S_n(\boldsymbol{\lambda})$ is asymptotically negligible compared to DDIMSE, then we may just use $DCp(\boldsymbol{\lambda})$ as the proxy.

Theorem 3.3.2 *Suppose X_i has a continuous probability density function f which is bounded away from 0 and its CDF F is twice differentiable on support set , also k is chosen appropriately as $k = O(n^\alpha)$ where $1/4 < \alpha < 1/2$, and the conditions in proposition 3.3.1 hold, let*

$$F_n(\boldsymbol{\lambda}) = E\left[\sum_{i=1}^{n}\left(Y_i^{(1)} - \widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}\right)^2\right] + E\left[\sum_{i=1}^{n}\sum_{s=1}^{n} c_{is}(2l_{is} - c_{is})\sigma_s^2\right], \tag{3.27}$$

and

$$\widehat{\boldsymbol{\lambda}_n} = \underset{\boldsymbol{\lambda}}{argmin} F_n(\boldsymbol{\lambda}) \qquad\qquad \widehat{\boldsymbol{\lambda}_n^*} = \underset{\boldsymbol{\lambda}}{argmin} DDIMSE(\boldsymbol{\lambda}). \tag{3.28}$$

Then

$$\frac{DDIMSE(\widehat{\boldsymbol{\lambda}_n})}{DDIMSE(\boldsymbol{\lambda}_n^*)} \longrightarrow 1 \qquad as \qquad n \to \infty. \tag{3.29}$$

Proof: From (3.13), we have

$$\begin{aligned} b_i &= \sum_{j=1}^{k} w_{ij}\frac{\mu(X_{i+j}) - \mu(X_{i-j})}{X_{i+j} - X_{i-j}} - \mu'(X_i) \\ &= \sum_{j=1}^{k} w_{ij}\left[\frac{\mu(X_{i+j}) - \mu(X_{i-j})}{X_{i+j} - X_{i-j}} - \mu'(X_i)\right]. \end{aligned} \tag{3.30}$$

By Taylor expansions,

$$\begin{aligned} \mu(X_{i+j}) &= \mu(X_i) + (X_{i+j} - X_i)\mu'(X_i) + 1/2(X_{i+j} - X_i)^2\mu''(\phi_{i,i+j}), \\ \mu(X_{i-j}) &= \mu(X_i) + (X_{i-j} - X_i)\mu'(X_i) + 1/2(X_{i-j} - X_i)^2\mu''(\phi_{i,i-j}), \end{aligned} \tag{3.31}$$

where $\phi_{i,i+j}$ is between X_i and X_{i+j}, $\phi_{i,i-j}$ is between X_{i-j} and X_i. The difference of these two expansions will be

$$\begin{aligned} &\mu(X_{i+j}) - \mu(X_{i-j}) \\ &\quad = (X_{i+j} - X_{i-j})\mu'(X_i) + \frac{1}{2}\left[(X_{i+j} - X_i)^2\mu''(\phi_{i,i+j}) - (X_{i-j} - X_i)^2\mu''(\phi_{i,i-j})\right], \end{aligned} \tag{3.32}$$

Assuming $i - k \geq 1$ and $i + k \leq n$, then

$$\begin{aligned} &\left|\frac{\mu(X_{i+j}) - \mu(X_{i-j})}{X_{i+j} - X_{i-j}} - \mu'(X_i)\right| \qquad (3.33) \\ &\qquad = \left|\frac{1}{2}\frac{(X_{i+j} - X_i)^2\mu''(\phi_{i,i+j}) - (X_{i-j} - X_i)^2\mu''(\phi_{i,i-j})}{X_{i+j} - X_{i-j}}\right| \\ &\qquad \leq \frac{B[(X_{i+j} - X_i)^2 + (X_{i-j} - X_i)^2]}{|X_{i+j} - X_{i-j}|} \\ &\qquad \leq B\left[|X_{i+j} - X_i| + |X_{i-j} - X_i|\right] \\ &\qquad \leq B\left[|X_{i+k} - X_i| + |X_{i-k} - X_i|\right], \qquad (3.34) \end{aligned}$$

and B is a finite positive constant. The upper bound of the rate of b_i depends on the distance between X_{i+j} or X_{i-j} and X_i. Following argument will help us to bound b_i by Bahadur's representation theorem Bahadur [1966]. Because CDF of continuous random variable X is twice differentiable on its support set and the probability density function f is bounded away from 0. Assume $\frac{i_n}{n} = p$ and $j_n = O(n^\alpha)$, $\alpha < 1/2$. Also, $q_n = i_n + j_n$, $q'_n = i_n - j_n$. Here $1 \leq j_n \leq k_n$.

Then

$$\frac{q_n}{n} = p + O(n^{\alpha-1}) \qquad \frac{q'_n}{n} = p + O(n^{\alpha-1}). \tag{3.35}$$

By Bahadur [1966], we'll have

$$\begin{aligned} X_{i_n+j_n} &= \zeta_p + \frac{q_n/n - F_n(\zeta_p)}{f(\zeta_p)} + \tilde{R}_n, \\ X_{i_n-j_n} &= \zeta_p + \frac{q'_n/n - F_n(\zeta_p)}{f(\zeta_p)} + \tilde{R}_n, \end{aligned} \tag{3.36}$$

where $\tilde{R_n} = O_p(n^{-3/4}(\log n)^{1/2(\delta+1)})$, $\delta \geq \frac{1}{2}$ and ζ_p is the pth quantile of density f. Therefore

$$\begin{aligned} X_{i_n+j_n} - X_{i_n-j_n} &= \frac{q_n - q'_n}{nf(\zeta_p)} + \tilde{R}_n \\ &= O(n^{\alpha-1})\frac{1}{f(\zeta_p)} + O_p(n^{-3/4}(\log n)^{1/2(\delta+1)}) \\ &= \begin{cases} O_p(n^{\alpha-1}) & 1/4 < \alpha < 1/2 \\ O_p(n^{-3/4}(\log n)^{1/2(\delta+1)}) & 0 < \alpha \leq 1/4. \end{cases} \end{aligned} \tag{3.37}$$

By (3.34), if $1/4 < \alpha < 1/2$, we'll have

$$\begin{aligned} |b_i| &\leq \sum_{j=1}^{k} w_{ij} \left| \left[\frac{\mu(X_{i+j}) - \mu(X_{i-j})}{X_{i+j} - X_{i-j}} - \mu'(X_i) \right] \right| \\ &\leq \sum_{j=1}^{k} w_{ij} \left| \frac{B[(X_{i+j} - X_i)^2 + (X_{i-j} - X_i)^2]}{|X_{i+j} - X_{i-j}|} \right| \\ &\leq B \sum_{j=1}^{k} w_{ij} \left[|X_{i+j} - X_i| + |X_{i-j} - X_i| \right] \\ &\leq B \sum_{j=1}^{k} w_{ij} \left[|X_{i+k} - X_i| + |X_{i-k} - X_i| \right] \\ &\leq B \left[|X_{i+k} - X_i| + |X_{i-k} - X_i| \right] \\ &= O_p(n^{\alpha-1}), \end{aligned} \tag{3.38}$$

Thus $b_i^2 \leq B^2(X_{i+k} - X_{i-k})^2$, which means

$$E[b_i^2] \leq B^2 E\left[(X_{i+k} - X_{i-k})^2\right].$$

The probability density function f is bounded away from 0 implies that there is a constant $c_1 > 0$ such that $f(x) \geq \frac{1}{c_1}$ for any points inside the support set. Since the support set is bounded by $[-1, 1]$, then there should exist a constant $c_2 > 0$ such that

$$(u - v)^2 \leq c_2(F(u) - F(v))^2$$

holds for any u and v inside the support set. By the formula of joint density of two order statistics X_{i+k} and X_{i-k}, we would have

$$
\begin{aligned}
& E\left[(X_{i+k} - X_{i-k})^2\right] \\
& = \iint\limits_D \frac{n!}{(i-k-1)!(2k-1)!(n-i-k)!} \\
& (u-v)^2 F(v)^{i-k-1}(F(u)-F(v))^{2k-1}(1-F(v))^{n-i-k} dF(u)dF(v) \\
& = \frac{n!}{(i-k-1)!(2k-1)!(n-i-k)!} \\
& \iint\limits_D (u-v)^2 F(v)^{i-k-1}(F(u)-F(v))^{2k-1}(1-F(v))^{n-i-k} dF(u)dF(v) \\
& \leq \frac{c_2 n!}{(i-k-1)!(2k-1)!(n-i-k)!} \\
& \iint\limits_D F(v)^{i-k-1}(F(u)-F(v))^{2k+1}(1-F(v))^{n-i-k} dF(u)dF(v) \\
& = \frac{c_2 n!}{(i-k-1)!(2k-1)!(n-i-k)!} \\
& \iint\limits_D F(v)^{(i+1)-(k+1)-1}(F(u)-F(v))^{2(k+1)-1}(1-F(v))^{(n+2)-(i+1)-(k+1)} dF(u)dF(v) \\
& = \frac{c_2 n!}{(i-k-1)!(2k-1)!(n-i-k)!} \frac{(i-k-1)!(2k+1)!(n-i-k)!}{(n+2)!} \\
& = \frac{c_2(2k+1)(2k)}{(n+2)(n+1)},
\end{aligned}
$$

where $D = \{(u,v) : -1 \leq v \leq u \leq 1\}$. Since $k = O(n^\alpha)$,

$$E[b_i^2] \leq O(n^{2\alpha-2}). \tag{3.39}$$

Also, by Cauchy Schwarz inequality

$$\begin{aligned}
&\left| E\sum_{i=1}^{n}\left[b_i\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}-\mu'(X_i)\right)\right]\right| \\
&\leq \{E[\sum_{i=1}^{n} b_i^2]\}^{1/2}\left\{E\left[\sum_{i=1}^{n}\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}-\mu'(X_i)\right)^2\right]\right\}^{1/2} \\
&= \{\sum_{i=1}^{n} E[b_i^2]\}^{1/2}\left\{E\left[\sum_{i=1}^{n}\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}-\mu'(X_i)\right)^2\right]\right\}^{1/2} \\
&\leq O_p(n^{\alpha-\frac{1}{2}})\left\{E\left[\sum_{i=1}^{n}\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)}-\mu'(X_i)\right)^2\right]\right\}^{1/2}. \qquad (3.40)
\end{aligned}$$

Notice (3.40) holds for any $\boldsymbol{\lambda}_n \in \Lambda_n$, Λ_n is the parameter space. Now let

$$F_n(\boldsymbol{\lambda}_n) = E\left[\sum_{i=1}^{n}\left(Y_i^{(1)}-\widehat{\mu'_{\boldsymbol{\lambda}_n}(X_i)}\right)^2\right] + E\left[\sum_{i=1}^{n}\sum_{s=1}^{n} c_{is}(2l_{is}-c_{is})\sigma_s^2\right], \qquad (3.41)$$

$$S_n(\boldsymbol{\lambda}_n) = E\left[\sum_{i=1}^{n} 2b_i\left(\widehat{\mu'_{\boldsymbol{\lambda}_n}(X_i)}-\mu'(X_i)\right)\right] - E\left[\sum_{i=1}^{n} b_i^2\right]. \qquad (3.42)$$

Then $DDIMSE(\boldsymbol{\lambda}_n) = F_n(\boldsymbol{\lambda}_n) + S_n(\boldsymbol{\lambda}_n)$ for $\forall\ \boldsymbol{\lambda}_n \in \boldsymbol{\Lambda_n}$, $\boldsymbol{\Lambda_n}$ is the parameter space. Let

$$\widehat{\boldsymbol{\lambda}_n} = \underset{\boldsymbol{\lambda}_n\in\boldsymbol{\Lambda_n}}{\arg\min} F_n(\boldsymbol{\lambda}_n) \qquad \boldsymbol{\lambda}_n^* = \underset{\boldsymbol{\lambda}_n\in\boldsymbol{\Lambda_n}}{\arg\min} DDIMSE(\boldsymbol{\lambda}_n), \qquad (3.43)$$

and

$$M = \max\{|S_n(\widehat{\boldsymbol{\lambda}_n})|, |S_n(\boldsymbol{\lambda}_n^*)|\}. \qquad (3.44)$$

Therefore

$$\begin{aligned}
DDIMSE(\widehat{\boldsymbol{\lambda}_n}) &= F_n(\widehat{\boldsymbol{\lambda}_n}) + S_n(\widehat{\boldsymbol{\lambda}_n}) \\
&\leq F_n(\widehat{\boldsymbol{\lambda}_n}) + M \\
&\leq F_n(\boldsymbol{\lambda}_n^*) + M \\
&= DDIMSE(\boldsymbol{\lambda}_n^*) - S_n(\boldsymbol{\lambda}_n^*) + M \\
&\leq DDIMSE(\boldsymbol{\lambda}_n^*) + 2M. \qquad (3.45)
\end{aligned}$$

$\Rightarrow$

$$1 \leq \frac{DDIMSE(\widehat{\boldsymbol{\lambda}_n})}{DDIMSE(\boldsymbol{\lambda}_n^*)} \leq 1 + \frac{2M}{DDIMSE(\boldsymbol{\lambda}_n^*)}. \tag{3.46}$$

We need to show that

$$\frac{M}{DDIMSE(\boldsymbol{\lambda}_n^*)} \longrightarrow 0. \tag{3.47}$$

From Stone [1980], we know that the optimal convergence rate of first derivative estimation in nonparametric regression is $O_p(n^{-\frac{2J}{2J+3}})$ if the mean response is at least $(J+1)th$ times continuously differentiable. Therefore

$$DDIMSE(\boldsymbol{\lambda}_n) \geq n \cdot \Theta_p(n^{-\frac{2J}{2J+3}}) = \Theta_p(n^{\frac{3}{2J+3}}). \tag{3.48}$$

Without loss of generality, assuming that $|S_n(\widehat{\boldsymbol{\lambda}_n})| > |S_n(\boldsymbol{\lambda}_n^*)|$. Obviously we'll have

$$\begin{aligned} & \frac{M}{DDIMSE(\boldsymbol{\lambda}_n^*)} \\ = & \frac{|S_n(\widehat{\boldsymbol{\lambda}_n})|}{DDIMSE(\boldsymbol{\lambda}_n^*)} \\ = & \frac{|S_n(\widehat{\boldsymbol{\lambda}_n})|}{DDIMSE(\widehat{\boldsymbol{\lambda}_n})} \frac{DDIMSE(\widehat{\boldsymbol{\lambda}_n})}{DDIMSE(\lambda_n^*)} \end{aligned}$$

Let $Q_n = \frac{|S_n(\widehat{\lambda_n})|}{DDIMSE(\widehat{\lambda_n})}$ and $T_n = \frac{M}{DDIMSE(\lambda_n^*)}$. Then we get

$$\begin{aligned} T_n &\leq Q_n(1+2T_n) \\ \Rightarrow T_n &\leq \frac{Q_n}{1-2Q_n} \end{aligned}$$

From (3.39), (3.40), (3.42) and (3.48), we have

$$\begin{aligned}
Q_n &\leq \frac{\left|E\left[\sum_{i=1}^n 2b_i\left(\widehat{\mu'_{\widehat{\boldsymbol{\lambda}}_n}(X_i)} - \mu'(X_i)\right)\right]\right|}{DDIMSE(\widehat{\boldsymbol{\lambda}}_n)} + \frac{E\left[\sum_{i=1}^n b_i^2\right]}{DDIMSE(\widehat{\boldsymbol{\lambda}}_n)} \\
&\leq O_p(n^{\alpha-\frac{1}{2}})\left\{E\left[\sum_{i=1}^n\left(\widehat{\mu'_{\widehat{\boldsymbol{\lambda}}_n}(X_i)} - \mu'(X_i)\right)^2\right]\right\}^{-1/2} \\
&+ O_p(n^{2\alpha-2})\left\{E\left[\sum_{i=1}^n\left(\widehat{\mu'_{\widehat{\boldsymbol{\lambda}}_n}(X_i)} - \mu'(X_i)\right)^2\right]\right\}^{-1} \\
&\leq O_p(n^{\alpha-\frac{1}{2}-\frac{3}{4J+6}}) + O_p(n^{2(\alpha-1-\frac{3}{4J+6})}) \\
&= O_p(n^{\alpha-\frac{1}{2}-\frac{3}{4J+6}}).
\end{aligned}$$

Noting that $1/4 < \alpha < 1/2$. We'll have $\alpha - \frac{1}{2} - \frac{3}{4J+6} < 0$, therefore

$$\frac{M}{DDIMSE(\boldsymbol{\lambda}_n^*)} = T_n \leq O_p(n^{\alpha-\frac{1}{2}-\frac{3}{4J+6}}), \tag{3.49}$$

which means T_n goes to 0 as $n \to 0$, namely, (3.47) holds. ■

Proposition 3.3.3 *Let*

$$\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{n}} = diag\begin{pmatrix} \widehat{\lambda_{n1}} & & \\ & \ddots & \\ & & \widehat{\lambda_{ns}} \end{pmatrix}$$

and

$$\boldsymbol{\Lambda}_{\boldsymbol{n}}^{\boldsymbol{*}} = diag\begin{pmatrix} \lambda_{n1}^* & & \\ & \ddots & \\ & & \lambda_{ns}^* \end{pmatrix}$$

Also, let

$$\widehat{\boldsymbol{\lambda}_{\boldsymbol{n}}} = [\widehat{\lambda}_{n1}, \cdots, \widehat{\lambda}_{ns}]' \qquad \boldsymbol{\lambda}_{\boldsymbol{n}}^{\boldsymbol{*}} = [\lambda_{n1}^*, \cdots, \lambda_{ns}^*]', \tag{3.50}$$

$$\widehat{\boldsymbol{\phi}_{\boldsymbol{n}}} = [\log \widehat{\lambda}_{n1}, \cdots, \log \widehat{\lambda}_{ns}]' \qquad \widehat{\boldsymbol{\phi}^*_{\boldsymbol{n}}} = [\log \lambda^*_{n1}, \cdots, \log \lambda^*_{ns}]'.$$
$$\widehat{\boldsymbol{\Phi}_{\boldsymbol{n}}} = \log \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{n}} \qquad \widehat{\boldsymbol{\Phi}^*_{\boldsymbol{n}}} = \log \boldsymbol{\Lambda}^*_{\boldsymbol{n}}. \tag{3.51}$$

Let $T_n(\boldsymbol{\lambda}) = DDIMSE(\boldsymbol{\lambda})$. Put

$$U_n(\widehat{\boldsymbol{\phi}}) := T_n(\exp \widehat{\boldsymbol{\phi}}) \tag{3.52}$$

and assume $U_n(\widehat{\boldsymbol{\phi}})$ is at least twice continuously differentiable and $\frac{\boldsymbol{D}^2 U_n(\boldsymbol{\phi_n})}{U_n(\boldsymbol{\phi_n})}$ is positive definite and its determinant is bounded below by some positive constant A for $\boldsymbol{\phi_n}$ in its parameter space, that is $\boldsymbol{\phi_n}$ satisfied $\exp(\boldsymbol{\phi_n}) \in \boldsymbol{\Lambda_n}$. Then we'll have

$$\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{n}} \boldsymbol{\Lambda}^{*\,-1}_{\boldsymbol{n}} \to \boldsymbol{I}. \tag{3.53}$$

where $\boldsymbol{I}$ is the identity matrix.

This results strengthen the Corollary 1 in Charnigo et al. [2011] since we extend it to the case with multiple parameters and random design points.

Proof:

By Talyor expansion

$$U_n(\widehat{\boldsymbol{\phi_n}}) = U_n(\boldsymbol{\phi^*_n}) + (\widehat{\boldsymbol{\phi_n}} - \boldsymbol{\phi^*_n})' \boldsymbol{D} U_n(\boldsymbol{\phi^*_n}) + \frac{1}{2}(\widehat{\boldsymbol{\phi_n}} - \boldsymbol{\phi^*_n})' \boldsymbol{D}^2 U_n(\boldsymbol{\phi^{*\prime}_n})(\widehat{\boldsymbol{\phi_n}} - \boldsymbol{\phi^*_n}), \tag{3.54}$$

where $\boldsymbol{\phi^{*\prime}_n}$ is a vector lies on the line segment between $\widehat{\boldsymbol{\phi_n}}$ and $\boldsymbol{\phi^*_n}$. Divide (3.54) by $U_n(\boldsymbol{\phi^*_n})$ on each side,

$$\frac{U_n(\widehat{\boldsymbol{\phi_n}})}{U_n(\boldsymbol{\phi^*_n})} = 1 + (\widehat{\boldsymbol{\phi_n}} - \boldsymbol{\phi^*_n})' \frac{\boldsymbol{D} U_n(\boldsymbol{\phi^*_n})}{U_n(\boldsymbol{\phi^*_n})} + \frac{1}{2}(\widehat{\boldsymbol{\phi_n}} - \boldsymbol{\phi^*_n})' \frac{\boldsymbol{D}^2 U_n(\boldsymbol{\phi^{*\prime}_n})}{U_n(\boldsymbol{\phi^*_n})}(\widehat{\boldsymbol{\phi_n}} - \boldsymbol{\phi^*_n}). \tag{3.55}$$

Since $\boldsymbol{\phi^*_n}$ minimizes U_n, then $\boldsymbol{D} U_n(\boldsymbol{\phi^*_n}) = \boldsymbol{0}$, along with $\frac{U_n(\widehat{\boldsymbol{\phi_n}})}{U_n(\boldsymbol{\phi^*_n})} \longrightarrow 1$, we'll have

$$(\widehat{\boldsymbol{\phi_n}} - \boldsymbol{\phi^*_n})' \frac{\boldsymbol{D}^2 U_n(\boldsymbol{\phi^{*\prime}_n})}{U_n(\boldsymbol{\phi^*_n})}(\widehat{\boldsymbol{\phi_n}} - \boldsymbol{\phi^*_n}) \longrightarrow 0. \tag{3.56}$$

Because $\frac{\boldsymbol{D}^2 U_n(\boldsymbol{\phi}_{\boldsymbol{n}}^{*\prime})}{U_n(\boldsymbol{\phi}_{\boldsymbol{n}}^*)}$ is positive definite and its determinant is bound below by some constant,

$$(\widehat{\boldsymbol{\phi}_{\boldsymbol{n}}} - \boldsymbol{\phi}_{\boldsymbol{n}}^*) \longrightarrow \mathbf{0}, \tag{3.57}$$

⇒

$$\widehat{\boldsymbol{\Phi}}_{\boldsymbol{n}} - \widehat{\boldsymbol{\Phi}_{\boldsymbol{n}}^*} \longrightarrow \boldsymbol{O}, \tag{3.58}$$

⇒

$$\exp(\widehat{\boldsymbol{\Phi}}_{\boldsymbol{n}} - \widehat{\boldsymbol{\Phi}_{\boldsymbol{n}}^*}) \longrightarrow \boldsymbol{I}, \tag{3.59}$$

⇒

$$\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{n}} {\boldsymbol{\Lambda}_{\boldsymbol{n}}^*}^{-1} \longrightarrow \boldsymbol{I}. \tag{3.60}$$

■

Since the optimal rate of the first derivative estimation from nonparameteric regression is $O_p(n^{-\frac{J}{2J+3}})$, we will have

$$b_i = o_p(\left|\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)} - \mu'(X_i)\right|), \tag{3.61}$$

which implies that

$$\frac{\sum_{i=k+1}^{n-k}\left[2b_i\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)} - \mu'(X_i)\right) - b_i^2\right]}{\sum_{i=k+1}^{n-k}\left(\widehat{\mu'_{\boldsymbol{\lambda}}(X_i)} - \mu'(X_i)\right)^2} \xrightarrow{P} 0 \quad as \quad n \to \infty. \tag{3.62}$$

In simulation study, we need to use $DCp(\boldsymbol{\lambda})$, which is without expectation sign. Therefore (3.62) could be a justification of ignoring the $S_n(\boldsymbol{\lambda})$ in the simulation study. However, we should be aware of the variation of DCp criteria. If DCp has a large variance, then the difference between $DCp(\boldsymbol{\lambda})$ and $E[DCp(\boldsymbol{\lambda})]$ will be large and minimizing $DCp(\boldsymbol{\lambda})$ may not indicate that $E(DCp(\boldsymbol{\lambda}))$ attains its minimum, so is the DDIMSE.

3.4 simulation study

We did the simulation study for two different scenarios. One is to use DCp criteria to choose bandwidth in local regression for first derivative estimation. The other is to choose bandwidth and weight β_n in Gaussian convolutions simultanuously for first derivative compound estimation. In both scenarios, we compared the performance of DCp criteria and its several competitors, includes the Cp in the chapter 2; generalized cross validation applied to the fitted and empirical derivatives with the appropriate choice of k ($GCV1k$); ordinary cross validation applied to the first empirical derivatives with the appropriate choice of k ($CVE1k$). There are also some other criterias for tuning parameter selection in nonparametric regression. However, some of them were not designed for heteroskedasticity case or for the derivative estimation. All of the methods we used above will incorporate the consideration of heteroskedasticity.

Charnigo et al. [2011] used a quantity

$$Q1 := \sum_{i=1}^{n} s_i \left(\frac{d}{dx}\mu(X_i) - \widehat{\frac{d}{dx}\mu_{\widetilde{\boldsymbol{\lambda}}}(X_i)} \right)^2 \Bigg/ \min_{\boldsymbol{\lambda}\in\Lambda_n} \sum_{i=1}^{n} s_i \left(\frac{d}{dx}\mu(X_i) - \frac{d}{dx}\mu_{\boldsymbol{\lambda}}(X_i) \right)^2 \tag{3.63}$$

to compare the GCp criteria with other tuning parameter selection methods, where $\widetilde{\boldsymbol{\lambda}}$ is the tuning parameter chosen by the specific method and $\boldsymbol{\lambda}$ is the tuning parameter that minimize the bottom of the (3.63). Λ_n is the set of all the tuning candidate parameters.

In this section, we will also use $Q1$ as the quantity for making comparison between DCp and the methods mentioned above. In practice, it's impossible to calculate $DDIMSE$, we define $DDIMSE_{prac}$ as

$$DDIMSE_{prac} = \sum_{i=1}^{n} s_i(\widehat{\frac{d}{dx}\mu_{\boldsymbol{\lambda}}(X_i)} - \frac{d}{dx}\mu(X_i))^2. \tag{3.64}$$

Then $Q1$ could be written as

$$Q1 = \frac{DDIMSE_{prac}(\widetilde{\boldsymbol{\lambda}})}{DDIMSE_{prac}(\boldsymbol{\lambda})} \tag{3.65}$$

Local regression. Firstly, we use DCp criteria to pick an approriate bandwidth for local regression. We generate the data set 20 times from model (2.1), and for each data set, there are 800 $X's$ distributed as uniform distribution on $(-1, 1)$ and error term ϵ_i is distributed as $N(0, [\frac{1}{(0.8+|X_i|)^2}]^2)$. The mean response function is $\mu(X) = \sin(2\pi X)$. 50 different candidate bandwidths were evenly spaced from 0.1 to 0.6. We fit each data set with local regression 50 times with different h. Let $k \in \{10, 15, 20, 25, 30\}$, DCp values from different h and k was computed, and then we pick the bandwidth which minimize the sum of the DCp values over 6 different k's. We let s_i to be 1 when $21 \leq s_i \leq 780$, otherwise $s_i = 0$. Simulation results are shown in Table 3.1. $DCpSum$, $CVE1kSum$, $GCV1kSum$ means we minimize the sum of these three criteria over 6 different values of k respectively. Then we select the bandwidth which minimize the Sum's. The $Q1$ values in (3.65) are computed for each trial and each tuning parameter selection method. We also record the wining times of each method and their average $Q1$ value as in Table 3.2.

As shown in Table 3.1 and 3.2, $DCpSum$ performs much better than $CVE1k$ and $GCV1kSum$. Cp has three wining times, which is not a surprise. We mentioned that if the mean response is smooth, then the Cp may also give us moderate good tuning parameters even if we are looking for optimal derivative estimation. Nevertheless, $DCpSum$ still outperformed Cp a lot. Therefore, the DCp will be a better choice than Cp if we want specifically a nice derivative estimation.

However, there is a concern about DCp. When tried to approximate $DDIMSE$, we proved that (3.26) is negligible comparing to $DDIMSE$ as $n \to \infty$. If the sample

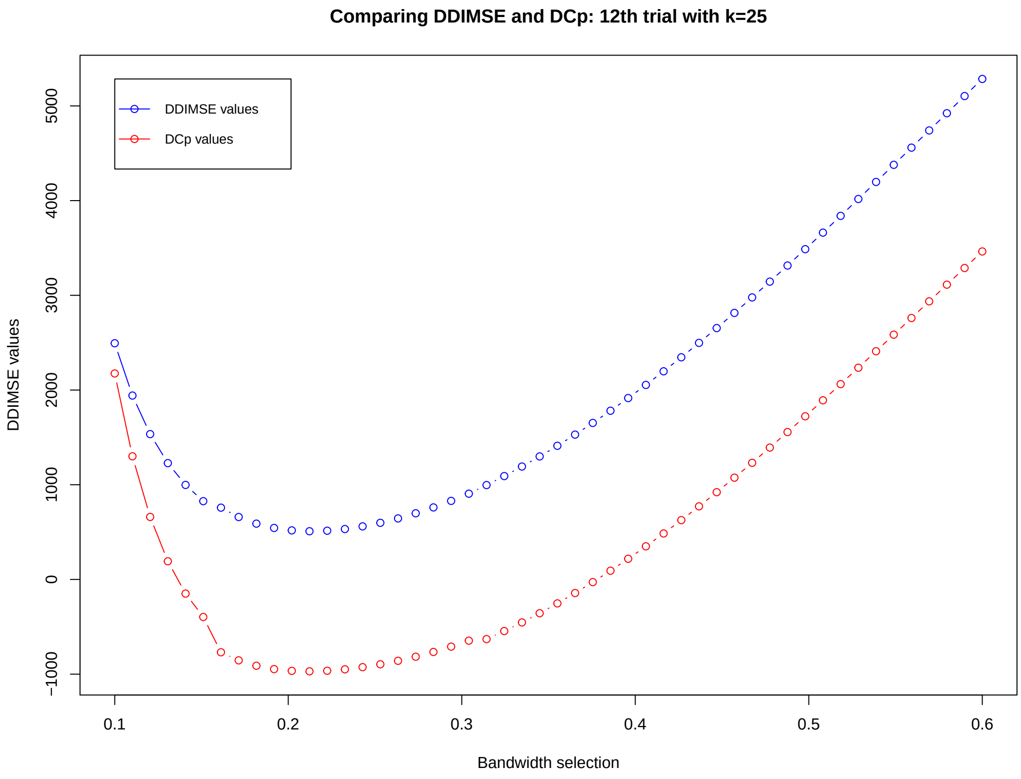
Table 3.1: Comparison for local regresion with bandwidth selection

Trials	$DCpSum$	Cp	$CVE1kSum$	$GCV1kSum$
1	1.2025	3.1254	8.1557	18.2407
2	1.0023	2.0721	8.0906	12.4772
3	1.7458	1.0171	6.1243	7.9359
4	1.1384	2.0319	4.5032	15.9404
5	1.2022	2.0875	4.1071	9.8594
6	1.3033	1.8424	5.0143	11.9014
7	1.0083	1.6542	3.0403	8.0436
8	1.0396	1.9145	6.4430	14.4529
9	1.0928	1.1534	5.6891	12.4176
10	1.2089	1.6436	3.4803	9.9332
11	1.8714	3.5124	6.7993	24.1581
12	1.0106	1.7787	4.8981	10.3792
13	2.0078	3.4274	8.1436	28.5114
14	1.2830	1.6668	2.0839	7.0453
15	3.7046	2.3706	8.8142	12.8036
16	1.0274	1.6893	3.9746	10.9566
17	1.0000	2.2344	7.5322	17.5417
18	1.1794	2.1717	4.2119	10.2884
19	1.2726	2.1774	5.4817	12.3370
20	1.1982	1.0386	4.7169	7.3177

Table 3.2: Comparison for local regresion with bandwidth selection

Methods	$DCpSum$	Cp	$CVE1kSum$	$GCV1kSum$
Wining times	17	3	0	0
Average $Q1$	1.3750	2.0305	5.5652	13.1271

Figure 3.3: Plot of bandwidth vs $DDIMSE$ or DCp



size is too small, this may not hold. Therefore, we recommend to use DCp for first derivative estimation when the sample size is moderately large. If the sample size n is too small and the mean response seems very smooth, Cp criteria in chapter 2 may be a better choice, because $DIMSE$ is exactly equal to $E[Cp]$ plus a constant, like (2.25). Also, in real data analysis, we won't be able to know the variance of each error term ϵ_i. We need to estimate the variance of σ_i^2 and then plug these estimators into (3.17) to obtain the DCp criteria.

Visualization of relations between $DDIMSE_{prac}$ and DCp values are shown in Figure 3.3. It is from the 12th trial with $k = 25$. From Figure 3.3, the bandwidth ($h \approx 0.2$) which minimize the DCp will also approximately minimize the corresponding $DDIMSE_{prac}$ value.

Compound estimation. Secondly, we applied DCp criteria to pick two tuning parameters simultaneously for compound estimation. Charnigo et al. [2011]. Again, we generate the data set 20 times from model (2.1) with mean response $\mu(X) = \sin(5\pi X) + \sin(3\pi X) + \cos(\pi X)$. For each time, there are 800 X's from uniform distribution on $(-1, 1)$. The error term ϵ_i is distributed as $N(0, [\frac{1}{(0.8+|X_i|)^2}]^2)$. To clarify, the error term ϵ_i is independent from X_i, we let ϵ_i has variance $[\frac{1}{(0.8+|X_i|)^2}]^2$ just for convenience. We used 15 centering points equally spaced on $[-0.95, 0.95]$ and local polynomial of degree 2 in the compound estimation. The parameters we need to choose are the bandwidth of local regression for pointwise estimation and a single weight β_n as in (2.4). The candidates for bandwidth h are 10 values equally spaced on $[0.05, 0.3]$ and for β_n are 10 values equally spaced on $[20, 100]$. Then we calculate the tuning parameters selection criteria 100 times for 100 pairs of $\{h, \beta_n\}$ for each $k \in \{10, 15, 20, 25\}$. We let s_i to be 1 when $21 \leq s_i \leq 780$, otherwise $s_i = 0$.

Very nicely, the Figure 3.4 shows us the similar behavior of $DDIMSE_{prac}$ and DCp criteria when they are varying with different pair of tuning parameter selections and appropriate k. From this figure, it is clear that DCp will be a good proxy for $DDIMSE_{prac}$, especially when $DDIMSE_{prac}$ is close to its minimum.

The $Q1$ values are as in Table 3.3. Table 3.4 displays the wining times and average $Q1$ for each tuning selection method. The results in Table 3.2 and 3.3 tell us that DCp will be the optimal choice of tuning parameter selection method. Surprisingly, $CVE1k$ works for several cases in simulation. However, it is not good at all for some of the trials. Therefore, even if it works very well for 3 trials, it is not stable. Also, Table 3.1 showed us $CVE1k$ is not good for local regression when we need to estimate first derivative. Our guess is that it may not work for general nonparametric first derivative estimation. Cp criteria is worse than DCp in most of the case, however it still have 5 times better than DCp criteria. We know that Cp will give a moderately good bandwidth for first derivative estimation when the function is smooth, but it is not good at choosing the convolution weight β_n. In most of the case, it will choose the biggest β_n for us. Here, Cp choose $\beta_n = 100$, which is the largest in our candidate set. Figure 3.5 shows the Cp value for different pair of $\{h, \beta_n\}$. Apparently, for each bandwidth h, Cp is always a decreasing function of β_n. If we don't restrict the upper bound of β_n value, we'll get plot like Figure 3.2 and a worse result for Cp. However, that is not the main purpose of simulation study. If we choose β_n based on our previous experience (β_n shouldn't be too large), Cp still can be applied for the choosing a bandwidth when we don't want to be bothered to select tuning parameters again after estimating the mean response. However, if we would like to obtain a good estimation of first derivative, DCp with appropriate k would be the best choice from the simulation study.

Figure 3.6 shows us in trial 12, the derivative estimation from the tuning parameters from DCp (red line) and from Cp (green line). Both of them select the same bandwidth h, however, choices of β_n are different. Cp pick the largest $\beta_n = 100$ and DCp pick $\beta_n = 64.4$. The DCp pick the 36th pair of parameters $\{0.133, 64.4\}$ and Cp pick the 40th pair of parameters $\{0.133, 100\}$. We can see from Figure 3.6 the derivative estimation is not that sensitive to a small β_n value since the green line and the red line are close, even if the green line is slightly bad than the red one. From the other side, we recognize that even if the estimation is not that sensitive to the tuning parameter choice, DCp may still give us a better choice of the tuning parameters.

Figure 3.4: Plot of $\{h, \beta_n\}$ vs $DDIMSE$ or DCp.

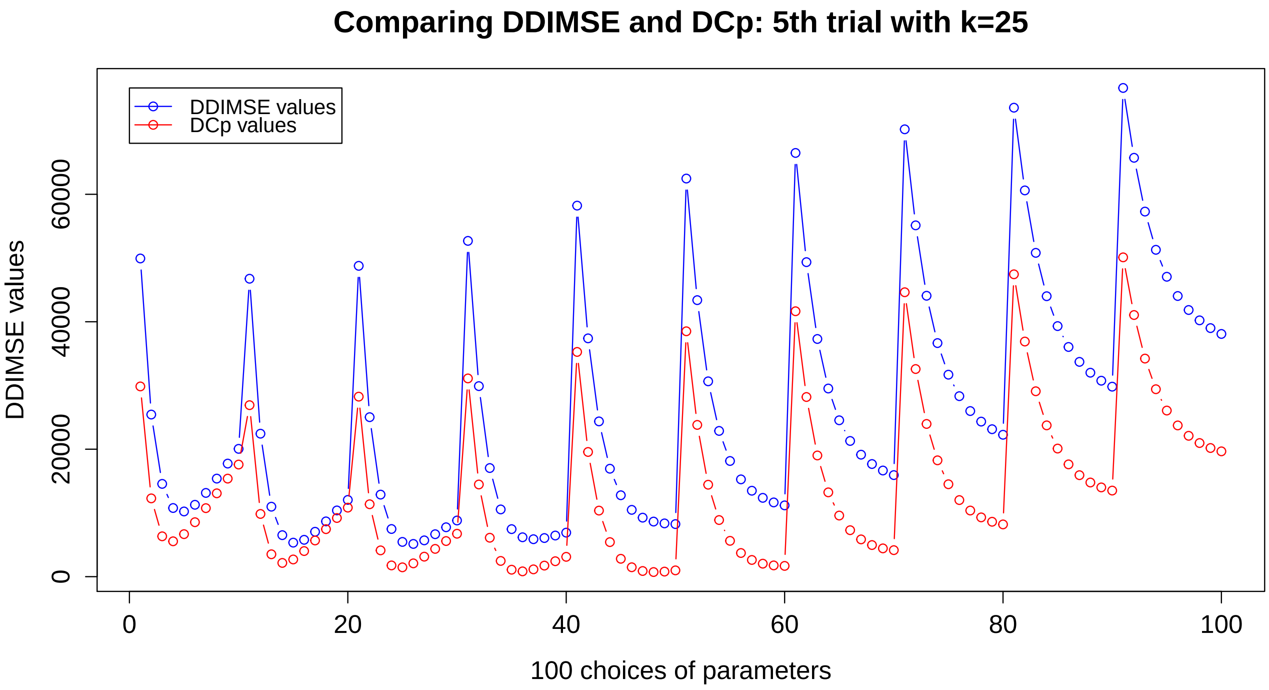


Figure 3.5: Plot of $\{h, \beta_n\}$ vs Cp.

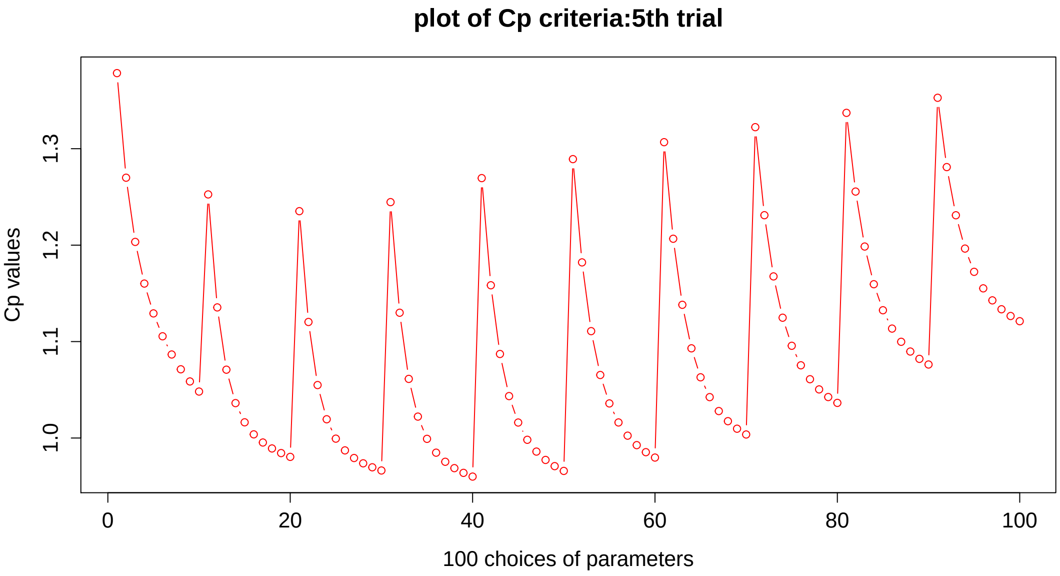


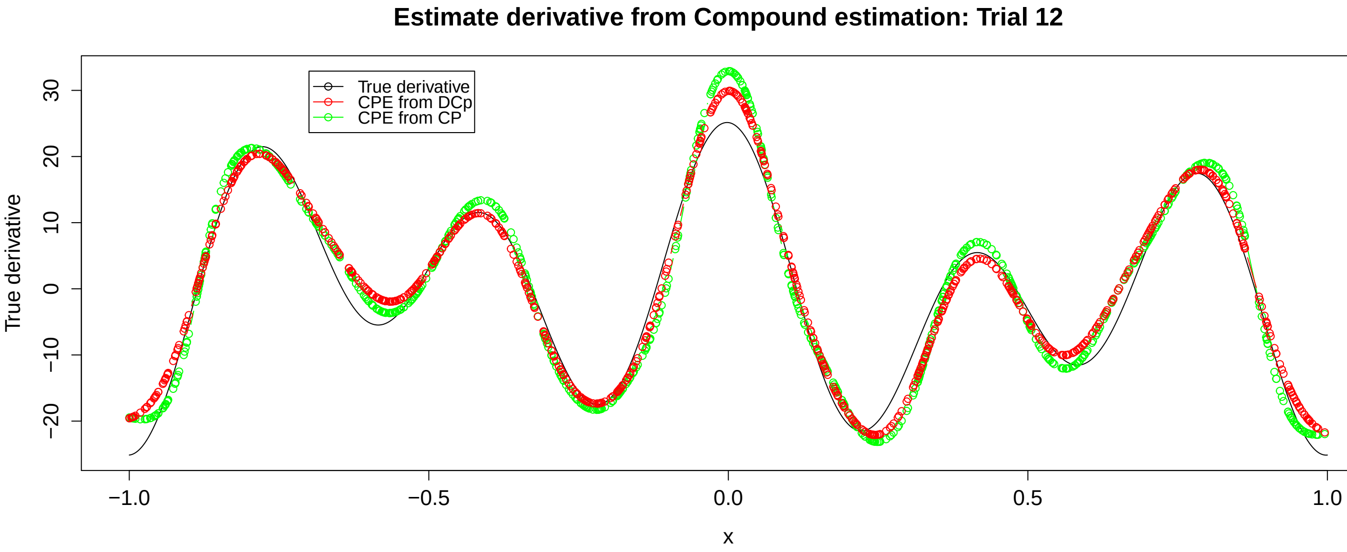
Table 3.3: Comparison for compound estimation with parameters selection

Trials	$DCpSum$	Cp	$CVE1kSum$	$GCV1kSum$
1	1.1283	1.6124	1.6124	20.018
2	1.4452	1.1194	1.1194	13.800
3	1.4718	1.7237	1.4230	21.407
4	1.4895	1.7333	7.5435	19.668
5	1.1440	1.3450	1.3450	14.929
6	1.0724	1.3279	1.3279	19.653
7	1.9226	1.7237	1.4230	21.407
8	1.4609	1.1735	1.1735	15.006
9	1.6091	1.8617	1.8617	27.177
10	1.7681	2.3759	6.6191	31.068
11	1.1340	1.5248	6.2000	14.415
12	1.0000	1.4814	1.4814	19.053
13	1.2599	1.6586	1.2411	22.090
14	1.3972	1.1392	1.1392	24.904
15	1.0154	1.3331	1.3331	14.393
16	1.3493	1.2718	1.2718	19.827
17	1.2756	1.3137	1.3137	27.535
18	1.1535	1.1860	1.1860	14.950
19	1.7379	1.3265	1.3265	22.814
20	1.1809	1.9607	1.9607	32.568

Table 3.4: Comparison for compound estimation with parameters selection

Methods	$DCpSum$	Cp	$CVE1kSum$	$GCV1kSum$
Wining times	12	5	3	0
Average $Q1$	1.3508	1.5096	2.1951	20.8342

Figure 3.6: Plot of estimated first derivative when choosing parameters from DCp or Cp.



Chapter 4 Jump Points Detection via Compound Estimation and Empirical Derivatives

4.1 Almost smooth function

An almost smooth mean response piecewise function is continuous with only finitely many discontinuities. The goal of this chapter is to find the location of a single discontinuity point. For example, in Figure 4.1, there are 500 data points that are generated from an almost smooth function $\mu(x) = \sin(2\pi x) + I(x \geq 0)[0.5 + \sin(2\pi x) + \cos(\pi x)]$ with random error $\epsilon_i \sim N(0, 0.9^2)$ in model (4.1). There is a discontinuity at $x = 0$ if we look at Figure 4.2. The red line is the mean response $\mu(x)$. However it is hard to visulize from the scatter plot. We will propose a method for detecting the jump point from the scatter plot.

For convenience, we continue to use some of the notations from Charnigo and Srinivasan [2011] and Chapter 2. Suppose $\mu(x)$ is the mean response of the nonparametric regression model with fixed design points x_i's that equally spaced on a compact interval $\mathcal{X} \subset \mathbb{R}$.

$$Y_i = \mu(x_i) + \epsilon_i, \tag{4.1}$$

where $\mu(x) = h(x) + I(x > x_0)f(x)$, and $h(x)$, $f(x)$ are continuous functions with at least $J + 1$th derivative and bounded away from 0 when x is close to x_0. Also, ϵ's are independent zero-mean random errors with equal variance σ^2. With out loss of generality, we assume $\mathcal{X} = [-1, 1]$.

4.2 Properties of "naive" compound estimation

From Figure 4.2, we can see there are three functions, $h(x)$, $h(x) + f(x)$ and $h(x) + I(x > 0)f(x)$, for simplicity, let $g(x) = h(x) + f(x)$ and $\mu(x) = h(x) + I(x > 0)f(x)$. And $x_0 = 0$ as in Figure 4.2. Suppose the compound estimator for functions $\mu(x)$ is

Figure 4.1: Scatter plot of the data points

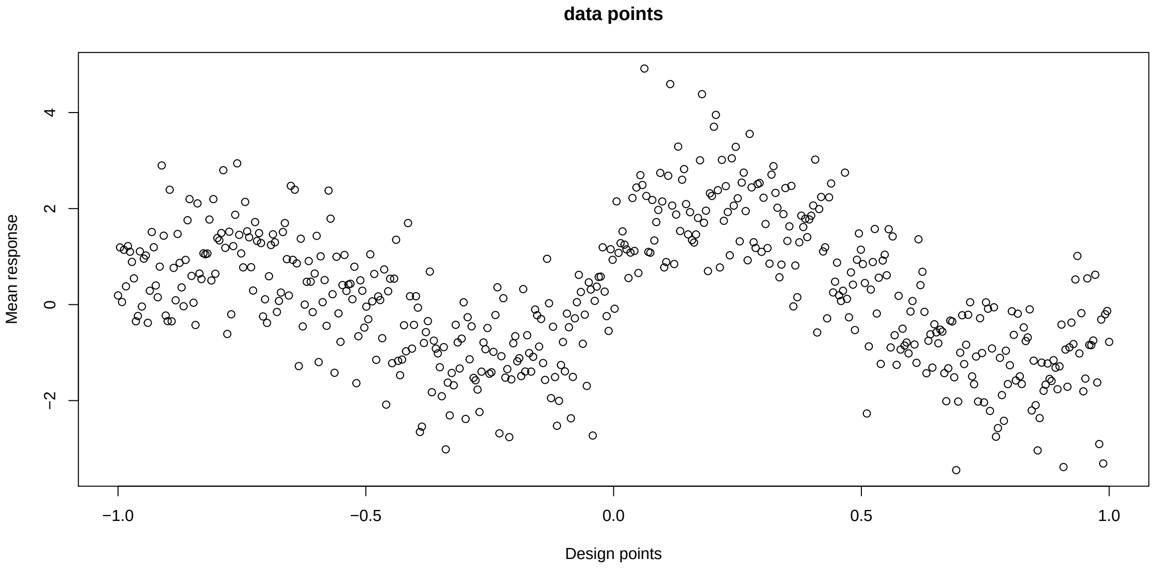
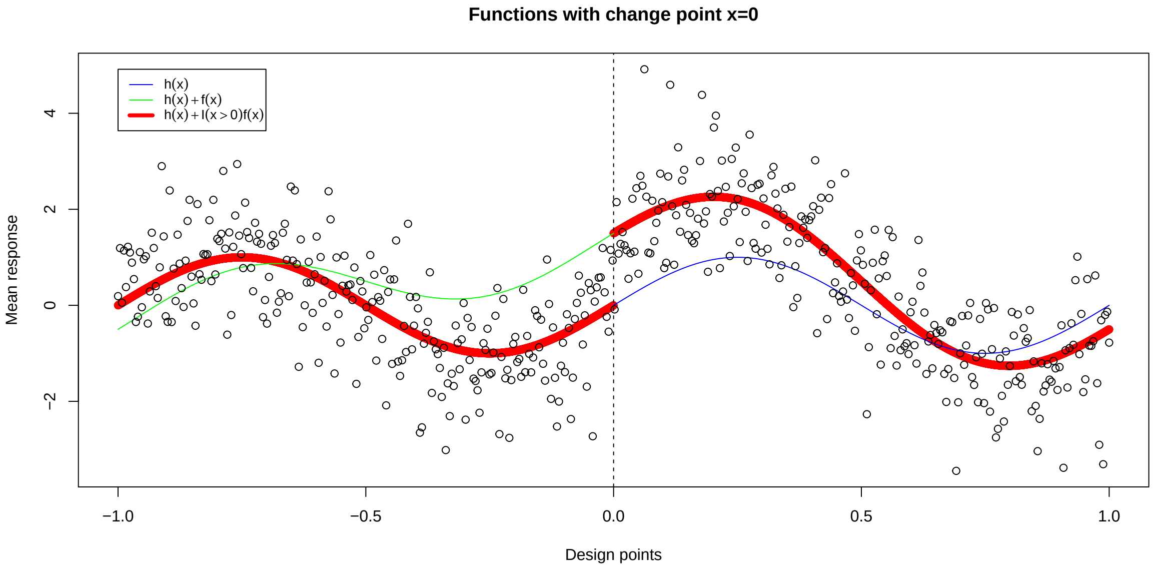


Figure 4.2: Almost smooth function



denoted by $\mu^\star(x)$. Also, suppose that conceptually we had access to two compound estimators for functions $h(x)$ and $g(x)$, we denote them as $h^\star(x)$ and $g^\star(x)$. Let $\widehat{h(a)}$, $\widehat{g(a)}$, $\widetilde{\mu(a)}$ be the pointwise estimators for $h(a)$, $g(a)$, and $\mu(a)$ at centering points $a's$. Then the compound estimators for these three functions are:

$$h^\star(x) = \sum_a W_n(x-a) \sum_{j=0}^{J} \widehat{h^{(j)}(a)}(x-a)^j/j! \tag{4.2}$$

$$g^\star(x) = \sum_a W_n(x-a) \sum_{j=0}^{J} \widehat{g^{(j)}(a)}(x-a)^j/j! \tag{4.3}$$

$$\mu^\star(x) = \sum_a W_n(x-a) \sum_{j=0}^{J} \widetilde{\mu^{(j)}(a)}(x-a)^j/j! \tag{4.4}$$

The goal for this section is to study some interesting properties of these compound estimators for detection of jump points. Suppose there is a change point x_0 in $\mu(x)$. First step of compound estimation is to get a set of pointwise estimators for each centering point. Therefore, we need to make some appropriate assumptions of these estimators to get valuable compound estimation. Even though the compound estimator of $\mu(x)$ will be naive when there is a discontinuity in the mean response, we still can get a "useful" naive estimator when we want to detect the jump point.

We assume the compound estimator of $h(x)$ and $g(x)$ is essentially optimal,then from Charnigo and Srinivasan [2011]

$$\sup_{x\in I} \left| \widehat{h^{(j)}(x)} - h^{(j)}(x) \right| \leq O_p(n^{-\frac{J-j+1}{2J+3}+v}) \tag{4.5}$$

$$\sup_{x\in I} \left| \widehat{g^{(j)}(x)} - g^{(j)}(x) \right| \leq O_p(n^{-\frac{J-j+1}{2J+3}+v}) \tag{4.6}$$

where v is an arbitrary small positive number, and $I \subset (-1, 1)$ is a compact interval. The assumptions about the convergence rate of $\mu(x)$ will be different because there is a discontinuity point in the function. When $a's$ are very close to x_0, the convergence rate of $\widetilde{\mu(a)}$ may not be very good because the pointwise estimators are obtained when assuming the mean response is continuous, which is actually not true at $x = x_0$. Thus, we need to consider the convergence rate when a is close to x_0. We incorporate this by assuming that the pointwise estimators at a small neighborhood of x_0 will be $O_p(1)$. Let a_1 be all the $a's$ on the left side of x_0 (or $a_1 \leq x_0$) and a_2 be all the $a's$ on the right side of x_0, and we assume that the pointwise estimators of $\mu(a)$ satisfy that if $|x_0 - a_1| > n^{-r_1}$ or $|x_0 - a_2| > n^{-r_1}$, then

$$\sup_{a_1 \in I_n} \left| \widetilde{\mu^{(j)}(a_1)} - h^{(j)}(a_1) \right| = O_p(n^{-d_j}) \tag{4.7}$$

$$\sup_{a_2 \in I_n} \left| \widetilde{\mu^{(j)}(a_2)} - g^{(j)}(a_2) \right| = O_p(n^{-d_j}) \tag{4.8}$$

$$\sup_{a_1 \in I_n} \left| \widetilde{\mu^{(j)}(a_1)} + f^{(j)}(a_1) - \widehat{g^{(j)}(a_1)} \right| = O_p(n^{-\alpha_j}) \tag{4.9}$$

$$\sup_{a_2 \in I_n} \left| \widetilde{\mu^{(j)}(a_2)} - f^{(j)}(a_2) - \widehat{h^{(j)}(a_2)} \right| = O_p(n^{-\alpha_j}), \tag{4.10}$$

where d_j and α_j are positive numbers satisfy that $0 < d_j \leq \frac{J-j+1}{2J+3} - v$ and $0 < \alpha_j \leq \frac{J-j+1}{2J+3} - v$, since $-\frac{J-j+1}{2J+3} + v$ is the essentially optimal convergence rate for compound estimator. Also, if $|x_0 - a_1| < n^{-r_1}$ or $|x_0 - a_2| < n^{-r_1}$, we assume

$$\sup_{a_1} \left| \widetilde{\mu^{(j)}(a_1)} - h^{(j)}(a_1) \right| = O_p(1) \tag{4.11}$$

$$\sup_{a_2} \left| \widetilde{\mu^{(j)}(a_2)} - g^{(j)}(a_2) \right| = O_p(1) \tag{4.12}$$

$$\sup_{a_1} \left| \widetilde{\mu^{(j)}(a_1)} + f^{(j)}(a_1) - \widehat{g^{(j)}(a_1)} \right| = O_p(1) \tag{4.13}$$

$$\sup_{a_2} \left| \widetilde{\mu^{(j)}(a_2)} - f^{(j)}(a_2) - \widehat{h^{(j)}(a_2)} \right| = O_p(1), \tag{4.14}$$

The above assumptions (4.11)-(4.14) say that in the neighborhood of x_0, the "naive" compound estimator may have a bad performance.

Theorem 4.2.1 *Suppose (4.1)-(4.14) hold, let I_n be the set of all the gridding points a's, $\delta = \frac{v}{4J+6}$, $r = \frac{1}{2J+3}$, $r_1 > 0$, then the naive compound estimator $\mu^*(x)$ in (4.4)* *satisfy*

$$|\frac{d}{dx}\mu^\star(x) - \frac{d}{dx}\mu(x)| = O(n^{4\delta + \frac{1}{2J+3}}) + O_p(n^{4\delta + \frac{2}{2J+3} - r_1}), \tag{4.15}$$

for $x \neq x_0$. However, we need to notice that formula (4.15) is obviously not sharp in any neighborhood away from x_0.

Proof:

First, suppose $x < x_0$, Then we'll have

$$\begin{aligned}
&\mu^\star(x) - \mu(x) \\
=&\mu^\star(x) - h^\star(x) + h^\star(x) - h(x) \\
=&\sum_a W_n(x-a) \sum_{j=0}^{J} [\widetilde{\mu^{(j)}(a)} - \widehat{h^{(j)}(a)}](x-a)^j + [h^\star(x) - h(x)] \\
=&\sum_{a_1} W_n(x-a_1) \sum_{j=0}^{J} [\widetilde{h^{(j)}(a_1)} - \widehat{h^{(j)}(a_1)}](x-a_1)^j + \sum_{a_2} W_n(x-a_2) \sum_{j=0}^{J} [\widetilde{g^{(j)}(a_2)} - \widehat{h^{(j)}(a_2)}](x-a_2)^j \\
&+ [h^\star(x) - h(x)] \\
=&\sum_{a_1} W_n(x-a_1) \sum_{j=0}^{J} [\widetilde{h^{(j)}(a_1)} - \widehat{h^{(j)}(a_1)}](x-a_1)^j \\
&+ \sum_{a_2} W_n(x-a_2) \sum_{j=0}^{J} [\widetilde{g^{(j)}(a_2)} - \widehat{h^{(j)}(a_2)} - f^{(j)}(a_2)](x-a_2)^j \\
&+ \sum_{a_2} W_n(x-a_2) \sum_{j=0}^{J} f^{(j)}(a_2)(x-a_2)^j + [h^\star(x) - h(x)].
\end{aligned} \tag{4.16}$$

From Charnigo and Srinivasan [2011], let $r = \frac{1}{2J+3}$, $I_{1n}(x) = \{a \in I_n : |a-x| < n^{-r}\}$, $\phi = \delta = v/(4J+6)$ and v is an abitrary small positive value, also, the number of centering points $L_n = \Theta(n^{r+\delta+\phi})$. The Θ sign means that there exist two positive constant k_1 and k_2 such that $k_1 n^{r+\delta+\phi} \leq |L_n| \leq k_2 n^{r+\delta+\phi}$. They showed the following result:

$$\sup_{a \in I_{1n}(x)} |\frac{d^k}{dx^k} W_n(x-a)| = O(n^{3k\delta - \phi + kr}). \tag{4.17}$$

Also following Charnigo and Srinivasan (2011), we know that

$$\sup_{a \in \overline{I}_{1n}(x)} |\frac{d^k}{dx^k} W_n(x-a)| = \frac{\Theta(L_n^k \beta_n^k) \exp[-\beta_0 n^{2\delta}]}{\left(\sum_{c \in I_n} \exp[-\beta_n (x-c)^2]\right)^{(k+1)}}, \tag{4.18}$$

which will decay in exponential rate, and the numbers of $a's$ in I_n is $\Theta(n^{r+\delta+\phi})$. Thus, summand of $|\frac{d^k}{dx^k} W_n(x-a)|$ when $a \in \overline{I}_{1n}(x)$ could be negligible.

Let $I_{2n}(x) = \{a \in I_n : |a-x| < n^{-r_1}\}$. Then

$$\begin{aligned}
&\sum_{a_1} W_n'(x-a_1)[\widetilde{h^{(j)}(a_1)} - \widehat{h^{(j)}(a_1)}](x-a_1)^j && (4.19)\\
&= \sum_{a_1 \in I_{2n}(x)} W_n'(x-a_1) O_p(n^{-d_j})(x-a_1)^j + \sum_{a_1 \in \overline{I}_{2n}(x)} W_n'(x-a_1) O_p(1)(x-a_1)^j \\
&= O(n^{-r}) \times \Theta(n^{r+\phi+\delta}) \times O(n^{3\delta-\phi+r}) \times O_p(n^{-d_j}) \times O(n^{-jr}) \\
&+ O(n^{-r_1}) \times \Theta(n^{r+\phi+\delta}) \times O(n^{3\delta-\phi+r}) \times O_p(1) \times O(n^{-jr_1}) \\
&= O_p(n^{4\delta-d_j-(j-1)r}) + O_p(n^{4\delta+2r-(j+1)r_1}). && (4.20)
\end{aligned}$$

Also,

$$\begin{aligned}
&\sum_{a_1} W_n(x-a_1)[\widetilde{h^{(j)}(a_1)} - \widehat{h^{(j)}(a_1)}] j(x-a_1)^{j-1} && (4.21)\\
&= \sum_{a_1 \in I_{2n}(x)} W_n(x-a_1) O_p(n^{-d_j}) j(x-a_1)^{j-1} + \sum_{a_1 \in \overline{I}_{2n}(x)} W_n(x-a_1) O_p(1) j(x-a_1)^{j-1} \\
&= O(n^{-r}) \times \Theta(n^{r+\phi+\delta}) \times O(n^{-\phi}) \times O_p(n^{-d_j}) \times O(n^{-(j-1)r}) \\
&+ O(n^{-r_1}) \times \Theta(n^{r+\phi+\delta}) \times O(n^{-\phi}) \times O_p(1) \times O(n^{-jr_1+r}) \\
&= O_p(n^{\delta-d_j-(j-1)r}) + O_p(n^{\delta-jr_1+r}). && (4.22)
\end{aligned}$$

In fact, (4.20) and (4.22) hold for all $1 \leq j \leq J$. These two results lead to the convergence rate of following term,

$$\begin{aligned}
&\frac{d}{dx}\left[\sum_{a_1} W_n(x-a_1)[\widetilde{h^{(j)}(a_1)}-\widehat{h^{(j)}(a_1)}](x-a_1)^j\right] \qquad (4.23)\\
&=\sum_{a_1} W_n'(x-a_1)[\widetilde{h^{(j)}(a_1)}-\widehat{h^{(j)}(a_1)}](x-a_1)^j+\sum_{a_1} W_n(x-a_1)[\widetilde{h^{(j)}(a_1)}-\widehat{h^{(j)}(a_1)}]j(x-a_1)^{j-1}\\
&=O_p(n^{4\delta-d_j-(j-1)r})+O_p(n^{4\delta+2r-(j+1)r_1})+O_p(n^{\delta-d_j-(j-1)r})+O_p(n^{\delta-jr_1+r})\\
&=O_p(n^{4\delta-d_j-(j-1)r})+O_p(n^{4\delta+2r-(j+1)r_1})+O_p(n^{\delta-jr_1+r}). \qquad (4.24)
\end{aligned}$$

When $j=0$, term (4.23) will be

$$\begin{aligned}
&\sum_{a_1} W_n'(x-a_1)[\widetilde{h(a_1)}-\widehat{h(a_1)}]\\
&=\sum_{a_1\in I_{2n}(x)} W_n'(x-a_1)O_p(n^{-d_j})+\sum_{a_1\in \bar{I}_{2n}(x)} W_n'(x-a_1)O_p(1)\\
&=O(n^{-r})\times\Theta(n^{r+\phi+\delta})\times O(n^{3\delta-\phi+r})\times O_p(n^{-d_j})+O(n^{-r_1})\times\Theta(n^{r+\phi+\delta})\times O(n^{3\delta-\phi+r})\times O_p(1)\\
&=O_p(n^{4\delta-d_j+r})+O_p(n^{4\delta+2r-r_1}). \qquad (4.25)
\end{aligned}$$

The same deduction will show

$$\begin{aligned}
&\frac{d}{dx}\left[\sum_{a_2} W_n(x-a_2)[\widetilde{g^{(j)}(a_2)}-\widehat{h^{(j)}(a_2)}-f^{(j)}(a_2)](x-a_2)^j\right] \qquad (4.26)\\
&=\sum_{a_2} W_n'(x-a_2)[\widetilde{g^{(j)}(a_2)}-\widehat{h^{(j)}(a_2)}-f^{(j)}(a_2)](x-a_2)^j\\
&+\sum_{a_2} W_n(x-a_2)[\widetilde{g^{(j)}(a_2)}-\widehat{h^{(j)}(a_2)}-f^{(j)}(a_2)]j(x-a_2)^{j-1}\\
&=O_p(n^{4\delta-\alpha_j-(j-1)r})+O_p(n^{4\delta+2r-(j+1)r_1})+O_p(n^{\delta-jr_1+r}), \qquad (4.27)
\end{aligned}$$

where $1 \leq j \leq J$.

If $j = 0$, (4.26) will be

$$\begin{aligned}
&\sum_{a_2} W_n'(x-a_2)[\widetilde{g(a_2)} - \widehat{h(a_2)} - f(a_2)](x-a_2)^j \\
&= O(n^{-r}) \times \Theta(n^{r+\phi+\delta}) \times O(n^{3\delta-\phi+r}) \times O_p(n^{-\alpha_j}) + O(n^{-r_1}) \times \Theta(n^{r+\phi+\delta}) \times O(n^{3\delta-\phi+r}) \times O_p(1) \\
&= O_p(n^{4\delta-\alpha_j+r}) + O_p(n^{4\delta+2r-r_1}). \qquad (4.28)
\end{aligned}$$

Now, let's look at the third term at RHS of (4.16). Since each $|f^{(j)}(a)|$ is bounded above by some constant, let

$$T(x) := \sum_{a_2} W_n(x-a_2) \sum_{j=0}^{J} f^{(j)}(a_2)(x-a_2)^j \qquad (4.29)$$

$$T_j(x) := \sum_{a_2} W_n(x-a_2) f^{(j)}(a_2)(x-a_2)^j \qquad (4.30)$$

$$T(x) = \sum_{j=0}^{J} T_j(x). \qquad (4.31)$$

When $j = 0$,

$$\begin{aligned}
T_0'(x) =& \sum_{a_2} W_n'(x-a_2) f(a_2) \\
=& O(n^{-r})\Theta(n^{r+\phi+\delta}) \times O(n^{3\delta-\phi+r}) \\
=& O(n^{4\delta+r}). \qquad (4.32)
\end{aligned}$$

When $1 \leq j \leq J$,

$$T_j'(x) = \sum_{a_2} W_n(x-a_2) f^{(j)}(a_2) j (x-a_2)^{j-1} + \sum_{a_2} W_n'(x-a_2) f^{(j)}(a_2)(x-a_2)^j. \quad (4.33)$$

The first term on RHS of (4.33) is

$$\begin{aligned}
&\sum_{a_2} W_n(x-a_2)f^{(j)}(a_2)j(x-a_2)^{j-1}\\
=&\sum_{a_2} W_n(x-a_2)(x-a_2)^{j-1}f^{(j)}(a_2)\\
=&O(n^{-r})\times\Theta(n^{r+\phi+\delta})\times O(n^{-\phi})\times O(n^{-(j-1)r})\\
=&O(n^{\delta-(j-1)r}). \qquad (4.34)
\end{aligned}$$

The second term on RHS of (4.33) is

$$\begin{aligned}
&\sum_{a_2} W_n'(x-a_2)f^{(j)}(a_2)(x-a_2)^{j}\\
=&O(n^{-r})\times\Theta(n^{r+\phi+\delta})\times O(n^{3\delta-\phi+r})\times O(n^{-jr})\times O(n^{\delta+\phi})\\
=&O(n^{4\delta-(j-1)r}). \qquad (4.35)
\end{aligned}$$

By (4.34) and (4.35),

$$T_j'(x)=O(n^{4\delta-(j-1)r}). \qquad (4.36)$$

Because there are finitely many of $T_j'(x)'s$, we'll have

$$T'(x)=\sum_{j=0}^{J}T_j'(x)=O(n^{4\delta+r}). \qquad (4.37)$$

Combine (4.5), (4.24), (4.25),(4.27), (4.28),(4.29), (4.37), suppose $x < x_0$

$$
\begin{aligned}
&\mu^\star(x)' - \mu'(x) \\
=&\frac{d}{dx}\left[\sum_{a_1} W_n(x-a_1)\sum_{j=0}^{J}[\widetilde{h^{(j)}(a_1)} - \widehat{h^{(j)}(a_1)}](x-a_1)^j\right] \\
&+ \frac{d}{dx}\left[\sum_{a_2} W_n(x-a_2)\sum_{j=0}^{J}[\widetilde{g^{(j)}(a_2)} - \widehat{h^{(j)}(a_2)} - f^{(j)}(a_2)](x-a_2)^j\right] \\
&+ T'(x) + [h^\star(x)' - h(x)] \\
=&\sum_{j=0}^{J}\left[O_p(n^{4\delta - d_j - (j-1)r}) + O_p(n^{4\delta+2r-(j+1)r_1})\right] + \sum_{j=1}^{J} O_p(n^{\delta - jr_1 + r}) \\
&+ \sum_{j=0}^{J}\left[O_p(n^{4\delta - \alpha_j - (j-1)r}) + O_p(n^{4\delta+2r-(j+1)r_1})\right] + \sum_{j=1}^{J} O_p(n^{\delta - jr_1 + r}) \\
&+ O(n^{4\delta + r}) + O_p(n^{-\frac{J-j+1}{2J+3}+v}) \\
=&O_p(n^{4\delta - \min_{j\in\{0,1\dots J\}}\{d_j + jr\} + r}) + O_p(n^{4\delta+2r-r_1}) + O_p(n^{\delta+r-r_1}) + O_p(n^{4\delta - \min_{j\in\{0,1\dots J\}}\{\alpha_j + jr\} + r}) \\
&+ O(n^{4\delta + r}) + O_p(n^{-\frac{J-j+1}{2J+3}+v}) \\
=&O(n^{4\delta+r}) + O_p(n^{4\delta+2r-r_1}). \qquad (4.38)
\end{aligned}
$$

The same procedure as before will also lead to the results when $x > x_0$. That is, if $x > x_0$, and we assume that (4.6), (4.8), (4.10),(4.12) and (4.14) hold, then

$$
\begin{aligned}
\mu^\star(x)' - \mu'(x) &= \mu^\star(x)' - g'(x) \\
&= O(n^{4\delta+r}) + O_p(n^{4\delta+2r-r_1}). \qquad (4.39)
\end{aligned}
$$

In Charnigo and Srinivasan [2011], $r = \frac{1}{2J+3}$. Then, from (4.38) and (4.39), (4.15) holds. ■

For convenience, let $\max\{r, (2r - r_1)\} = \eta$, we'll have

$$\mu^{\star}(x)' - \mu'(x) = O_p(n^{4\delta+\eta}). \tag{4.40}$$

Therefore, (4.39) and (4.40) show an upper bound for divergence when x is in the neighborhood of x_0 if compound estimation was performed naively. When x is away from jump point x_0, by Charnigo and Srinivasan [2011], $\mu'(x)$ could be estimated very well. Thus, if x is away from jump point x_0, then

$$\sup_{x \subset I \backslash [x_0 - \tau_n, x_0 + \tau_n]} \left| \widetilde{\mu^{(j)}(x)} - \mu^{(j)}(x) \right| = O_p(n^{-\frac{J-j+1}{2J+3}+v}). \tag{4.41}$$

Here, τ_n is a sequence of positive numbers depend on sample size n such that the estimator of $\mu^{(j)}(x)$ when x is outside of $[x_0 - \tau_n, x_0 + \tau_n]$ will not be affected by the change point x_0. For example, τ_n could be the bandwidth h_n of local regression. Equation (4.39) and (4.41) tell us the distance between "naive" compound estimator of first derivative and the true mean response function when x is away from x_0 or x is in the neighborhood of x_0. However, we don't know the true derivative. Suppose we could find another first derivative estimator and its distance from $\mu'(x)$ will behave different whether x is in the neighborhood of x_0 or not. Then we can look at the gap between this estimator and the naive compound estimator. Large gap may imply the location of the jump point. The next section will show the Empirical first derivative works for this purpose.

4.3 Properties of Empirical first derivatives

The Empirical first derivative was defined as (3.2) in chapter 3. From Charnigo, Hall and Srinivasan [2011], they let the weights of empirical derivative to be $w_j = \frac{j^2}{\sum_{m=1}^{k} m^2}$.

then

$$\frac{w_j}{x_{i+j} - x_{i-j}} = \frac{3}{2}\frac{nj}{k(k+1)(2k+1)}, \tag{4.42}$$

since x's are equally spaced on $[-1, 1]$. Also, alone with these weights and $k = \Theta(n^\alpha)$, they proved that

$$Var[Y_i^{(1)}] = O(n^{2-3\alpha}) \qquad\qquad Bias[Y_i^{(1)}] = O(n^{\alpha-1}), \tag{4.43}$$

uniformly for $k + 1 \le i \le n - k$.

Assume jump point $x_0 \in (x_s, x_{s+1})$ and $x_i \le x_s$, then from the equation above, the empirical first derivative of x_i will be

$$\begin{aligned}
Y_i^{(1)} &= \sum_{j=1}^{k} w_i \frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}} \\
&= \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} \left[h(x_{i+j}) - h(x_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} + f(x_{i+j}) I[x_0 \in (x_i, x_{i+j})]\right] \\
&= \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} \left[h(x_{i+j}) - h(x_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j}\right] + \sum_{j=1}^{k} \frac{w_j f(x_{i+j})}{x_{i+j} - x_{i-j}} I[x_0 \in (x_i, x_{i+j})] \\
&= \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} \left[h(x_{i+j}) - h(x_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j}\right] \\
&\quad + \frac{3}{2}\frac{n}{k(k+1)(2k+1)} \sum_{j=1}^{k} f(x_{i+j}) I\left(i \le s < s+1 \le i+j\right) \\
&= \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} \left[h(x_{i+j}) - h(x_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j}\right] \\
&\quad + \frac{3}{2}\frac{n}{k(k+1)(2k+1)} \sum_{j=s+1-i}^{k} j f(x_{i+j}).
\end{aligned} \tag{4.44}$$

Thus, if we let $k = \Theta(n^\alpha)$, then from (4.44),

$$\begin{aligned}
Y_i^{(1)} - \mu'(x_i) =& Y_i^{(1)} - E[Y_i^{(1)}] + E[Y_i^{(1)}] - h'(x_i) \\
=& Y_i^{(1)} - E[Y_i^{(1)}] + \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} \left[h(x_{i+j}) - h(x_{i-j})\right] - h'(x_i) \\
&+ \frac{3}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=s+1-i}^{k} j f(x_{i+j}) \\
=& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=s+1-i}^{k} j f(x_{i+j}). \quad (4.45)
\end{aligned}$$

Without loss of generality, we can assume $f(x_0) > 2C > 0$ where $2C$ is a fixed constant. Since $k = \Theta(n^\alpha)$ implies $|x_{i+k} - x_{i-k}| = O(n^{\alpha-1})$ and $f(x)$ is smooth, when n is sufficient large, we'll have $f(x_{s+j}) > f(x_0)/2 > C$ for all $j \in \{1, 2, ...k\}$.
If $i = s$, then from (4.45),

$$\begin{aligned}
Y_i^{(1)} - \mu'(x_i) =& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=1}^{k} j f(x_{s+j}) \\
\geq& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3C}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=1}^{k} j \\
=& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3C}{4} \frac{n}{(2k+1)} \\
=& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \Theta(n^{1-\alpha}). \quad (4.46)
\end{aligned}$$

If $i = s + 1 - k$, then

$$\begin{aligned}
Y_i^{(1)} - \mu'(x_i) =& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3}{2} \frac{n}{(k+1)(2k+1)} f(x_{i+k}) \\
=& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \Theta(n^{1-2\alpha}). \quad (4.47)
\end{aligned}$$

If $i \leq s - k$, then

$$Y_i^{(1)} - \mu'(x_i) = O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}). \tag{4.48}$$

Then let's look at the empirical first derivative of x_i if $x_i \geq x_{s+1}$. Again from (4.42) and (4.43), we'll have

$$\begin{aligned}
Y_i^{(1)} =& \sum_{j=1}^{k} w_i \frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}} \\
=& \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} \left[g(x_{i+j}) - g(x_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} + f(x_{i-j}) I[x_0 \in (x_{i-j}, x_i)] \right] \\
=& \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} \left[g(x_{i+j}) - g(x_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} \right] + \sum_{j=1}^{k} \frac{w_j f(x_{i-j})}{x_{i+j} - x_{i-j}} I[x_0 \in (x_{i-j}, x_i)] \\
=& \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} \left[g(x_{i+j}) - g(x_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} \right] \\
&+ \frac{3}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=1}^{k} f(x_{i-j}) I\left(i - j \leq s < s + 1 \leq i\right) \\
=& \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} \left[g(x_{i+j}) - g(x_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} \right] \\
&+ \frac{3}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=i-s}^{k} j f(x_{i-j}).
\end{aligned} \tag{4.49}$$

From (4.49) and the same deduction as (4.45),

$$\begin{aligned}Y_i^{(1)} - \mu'(x_i) =& Y_i^{(1)} - E[Y_i^{(1)}] + E[Y_i^{(1)}] - g'(x_i) \\ =& Y_i^{(1)} - E[Y_i^{(1)}] + \sum_{j=1}^{k} \frac{w_j}{x_{i+j} - x_{i-j}} [g(x_{i+j}) - g(x_{i-j})] - g'(x_i) \\ &+ \frac{3}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=i-s}^{k} jf(x_{i-j}) \\ =& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=i-s}^{k} jf(x_{i-j}). \qquad (4.50)\end{aligned}$$

Again, when n is sufficient large, we'll have $f(x_{s+1-j}) > f(x_0)/2 > C$ for all $j \in \{1, 2, ...k\}$. For the case $x_i > x_{s+1}$, we'll have the following results.
If $i = s + 1$, then from (4.50),

$$\begin{aligned}Y_i^{(1)} - \mu'(x_i) =& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=1}^{k} jf(x_{s+1-j}) \\ \geq& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3C}{2} \frac{n}{k(k+1)(2k+1)} \sum_{j=1}^{k} j \\ =& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3C}{4} \frac{n}{(2k+1)} \\ =& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \Theta(n^{1-\alpha}). \qquad (4.51)\end{aligned}$$

If $i = s + k$, then

$$\begin{aligned}Y_i^{(1)} - \mu'(x_i) =& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \frac{3}{2} \frac{n}{(k+1)(2k+1)} f(x_{i-k}) \\ =& O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \Theta(n^{1-2\alpha}) \qquad (4.52)\end{aligned}$$

If $i \geq s + k + 1$, then

$$Y_i^{(1)} - \mu'(x_i) = O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}). \qquad (4.53)$$

The above results gave us the hint about the gap between Empirical first derivative and $\mu'(x)$ whether x_i is away from x_0 or not. These results will be applied in the next section for the purpose of detecting the jump point.

4.4 Detection of the location of jump point

The last two sections showed us some properties of naive compound estimator and empirical first derivatives when there is a discontinuity in the function. In this section, we will try to detect the interval estimator of the jump point x_0. The idea is that if we appropriately choose k for the empirical first derivative, then the distance between naive compound estimator and empirical first derivative will behave differently when x is outside of neighborhood of x_0 comparing to x is in the neighborhood of x_0.

Theorem 4.4.1 *Suppose the change point x_0 is in the interval $[x_{s_n}, x_{s_n+1})$, and the order of Empirical first derivative is $k_n = \Theta(n^\alpha)$. Let t_n be the integer such that*

$$|Y_{t_n}^{(1)} - \mu^\star(x_{t_n})'| = \max_{i \in \{k+1,\ldots,n-k\}} \left\{ |Y_i^{(1)} - \mu^\star(x_i)'| \right\}, \tag{4.54}$$

and $I_n = (x_{t_n-k_n}, x_{t_n+k_n})$. Then there exist α such that $P(x_0 \in I_n) \longrightarrow 1$.

Proof: By triangle inequality,

$$\left| Y_i^{(1)} - \mu^\star(x_i)' \right| \geq \left| Y_i^{(1)} - \mu'(x_i) \right| - |\mu^\star(x_i)' - \mu'(x_i)| \tag{4.55}$$

$$\left| Y_i^{(1)} - \mu^\star(x_i)' \right| \leq \left| Y_i^{(1)} - \mu'(x_i) \right| + |\mu^\star(x_i)' - \mu'(x_i)|\,. \tag{4.56}$$

Choose α such that $1 - \frac{3}{2}\alpha > 4\delta + \eta$, or equivalently, $\alpha < \frac{2}{3}(1 - 4\delta - \eta)$, then $\alpha - 1 - (1 - \frac{3}{2}\alpha) = \frac{5}{2}\alpha - 2 < -\frac{1}{3} < 0$.

If $i = s_n$, from (4.40) and (4.46) we'll have

$$\begin{aligned} \left| Y_i^{(1)} - \mu^\star(x_i)' \right| &\geq O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + \Theta(n^{1-\alpha}) - O_p(n^{4\delta+\eta}) \\ &= O_p(n^{1-\frac{3}{2}\alpha}) + \Theta_p(n^{1-\alpha}). \end{aligned} \tag{4.57}$$

Also, if $i \geq s_n + k_n$ or $i \leq s_n - k_n + 1$, then by (4.47), (4.48), (4.52) and (4.53),

$$\begin{aligned} \left| Y_i^{(1)} - \mu^\star(x_i)' \right| &\leq O_p(n^{1-\frac{3}{2}\alpha}) + O(n^{\alpha-1}) + O(n^{4\delta+\eta}) \\ &= O_p(n^{1-\frac{3}{2}\alpha}). \end{aligned} \tag{4.58}$$

Suppose $t_n \leq s_n - k_n + 1$ or $t_n \geq s_n + k_n$, then (4.58) tell us $\left| Y_{t_n}^{(1)} - \mu^\star(x_{t_n})' \right| = O_p(n^{1-\frac{3}{2}\alpha}) < \Theta_p(n^{1-\alpha}) \leq \left| Y_{s_n}^{(1)} - \mu^\star(x_{s_n})' \right|$, which is contradictory to that (4.54). Therefore we must have

$$P(s_n - k_n < t_n < s_n + k_n) \to 1, \tag{4.59}$$

$\Rightarrow$

$$P(t_n - k_n < s_n < t_n + k_n) \to 1, \tag{4.60}$$

and

$$P(t_n - k_n + 1 < s_n + 1 < t_n + k_n + 1) \to 1, \tag{4.61}$$

$\Rightarrow$

$$P(x_{s_n} \in I_n) \to 1, \tag{4.62}$$

and

$$\begin{aligned}
P(x_{s_n+1} \in I_n) &= P(s_n - k_n < s_n + 1 < s_n + k_n) \\
&\geq P(s_n - k_n + 1 < s_n + 1 < s_n + k_n) \\
&= P(s_n - k_n + 1 < s_n + 1 < s_n + k_n + 1) - P(s_n = t_n + k_n - 1) \\
&= 1 - P(t_n = s_n + 1 - k_n) \\
&= 1. \qquad (4.63)
\end{aligned}$$

Then (4.62) and (4.63) lead to $P(x_0 \in I_n) \to 1$. ■

Corollary 4.4.2 *The length of inteval I_n, $len(I_n) \to 0$ as $n \to \infty$.*

Proof: It is easy to see that $len(I_n) = x_{t_{2n}} - x_{t_{1n}} \leq \Theta(n^\alpha)/n = \Theta(n^{\alpha-1}) \to 0$. ■

4.5 Simulation study

Simulation study was done by two different scenarios. In the first scenarios, we let $f(x)$ be a fixed constant, specifically, $f(x) = 1$ for any $x \in [-1, 1]$ and 1000 samples were generated from the model (4.1). The mean response is $\mu(x) = \sin(2\pi x) + \cos(\pi x) + I(x \geq -0.2)$. The random error $\epsilon_i \sim N(0, 0.8^2)$. Figure 4.4 displays the location of jump point and the mean response. However, that is not clear at all if we look at Figure 4.3. In real case, we can only observe the scatter plot like Figure 4.3, then the jump detection method could be applied for the situation like this. For the purpose of finding interval estimator of x_0, we need to fit the "naive" compound estimator of $\mu'(x)$. The pointwise estimators a's could be fit with different nonparameteric regression method in the first step of compound estimation. We choose local regression as before in chapter 2 and chapter 3. Then the tuning parameter selection problem arise because we need to choose bandwidth h in the pointwise estimation step and the gaussian convolution weight β_n in formula (2.4). In chapter 3, we develop the DCp criteria for tuning parameter selection of nonparametric first derivative estimation. However, this should not be used for the "naive" compound

Figure 4.3: The observed data points

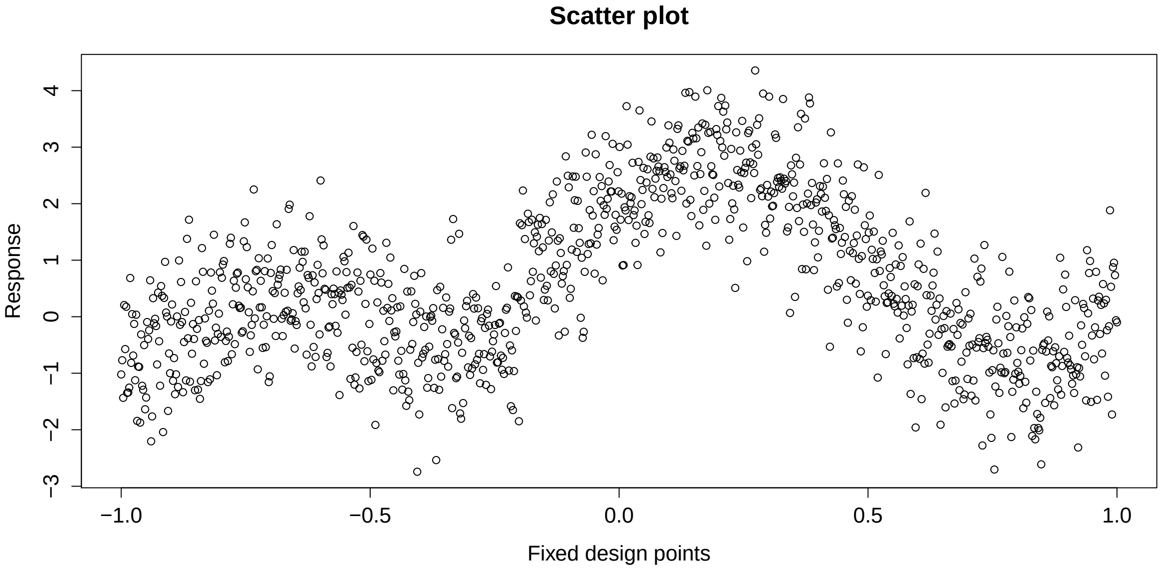
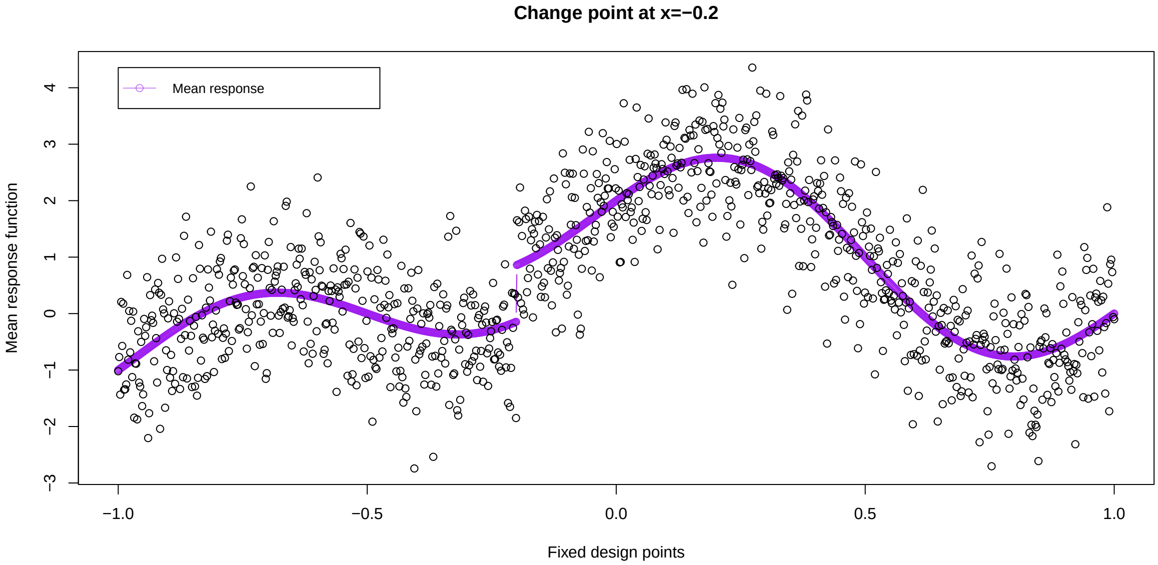


Figure 4.4: The mean response function



estimation because the $S_n(\lambda)$ in (3.26) is not negligible when there is a jump point x_0. Then the proxy like (3.17) won't work since the DCp is highly affected by the part we ignored. Nevertheless, we still have a way to obtain a fairly good "naive" compound estimator. In chapter 3, we mentioned that if the function $\mu(x)$ is smooth, the derivative estimation will be competitive if we choose bandwidth from Cp criteria and β_n from our experience ($\beta_n < 100$). Here, even if the mean response $\mu(x)$ has one jump point, the Cp could still work. In theorem 2.3.2, we can see that the $DIMSE$ is strictly equal to $E(Cp)$ plus a constant and this does not require $\mu(x)$ to be a strictly smooth function. Thus we can still approximately minimize $DIMSE$ by minimizing C_p, which tells us that if x is away from the jump point x_0, $\mu^*(x)$ should be very close to $\mu(x)$. Since $\mu(x)$ is almost smooth and compound estimator is self-consistent, then the derivative of both should also be close when x is not in the neighborhood of x_0. Therefore, we fixed $\beta_n = 30$ for the compound estimation and applied C_p to pick the bandwidth h. These may not give us the best tuning parameters for the "naive" compound estimation, but it will give us a moderately good one to present a big gap from the Empirical first derivative when x is in the neighborhood of x_0. Figure 4.5 and Figure 4.6 show us the "naive" compound estimator for the mean response and the first derivative. We used 40 a's equally spaced on the interval $[-0.98, 0.98]$ and the order of the local regression is 2. The estimation behaves very well except in the neighborhood of the jump $x_0 = -0.2$.

Next step is to attain the Empirical first derivative, which depends on the order k. We let $k = n^\alpha$ and $n = 1000$ is the sample size, then we pick α by maximizing a quantity Q_n. Let $l = \arg\max_{i \in \{k+1, \ldots n-k\}} |Y_i^{(1)} - \mu^*(x_i)|$, then Q_n is defined as

$$Q_n = \frac{|Y_l^{(1)} - \mu^*(x_l)|}{\max_{j \notin \{l-k, \ldots l+k\} \cup \{1, \ldots k\} \cup \{n-k+1, \ldots n\}} |Y_j^{(1)} - \mu^*(x_j)|} \tag{4.64}$$

This is for picking α which can present the biggest signal of the jump point. Since the Empirical derivative will behave badly on the boundary, we ignore the first and the last k data points when calculating Q_n. In our simulation, we select α from 100

Figure 4.5: CPE from the bandwidth chosen from Cp

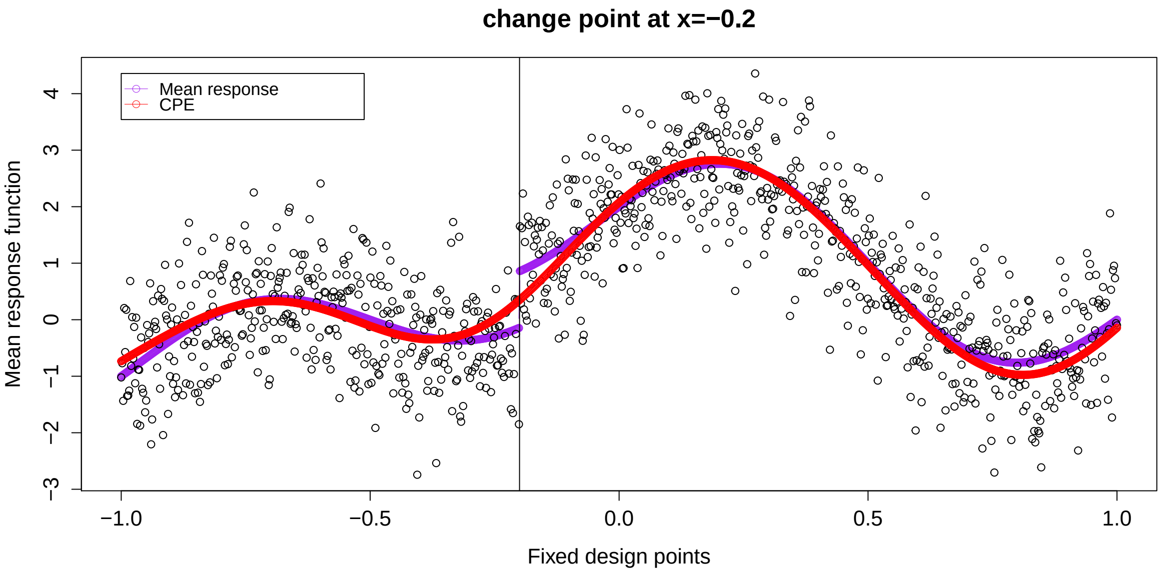


Figure 4.6: CPE for derivative from the bandwidth chosen from Cp

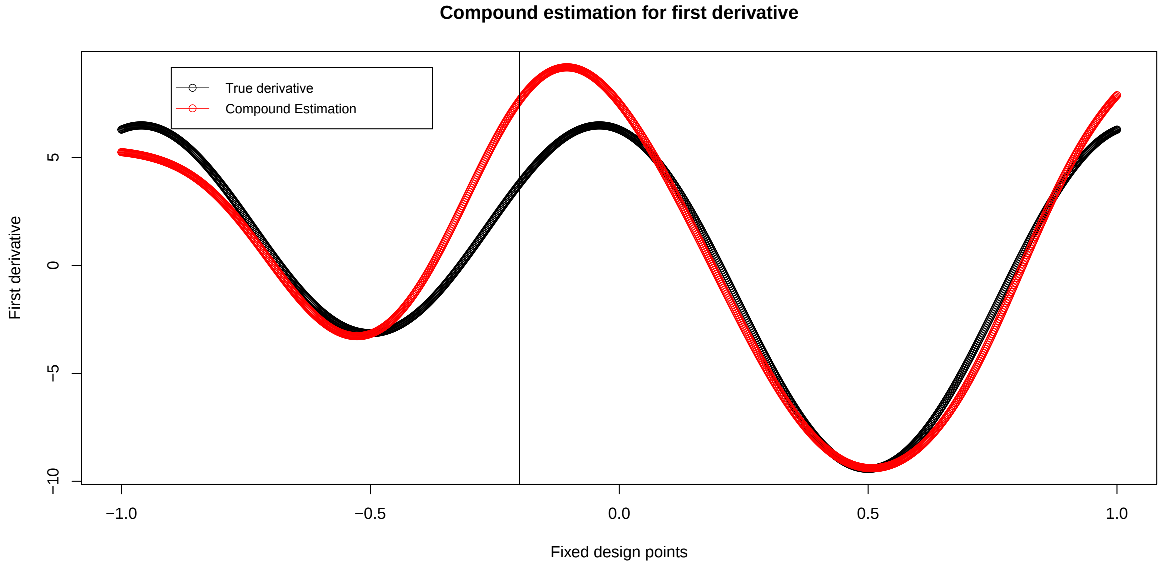
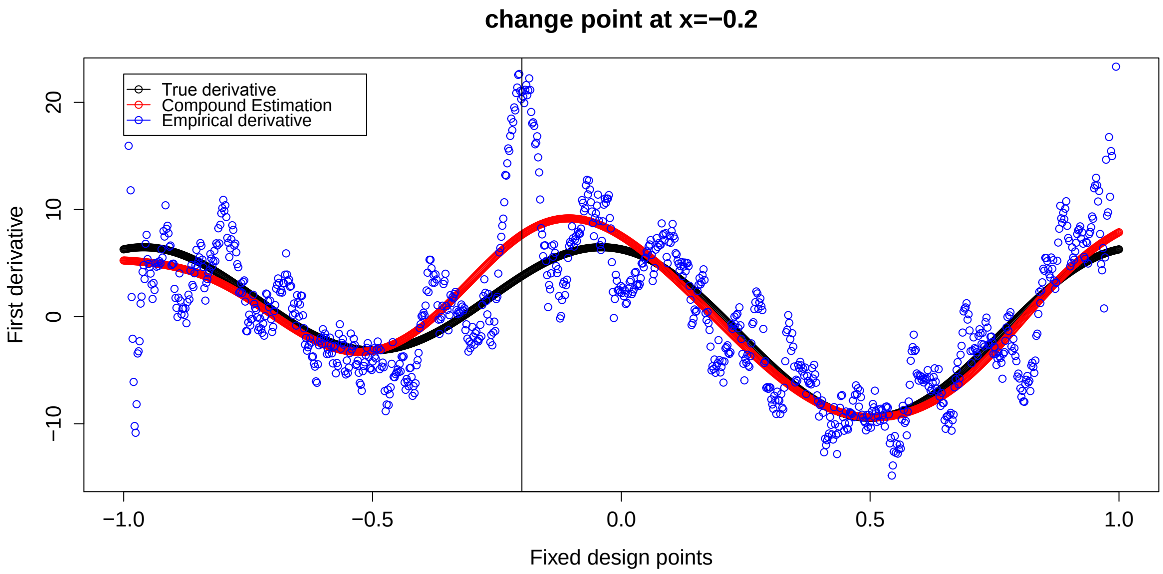


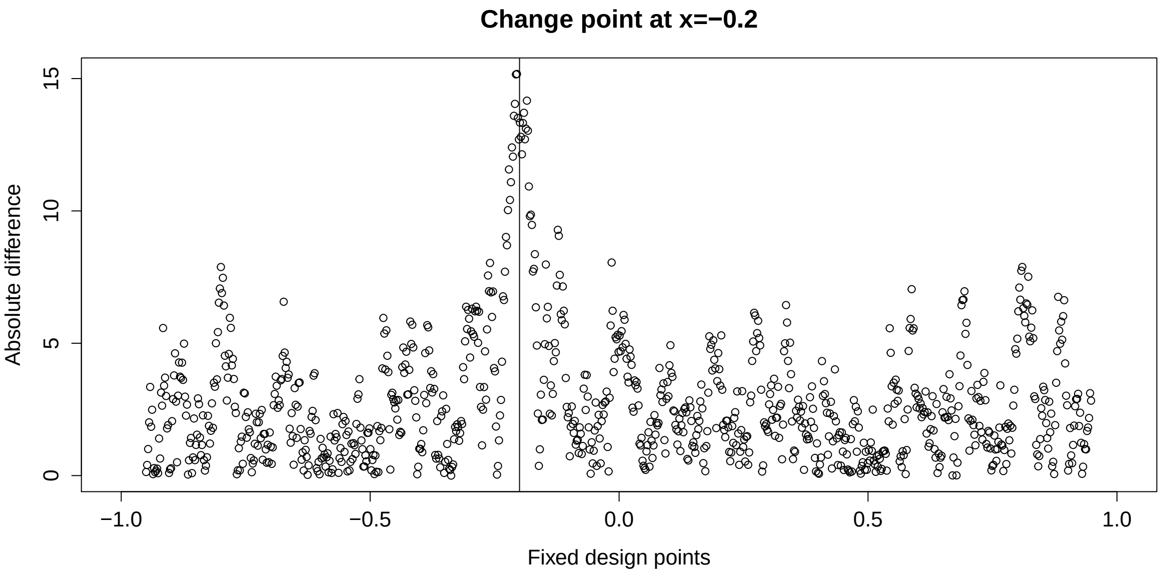
Figure 4.7: True Derivative, CPE, and Empirical derivative



candidate values equally spaced on $[0.4, 0.6]$. Once we fixed k, location of where the maximum of the difference between "naive" compound estimator and the Empirical first derivative could be detected. Then the interval estimator will be as in Theorem 4.4.1. Figure 4.7 shows that the Empirical first derivative will present an abnormally high spike when there is a jump point, which is expected from the deduction in section 4.3. Figure 4.8 plots the x_i vs $|Y_i^{(1)} - \mu^*(x_i)|$, it is very clear that there should be a jump point x_0 around -0.2.

The procedure of the simulation described above could be exactly the guide when doing the real data analysis. We actually simulated the data 20 times, each time we record different bandwidth h, α, interval estimate etc. The results are displayed as in Table 4.1. Table 4.1 shows that our method works pretty well, the interval estimate contains the jump point $x_0 = -0.2$ for all 20 times of simulations. The IndMax stands for the index of x_i that have the largest gap between Empirical derivative and Compound estimator and LocMax means the location of the highest jump. From the table, we can see the location of the maximum gap is always around -0.2. The

Figure 4.8: Distance between CPE and Empirical derivative



bandwidths h we chosen at each time are close, which indicates that C_p is stable and works well for picking h. The best α value seems to be between 0.45 and 0.55, which intuitively makes sense. First, we would like to have a large α such that the variance of Empirical first derivative will increase as slowly as possible compared to the gap when x is in the neighborhood of x_0. However, the proof of Theorem 4.4.1 tells us that there should be a upper bound for α value, which is $\frac{2}{3}(1 - 4\delta - \eta)$.

In the second scenario, we let $f(x)$ be a continuous function on $[-1, 1]$. The data was simulated from model (4.1) with $\mu(x) = sin(2\pi x) + cos(\pi x)I(x \geq 0.2)$ and error term $\epsilon_i \sim N(0, 0.4^2)$. We generate 1000 data points 20 times, Figure 4.9 is the scatter plot from one of the generated data sets, and the mean response was added to the scatter plot in Figure 4.10. It is clear there is a big jump at $x_0 = 0.2$ from Figure 4.10. However, the noise ϵ's blurred the jump in Figure 4.9. The same as in first scenario, we pick bandwidth h from the Cp criteria and let $\beta_n = 30$ for the "naive" compound estimation. The 40 griding points a's are equally spaced on the interval

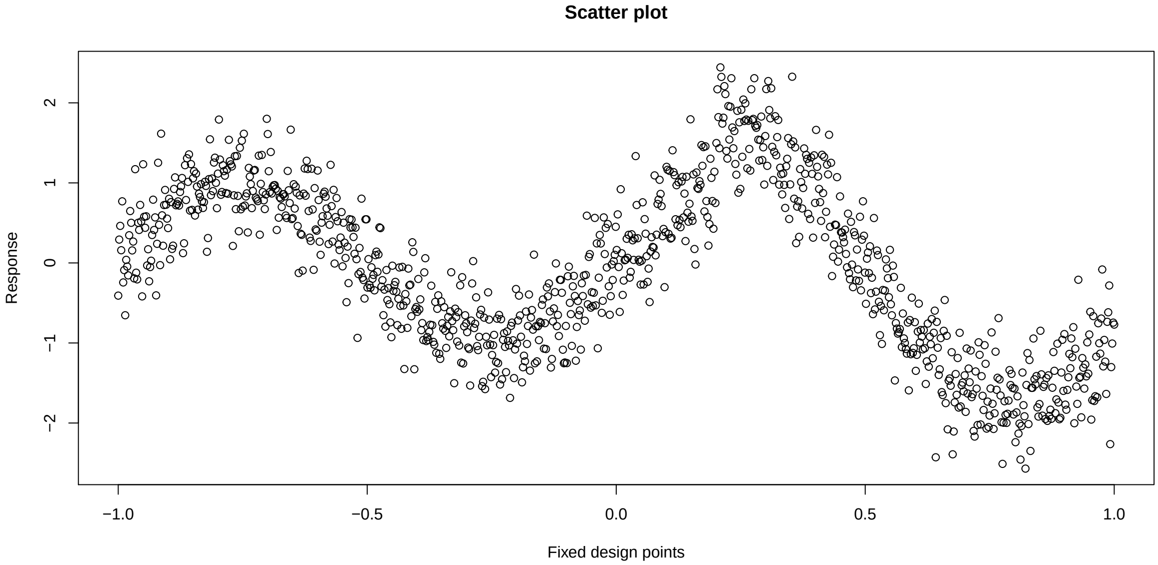
Table 4.1: Results from 20 trials.

Trials	h	α	k	IndMax	LocMax	Interval estimate	percent
1	0.378	0.512	34	413	-0.175	(-0.243, -0.107)	6.81%
2	0.364	0.524	37	407	-0.187	(-0.261, -0.113)	7.41%
3	0.351	0.473	26	398	-0.205	(-0.257, -0.153)	5.21%
4	0.365	0.427	19	398	-0.205	(-0.243, -0.167)	3.80%
5	0.343	0.512	34	408	-0.185	(-0.253, -0.117)	6.81%
6	0.363	0.467	25	405	-0.191	(-0.241, -0.141)	5.01%
7	0.363	0.493	30	414	-0.173	(-0.233, -0.113)	6.01%
8	0.347	0.400	15	406	-0.189	(-0.219, -0.159)	3.00%
9	0.368	0.511	34	403	-0.195	(-0.263, -0.127)	6.81%
10	0.371	0.402	16	398	-0.205	(-0.237, -0.173)	3.20%
11	0.358	0.564	49	399	-0.203	(-0.301, -0.105)	9.81%
12	0.355	0.467	25	411	-0.179	(-0.229, -0.129)	5.01%
13	0.371	0.412	17	395	-0.211	(-0.245, -0.177)	3.40%
14	0.361	0.412	17	394	-0.213	(-0.247, -0.179)	3.40%
15	0.384	0.461	24	406	-0.189	(-0.237, -0.141)	4.81%
16	0.347	0.515	35	404	-0.193	(-0.263, -0.123)	7.01%
17	0.354	0.479	27	389	-0.223	(-0.277, -0.169)	5.41%
18	0.354	0.588	58	387	-0.227	(-0.343, -0.111)	11.6%
19	0.370	0.519	36	404	-0.193	(-0.265, -0.121)	7.21%
20	0.355	0.473	26	407	-0.187	(-0.239, -0.135)	5.21%

$[-0.98, 0.98]$. The compound estimates for the mean response and first derivative are shown as Figure 4.11 and Figure 4.12. Figure 4.12 shows us the derivative is also discontinuous at the jump point $x_0 = 0.2$. However, the compound estimator still behaves very well when x is not in the neighborhood of change point. Also, we let $k = n^\alpha$ and pick α from $[0.4, 0.6]$ such that quantity (4.64) will be maximized. Visualization of the relations among mean response, compound estimate and the Empirical first derivative is in Figure 4.13. Obviously, the Empirical first derivative will have a larger bias when it is close to the change point $x_0 = 0.2$. Figure 4.14 displays the absolute difference between compound estimation and the Empirical first derivative, it gives us a very strong signal about where the change point is.

Table 4.2 presents the simulation results from 20 replications. For each replication, we proceed with the same methods of choosing h and α as before, and then obtain the interval estimate from that h and α. Again, the change point $x_0 = 0.2$ was contained

Figure 4.9: The observed data points



in the interval estimate for all 20 replications. The index of x_i that has the largest gap is always around 600, which is the true index of the jump point. Therefore, the method works well in this case.

Figure 4.10: The mean response function

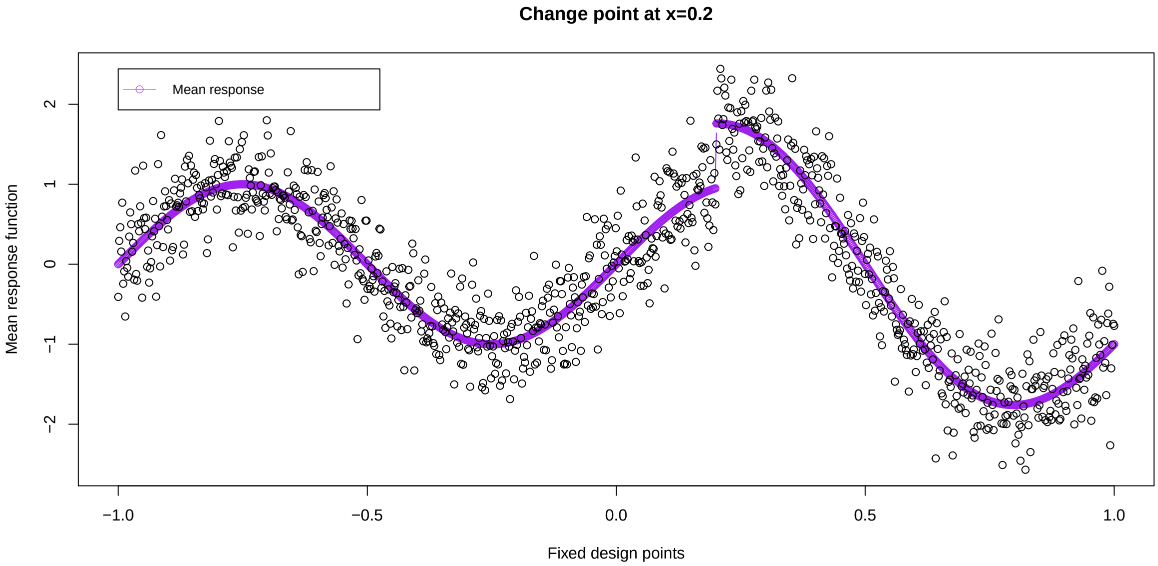


Figure 4.11: CPE from the bandwidth chosen from Cp

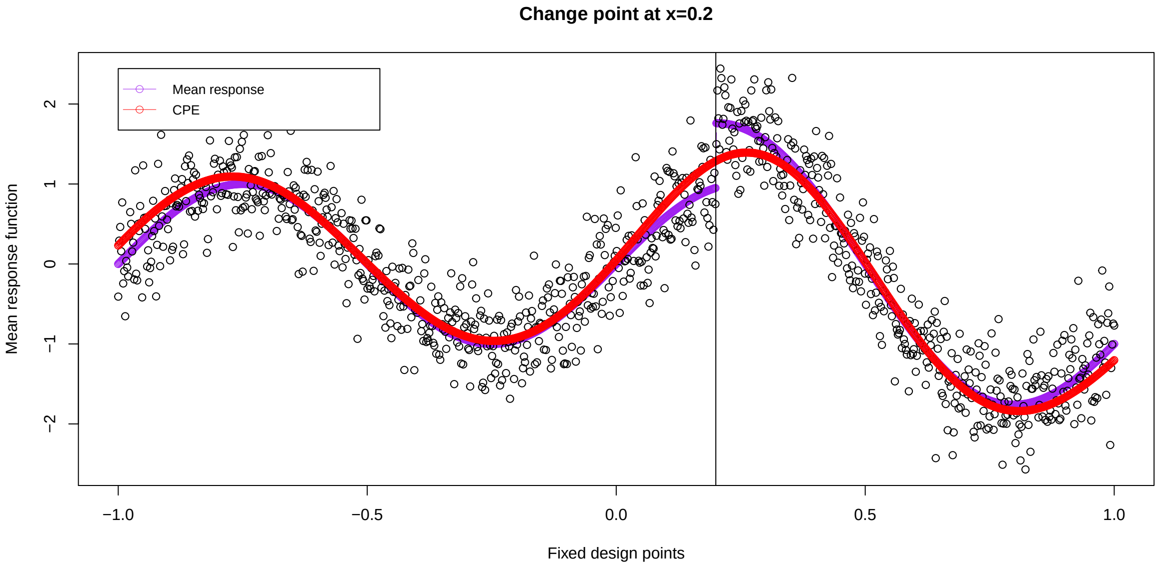


Figure 4.12: CPE from the bandwidth chosen from Cp

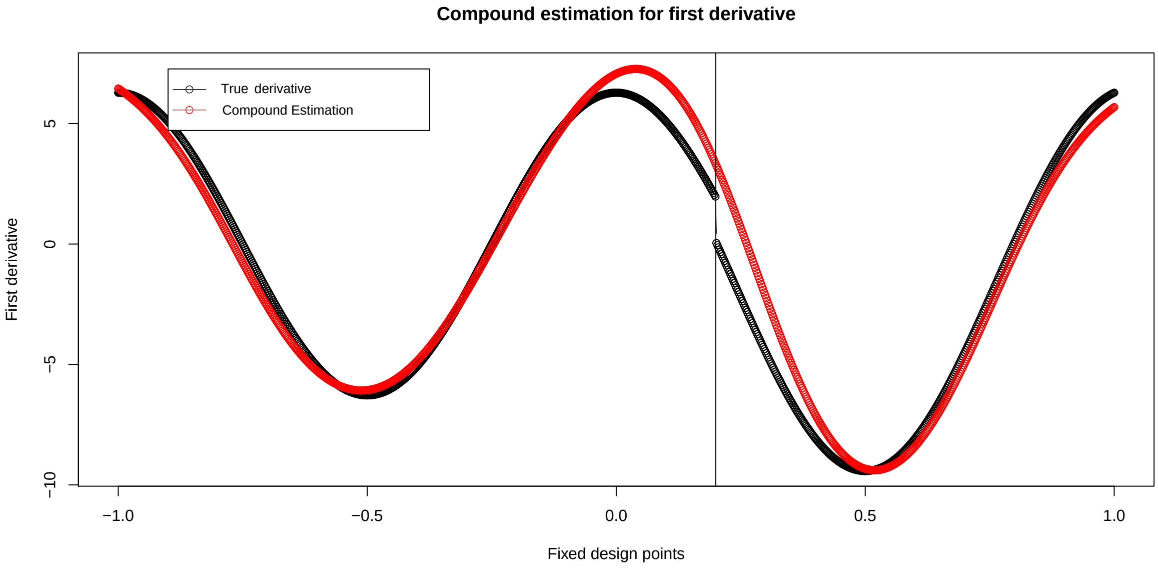


Figure 4.13: True Derivative, CPE, and Empirical derivative

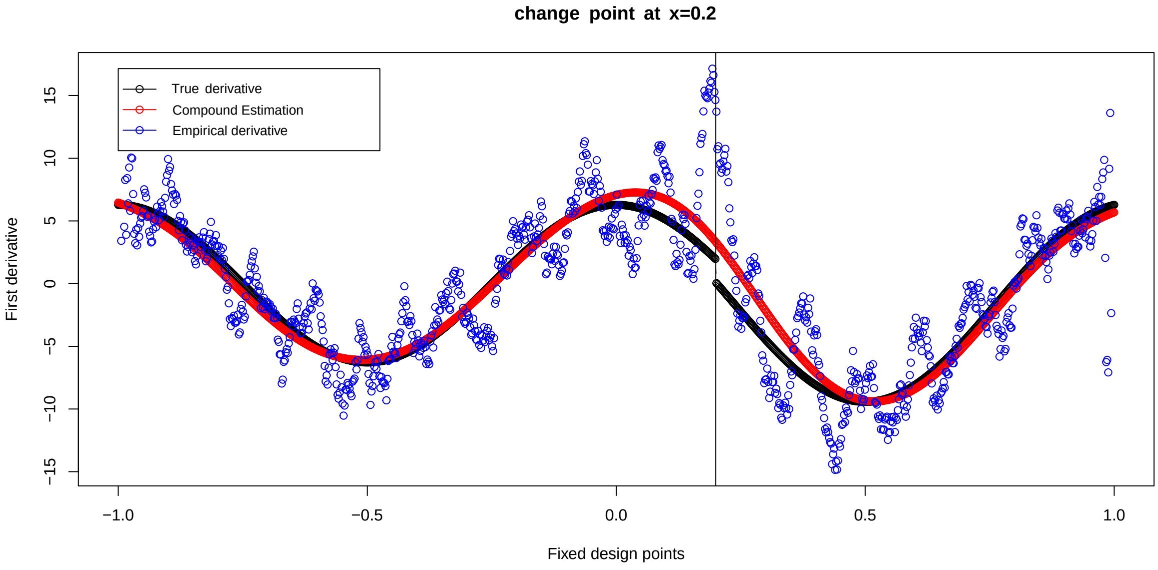


Figure 4.14: Distance between CPE and Empirical derivative

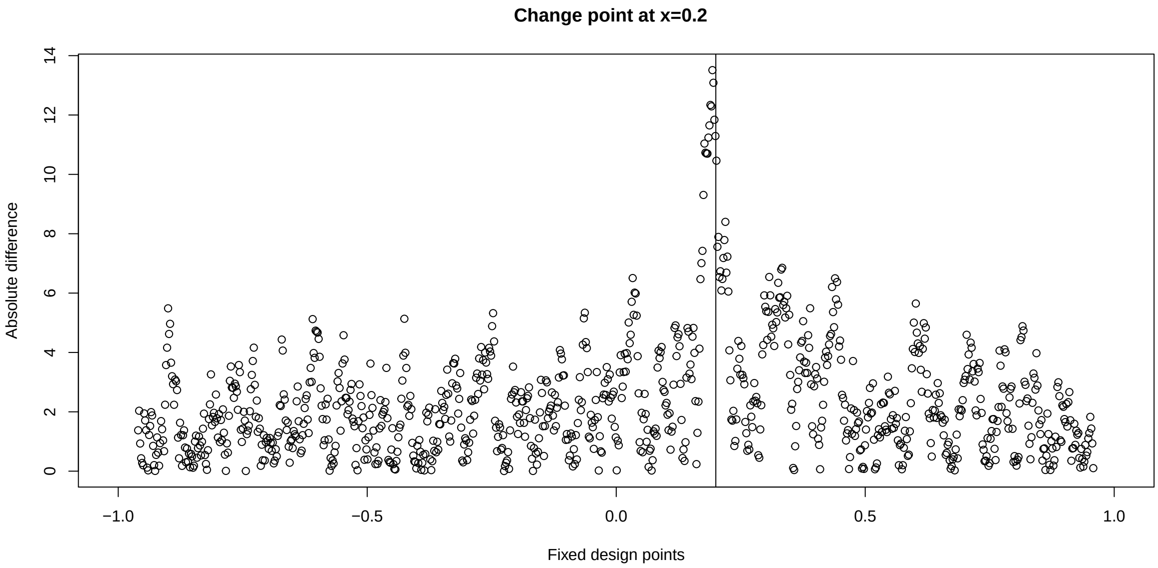


Table 4.2: Results from 20 trials.

Trials	h	α	k	IndMax	LocMax	Interval estimator	percent
1	0.354	0.503	32	595	0.189	(0.125, 0.253)	6.41%
2	0.350	0.442	21	597	0.193	(0.151, 0.235)	4.20%
3	0.351	0.503	32	591	0.181	(0.117, 0.245)	6.41%
4	0.347	0.427	19	604	0.207	(0.169, 0.245)	3.80%
5	0.358	0.564	49	600	0.199	(0.101, 0.297)	9.81%
6	0.354	0.448	22	596	0.191	(0.147, 0.235)	4.40%
7	0.355	0.479	27	608	0.215	(0.161, 0.269)	5.41%
8	0.360	0.403	16	603	0.205	(0.173, 0.237)	3.20%
9	0.357	0.494	30	602	0.203	(0.143, 0.263)	6.01%
10	0.330	0.488	29	607	0.213	(0.155, 0.271)	5.81%
11	0.354	0.485	28	607	0.213	(0.157, 0.269)	5.61%
12	0.356	0.427	19	596	0.191	(0.153, 0.229)	3.80%
13	0.342	0.503	32	600	0.199	(0.135, 0.263)	6.41%
14	0.355	0.427	19	595	0.189	(0.151, 0.227)	3.80%
15	0.340	0.420	18	602	0.203	(0.167, 0.239)	3.60%
16	0.355	0.467	25	606	0.211	(0.161, 0.261)	5.01%
17	0.357	0.521	36	603	0.205	(0.133, 0.277)	7.21%
18	0.345	0.494	30	596	0.191	(0.131, 0.251)	6.01%
19	0.344	0.524	37	601	0.201	(0.127, 0.275)	7.41%
20	0.345	0.442	21	602	0.203	(0.161, 0.245)	4.20%

Chapter 5 Jump Point Detection with random design points and heteroskedasticity

5.1 Motivation

In chapter 4, the jump point detection procedure is followed from model (4.1), which assumes fixed design points and equal variance of error terms. However, in real case, the x's may not be equally spaced and the data may present heteroskedasticity. For example, some stock markets are closed on Sunday and some national holidays, if we would like to analyze them, the time span is not uniform. In this scenario, the x's is fixed but not equally spaced. Another example is the recording of earthquakes in a seismic zone. If we want to model the relation between time and the strength of earthquakes, timing of earthquakes will be random. Then the x's will be realizations of random points in this case. Therefore, we would like to extend the results in chapter 4 to the case with random designed X's and heteroskedasticity. The idea of detecting jump points will be the same as in chapter 4, there are only slight modifications for the model and methodology. Figure 5.1 and Figure 5.2 show us an example of scatter plot and mean response when X's are random and errors have non constant variance as model (5.1). We generated X's from truncated normal distribution with mean 0 and standard deviation 1. The support set is $(-1, 1)$. The mean response function is $\mu(x) = \sin(2\pi x)I(x \leq 0.5) + [\sin(2\pi x) + x^3 + 1]I(x > 0.5)$. Also, the error ϵ_i are generated from distribution $N(0, \left[\frac{1}{(|X_i|+1)^2}\right]^2)$.

Suppose $\mu(x)$ is the mean response of the nonparametric regression model.

$$Y_i = \mu(X_i) + \epsilon_i, \tag{5.1}$$

where $\mu(x) = h(x) + I(x > x_0)f(x)$, and $h(x)$, $f(x)$ are continuous functions as in chapter 4 with at least $(J + 1)$ derivatives and $f(x)$ is bounded away from 0 when x

Figure 5.1: Scatter plot from observations

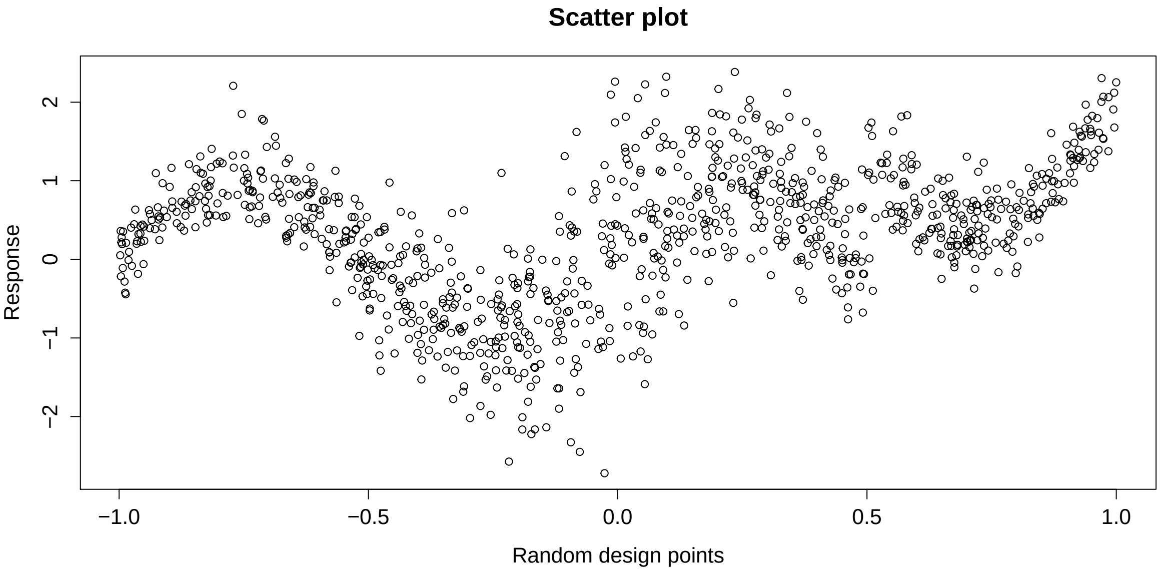
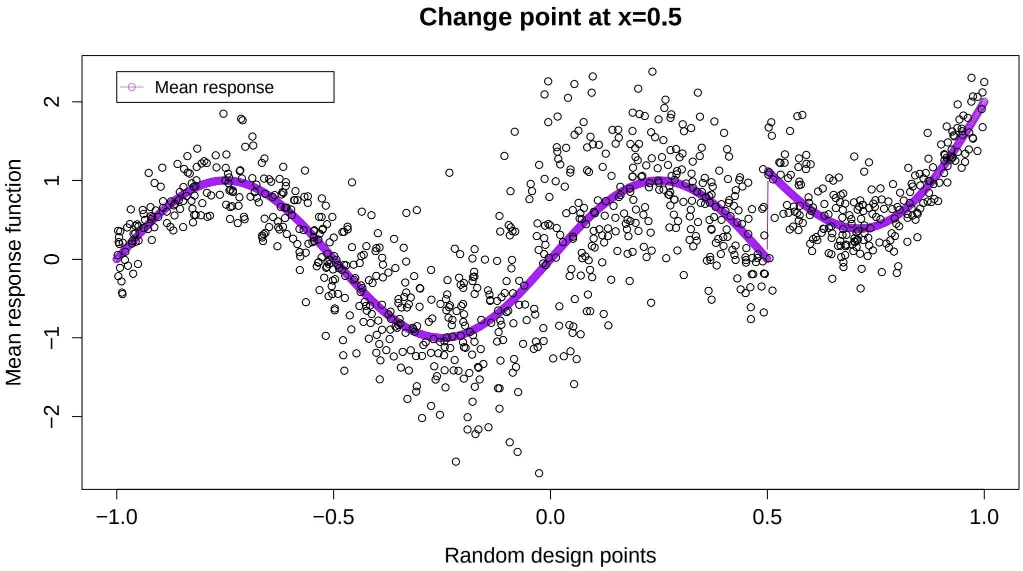


Figure 5.2: Mean response function



is close to x_0, and for $i \in \{1, 2, ..., n\}$. The random design points X_i's are identically distributed according to a continuous distribution F, which is twice differentiable, and X_i's belong to a compact interval $\mathcal{X} \subset \mathbb{R}$. The error terms ϵ_i are independent zero-mean random errors with variance σ_i^2. With out loss of generality, we assume $\mathcal{X} = [-1, 1]$. Also, the variance of error terms are finite, bounded away from 0, and bounded above. That is there are some positive constants C_1 and C_2 such that,

$$C_1 < \min(\sigma_i^2) < \max(\sigma_i^2) < C_2. \tag{5.2}$$

5.2 Properties of "naive" compound estimation with heteroskedasticity and random design points

Chapter 2 has already showed us the properties of compound estimation of a smooth function with random design points and heteroskedasticity. Chapter 4 made some assumption for "naive" compound estimation with fixed designs. However, most of these assumption could be the same and part of the conclusions during the proof will follow from these assumptions. Therefore, some of the conclusion from chapter 2 and chapter 4 could be applied immediately. In this section, we'll briefly describe the proof of the following theorem instead of reinventing the wheel. We will use the same notation of three mean response functions as in chapter 4. The only difference is that the X's will be random design points, which are not equally spaced over the interval $[-1, 1]$. Thus, the assumptions (4.5) to (4.14) will still hold. However, the assumptions (4.2) to (4.4) need to be changed on account of heteroskedasticity. In chapter 2, we proved that the compound estimation still achieves near optimal convergence rate with adaptive Gaussian convolution weights. Therefore, we can assume

$$h^{\star}(x) = \sum_{a} W_{a,n}(x-a) \sum_{j=0}^{J} \widehat{h^{(j)}(a)}(x-a)^j \tag{5.3}$$

$$g^{\star}(x) = \sum_{a} W_{a,n}(x-a) \sum_{j=0}^{J} \widehat{g^{(j)}(a)}(x-a)^j \tag{5.4}$$

$$\mu^{\star}(x) = \sum_{a} W_{a,n}(x-a) \sum_{j=0}^{J} \widetilde{\mu^{(j)}(a)}(x-a)^j, \tag{5.5}$$

instead of (4.2)-(4.4). The weight function $W_{a,n}(x-a)$ depends on both x and grid point a. Also, the pointwise estimators $\widehat{h^{(j)}(a)}$, $\widehat{g^{(j)}(a)}$, and $\widetilde{\mu^{(j)}(a)}$ will be adapted for heteroskedasticity by some nonparametric regression method in the first step of compound estimation.

Theorem 5.2.1 *Suppose we have a nonparametric regression model (5.1), and the assumptions (5.2)-(5.5) and (4.5)-(4.14) hold, let I_n be the set of all the grid points a's, $\delta = \frac{v}{4J+6}$, $r = \frac{1}{2J+3}$, $r_1 > 0$, then the naive compound estimator $\mu'(x)$ in (5.5) satisfy*

$$|\frac{d}{dx}\mu^{\star}(x) - \frac{d}{dx}\mu(x)| = O(n^{4\delta+\frac{1}{2J+3}}) + O_p(n^{4\delta+\frac{2}{2J+3}-r_1}) \tag{5.6}$$

Proof:

The proof of theorem 5.2.1 will be very similar to the proof of theorem 4.2.1. The only difference is that we need to find the convergence rate of the adaptive weights $W_{a,n}(x-a)$, which could be easily obtained from (2.11) and (2.14). Following from the notation of chapter 4, we let $L_n = \Theta(n^{r+\delta+\phi})$ and $I_{1n}(x) = \{a \in I_n : |a-x| < n^{-r}\}$. Also, the upper bound of Gaussian convolution weights is

$$\beta_n^{\star} = \beta_{0n}^{\star} n^{2(\gamma+\delta)} = \Theta(n^{2(r+\delta)}) \tag{5.7}$$

then (2.11), (2.14) and (5.7) immediately yields

$$\begin{aligned}\sup_{a\in \overline{I}_{1n}(x)}\left|\frac{d^k}{dx^k}W_{a,n}(x)\right| &\leq \frac{L_n^k C_2 exp^{[-\min\{\beta_{0n}(a)\}n^{2\delta}]}(\beta_n^\star)^k}{[L_n(\beta_n^\star)^{-1/2}]^{k+1}} \\ &= \frac{C_2(\beta_n^\star)^{(3k+1)/2}exp^{[-\min\{\beta_{0n}(a)\}n^{2\delta}]}}{L_n} \\ &= \Theta\left(n^{(3kr-\phi+3k\delta)}exp^{[-\min\{\beta_{0n}(a)\}n^{2\delta}]}\right), \qquad (5.8)\end{aligned}$$

and

$$\begin{aligned}\sup_{a\in I_{1n}(x)}\left|\frac{d^k}{dx^k}W_{a,n}(x)\right| &= O\left(\frac{n^{(3k+1)(\gamma+\delta)-2k\gamma}}{L_n}\right) \\ &= O(n^{3k\delta-\phi+kr}). \qquad (5.9)\end{aligned}$$

As (4.18), formula (5.8) will decay in exponential rate, and the numbers of $a's$ in I_n is $\Theta(n^{r+\delta+\phi})$. Therefore, summand of $|\frac{d^k}{dx^k}W_n(x-a)|$ when $a \in \overline{I}_{1n}(x)$ could be negligible. Equation (5.9) is identical to (4.17) as in the proof of theorem 4.2.1. The rest of proof will be exactly the same as proof of theorem 4.2.1. ■

Let $\max\{r,(2r-r_1)\} = \eta$, we'll have

$$\mu^\star(x)' - \mu'(x) = O_p(n^{4\delta+\eta}). \qquad (5.10)$$

If x is away from the jump point x_0, then by (2.16) in chapter 2, we'll have

$$\left|\widetilde{\mu^{(j)}(x)} - \mu^{(j)}(x)\right| = o_p(n^{-\frac{J-j+1}{2J+3}+v}), \qquad (5.11)$$

for $x \subset I\backslash[x_0-\tau_n, x_0+\tau_n]$ and τ_n is a sequence of positive numbers such that the estimator of $\mu^{(j)}(x)$ when x is outside of $[x_0-\tau_n, x_0+\tau_n]$ will not be affected by the change point x_0. For example, if we apply the Nearest Neighbor regression, τ_n will be decided by the number of nearest design points of x_0, or if we use local regression, τ_n is dependent on the bandwidth h.

5.3 Properties of Empirical first derivatives with random designed points and heteroskedasticity

Suppose the design points X's are random, from chapter 3, the Empirical first derivative is defined as

$$Y_i^{(1)} = \sum_{j=1}^{k} w_{ij} \frac{Y_{i+j} - Y_{i-j}}{X_{i+j} - X_{i-j}}. \tag{5.12}$$

The X's are sorted from smallest to the largest. And the weights are

$$w_{ij} = \frac{\eta_{ij}}{\sum_{j=1}^{k} \eta_{ij}} \tag{5.13}$$

for $k+1 \leq i \leq n-k$, where $\eta_{ij} = \frac{(X_{i+j}-X_{i-j})^2}{\sigma_{i+j}^2+\sigma_{i-j}^2}$. Then

$$\frac{w_{ij}}{X_{i+j} - X_{i-j}} = \frac{X_{i+j} - X_{i-j}}{\sigma_{i+j}^2 + \sigma_{i-j}^2} \Bigg/ \sum_{j=1}^{k} \frac{(X_{i+j} - X_{i-j})^2}{\sigma_{i+j}^2 + \sigma_{i-j}^2}. \tag{5.14}$$

By (5.2)

$$\frac{C_1}{C_2} \frac{X_{i+j} - X_{i-j}}{\sum_{j=1}^{k} (X_{i+j} - X_{i-j})^2} \leq \frac{w_{ij}}{X_{i+j} - X_{i-j}} \leq \frac{C_2}{C_1} \frac{X_{i+j} - X_{i-j}}{\sum_{j=1}^{k} (X_{i+j} - X_{i-j})^2} \tag{5.15}$$

Assume jump point $x_0 \in (X_s, X_{s+1})$ and $X_i \leq X_s$, then

$$\begin{aligned}
Y_i^{(1)} =& \sum_{j=1}^{k} w_{ij} \frac{Y_{i+j} - Y_{i-j}}{X_{i+j} - X_{i-j}} \\
=& \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[h(X_{i+j}) - h(X_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} + f(X_{i+j}) I[X_0 \in (X_i, X_{i+j})] \right] \\
=& \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[h(X_{i+j}) - h(X_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} \right] + \sum_{j=1}^{k} \frac{w_j f(X_{i+j})}{X_{i+j} - X_{i-j}} I[x_0 \in (X_i, X_{i+j})] \\
=& \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[h(X_{i+j}) - h(X_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} \right] \\
&+ \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i+j}) I[i \leq s < s+1 \leq i+j] \\
=& \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[h(X_{i+j}) - h(X_{i-j}) \right] + \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[\epsilon(X_{i+j}) - \epsilon(X_{i-j}) \right] \\
&+ \sum_{j=s+1-i}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i+j}) I[i \leq s < s+1 \leq i+j]. \qquad (5.16)
\end{aligned}$$

Since $X_i < X_s$, we have $\mu'(X_i) = h'(X_i)$, then from the equation above,

$$\begin{aligned}
Y_i^{(1)} - \mu'(X_i) &= Y_i^{(1)} - h'(X_i) \\
&= E(Y_i^{(1)} | \boldsymbol{X}) - h'(X_i) + Y_i^{(1)} - E(Y_i^{(1)} | \boldsymbol{X}) \\
&= \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[h(X_{i+j}) - h(X_{i-j}) \right] - h'(X_i) \\
&+ \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[\epsilon(X_{i+j}) - \epsilon(X_{i-j}) \right] \\
&+ \sum_{j=s+1-i}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i+j}) \qquad (5.17)
\end{aligned}$$

Following from the same assumption of k as in Chapter 3, we let $k = \Theta(n^{\alpha})$. Suppose $\frac{1}{4} < \alpha < \frac{1}{2}$, then from (3.38),

$$\sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[h(X_{i+j}) - h(X_{i-j})\right] - h'(X_i) = O_p(n^{\alpha-1}) \tag{5.18}$$

From (3.37), we adopt the notations from Chapter 3 that $p = \frac{i_n}{n}$, $q_n = i_n + j_n$, $q'_n = i_n - j_n$. Also, let the density function be denoted by f_1, then

$$\begin{aligned} X_{i_n+j_n} - X_{i_n-j_n} &= \frac{q_n - q'_n}{n f_1(\zeta_p)} + \tilde{R}_n \\ &= \frac{2j_n}{n}\frac{1}{f_1(\zeta_p)} + O_p(n^{-3/4}(\log n)^{1/2(\delta+1)}) \end{aligned}$$

and $j_n \in \{1, 2, 3 ... k_n\}$.

Since $k_n = \Theta(n^{\alpha})$ and $\frac{1}{4} < \alpha < \frac{1}{2}$, then there are $\Theta(n^{\alpha})$ of j_n such that $X_{i_n+j_n} - X_{i_n-j_n} = \Theta_p(n^{\alpha-1})$ and others will be $X_{i_n+j_n} - X_{i_n-j_n} < \Theta_p(n^{\alpha-1})$, number of j_n satisfy later condition will be $O(n^{\alpha})$. These lead to the following results:

$$\begin{aligned} \sum_{j=1}^{k}(X_{i+j} - X_{i-j}) &= \Theta(n^{\alpha})\Theta_p(n^{\alpha-1}) + O(n^{\alpha})O_p(n^{\alpha-1}) \\ &= \Theta_p(n^{2\alpha-1}), \end{aligned} \tag{5.19}$$

and

$$\begin{aligned} \sum_{j=1}^{k}(X_{i+j} - X_{i-j})^2 &= \Theta(n^{\alpha})\Theta_p(n^{2\alpha-2}) + O(n^{\alpha})O_p(n^{2\alpha-2}) \\ &= \Theta_p(n^{3\alpha-2}) \end{aligned} \tag{5.20}$$

Then by (3.6), (5.15) and (5.20),

$$\begin{aligned}
Var(Y_i^{(1)}|\boldsymbol{X}) &= \sum_{j=1}^{k} \frac{\sigma_{i+j}^2 + \sigma_{i-j}^2}{(X_{i+j} - X_{i-j})^2} w_{ij}^2 \\
&\leq 2C_2 \sum_{j=1}^{k} \left[\frac{w_{ij}}{(X_{i+j} - X_{i-j})} \right]^2 \\
&\leq 2C_2 \sum_{j=1}^{k} \left\{ \frac{(X_{i+j} - X_{i-j})^2}{4C_1^2} \Bigg/ \left(\sum_{j=1}^{k} \frac{(X_{i+j} - X_{i-j})^2}{2C_2} \right)^2 \right\} \\
&= \frac{8C_2^3}{4C_1^2} \frac{\sum_{j=1}^{k} (X_{i+j} - X_{i-j})^2}{\left[\sum_{j=1}^{k} (X_{i+j} - X_{i-j})^2 \right]^2} \\
&= \frac{2C_2^3}{C_1^2} \frac{1}{\sum_{j=1}^{k} (X_{i+j} - X_{i-j})^2} \\
&= \Theta_p(n^{2-3\alpha}) \qquad (5.21)
\end{aligned}$$

From (5.21), we can infer the rate of $\sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[\epsilon(X_{i+j}) - \epsilon(X_{i-j}) \right]$ in formula (5.17). We let

$$W_n = \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[\epsilon(X_{i+j}) - \epsilon(X_{i-j}) \right], \qquad (5.22)$$

then

$$Var(W_n|\boldsymbol{X}) = \sum_{j=1}^{k} \frac{\sigma_{i+j}^2 + \sigma_{i-j}^2}{(X_{i+j} - X_{i-j})^2} w_{ij}^2 \qquad (5.23)$$

where

$$w_{ij} = \frac{\eta_{ij}}{\sum_{j=1}^{k} \eta_{ij}} \qquad \text{and} \qquad \eta_{ij} = \frac{(X_{i+j} - X_{i-j})^2}{\sigma_{i+j}^2 + \sigma_{i-j}^2} \qquad (5.24)$$

Therefore, we will have

$$\begin{aligned}
Var(W_n|\boldsymbol{X}) &= \sum_{j=1}^{k} \frac{\sigma_{i+j}^2 + \sigma_{i-j}^2}{(X_{i+j} - X_{i-j})^2} \frac{\eta_{ij}^2}{(\sum_{j=1}^{k} \eta_{ij})^2} \\
&= \frac{1}{\sum_{j=1}^{k} \eta_{ij}}. \qquad (5.25)
\end{aligned}$$

From (5.2), we'll have

$$Var(W_n|\boldsymbol{X}) = \frac{1}{\sum_{j=1}^{k} \frac{(X_{i+j}-X_{i-j})^2}{\sigma_{i+j}^2+\sigma_{i-j}^2}} \leq C\frac{1}{(X_{i+j}-X_{i-j})^2}. \tag{5.26}$$

Where C is a fixed constant. Thus,

$$\begin{aligned} E\left[Var(W_n|\boldsymbol{X})\right] &\leq CE\left[\frac{1}{(X_{i+j}-X_{i-j})^2}\right] \\ &\leq CE\left[\frac{1}{\frac{k}{2}(X_{i+k/2}-X_{i-k/2})^2}\right] \\ &\leq \frac{2C}{k}E\left[\frac{1}{(X_{i+t}-X_{i-t})^2}\right], \end{aligned} \tag{5.27}$$

where $t = \left[\frac{k}{2}\right] = \Theta(n^{\alpha})$. Now we look at the rate of $E\left[\frac{1}{(X_{i+t}-X_{i-t})^2}\right]$. Suppose the CDF of each random variable X is F, then by the formula of joint distribution of two order statistics, we could have

$$\begin{aligned} &E\left[\frac{1}{(X_{i+t}-X_{i-t})^2}\right] \\ &= \iint_D \frac{n!(u-v)^{-2}}{(i-t-1)!(2t-1)!(n-i-t)!}F(u)^{i-t-1}[F(v)-F(u)]^{2t-1}[1-F(v)^{n-i-t}]dF(u)dF(v) \\ &= \iint_D \frac{n!F(v)^{i-t-1}[1-F(v)]^{n-i-t}}{(i-t-1)!(2t-1)!(n-i-t)!}\frac{[F(v)-F(u)]^{2t-1}}{(u-v)^2}dF(u)dF(v) \\ &\leq C_3\frac{n!}{(i-t-1)!(2t-1)!(n-i-t)!}\iint_D F(u)^{i-t-1}[F(v)-F(u)]^{2t-3}[1-F(v)^{n-i-t}]dF(u)dF(v) \\ &= C_3\frac{n!}{(i-t-1)!(2t-1)!(n-i-t)!}\frac{(i-t-1)!(2t-3)!(n-i-t)!}{(n-2)!} \\ &= C_3\frac{n(n+1)}{(2t-1)(2t-2)} \\ &= \Theta(n^{2-2\alpha}) \end{aligned} \tag{5.28}$$

Combine (5.27) and (5.28), we get $E[Var(W_n|\boldsymbol{X})] \leq \Theta(n^{2-3\alpha})$, which indicate that

$$\begin{aligned} Var(W_n) &= E[Var(W_n|\boldsymbol{X})] + Var[E(W_n|\boldsymbol{X})] \\ &\leq \Theta(n^{2-3\alpha}). \end{aligned} \tag{5.29}$$

Therefore,

$$W_n = E(W_n) + O_p\left(Var(W_n)\right) \leq \Theta_p(n^{1-\frac{3}{2}\alpha}). \tag{5.30}$$

Without loss of generality, again, we can assume $f(x_0) > 2C > 0$ where C is a fixed constant. Since $k = \Theta(n^\alpha)$ implies $|X_{i+k} - X_{i-k}| = O_p(n^{\alpha-1}) \to 0$ in probability and $f(x)$ is smooth, when n is sufficient large, we'll have $f(X_{s+j}) > f(x_0)/2 > C$ for all $j \in \{1, 2, ...k\}$.

If $i = s$, by (5.2), (5.19) and (5.20), we will have

$$\begin{aligned}
\sum_{j=s+1-i}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i+j}) &= \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i+j}) \\
&\geq C \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \\
&\geq \sum_{j=1}^{k} \frac{C_1}{C_2} \frac{X_{i+j} - X_{i-j}}{\sum_{j=1}^{k}(X_{i+j} - X_{i-j})^2} C \\
&= \frac{C_1 C}{C_2} \frac{\sum_{j=1}^{k}(X_{i+j} - X_{i-j})}{\sum_{j=1}^{k}(X_{i+j} - X_{i-j})^2} \\
&= \frac{\Theta_p(n^{2\alpha-1})}{\Theta_p(n^{3\alpha-2})} \\
&= \Theta_p(n^{1-\alpha})
\end{aligned} \tag{5.31}$$

By (5.17), (5.18), (5.30) and (5.31), if $i = s$

$$\begin{aligned}
Y_i^{(1)} - \mu'(X_i) &= Y_i^{(1)} - h'(X_i) \\
&= O_p(n^{\alpha-1}) + \Theta_p(n^{1-\frac{3}{2}\alpha}) + \Theta_p(n^{1-\alpha}) \\
&= \Theta_p(n^{1-\alpha}).
\end{aligned} \tag{5.32}$$

If $i = s + 1 - k$, then by (5.15)

$$\begin{aligned}\sum_{j=s+1-i}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i+j}) &= \frac{w_{ik}}{X_{i+k} - X_{i-k}} f(X_{i+k}) \\ &\leq \frac{C_2}{C_1} \frac{X_{i+j} - X_{i-j}}{\sum_{j=1}^{k} (X_{i+j} - X_{i-j})^2} f(X_{i+k}) \quad = \frac{\Theta_p(n^{\alpha-1})}{\Theta_p(n^{3\alpha-2})} \\ &= \Theta_p(n^{1-2\alpha}) \end{aligned} \tag{5.33}$$

By (5.17), (5.18), (5.30) and (5.33) $\Rightarrow$

$$\begin{aligned} Y_i^{(1)} - \mu'(X_i) &= Y_i^{(1)} - h'(X_i) \\ &= O_p(n^{\alpha-1}) + \Theta_p(n^{1-\frac{3}{2}\alpha}) + \Theta_p(n^{1-2\alpha}) \\ &= \Theta_p(n^{1-\frac{3}{2}\alpha}) \end{aligned} \tag{5.34}$$

If $i \leq s - k$, then by (5.17), (5.18) and (5.21),

$$\begin{aligned} Y_i^{(1)} - \mu'(X_i) &= Y_i^{(1)} - h'(X_i) \\ &= O_p(n^{\alpha-1}) + \Theta_p(n^{1-\frac{3}{2}\alpha}) \\ &= \Theta_p(n^{1-\frac{3}{2}\alpha}) \end{aligned} \tag{5.35}$$

From (5.32), (5.34) and (5.35), we could see that the gap between Empirical first derivative will be larger when X's is in the neighborhood of the jump point x_0. This is very similar to the procedure in chapter 4. Now let's look at the properties of the Empirical first derivative when $i \geq s + 1$.

$$
\begin{aligned}
Y_i^{(1)} &= \sum_{j=1}^{k} w_{ij} \frac{Y_{i+j} - Y_{i-j}}{X_{i+j} - X_{i-j}} \\
&= \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[g(X_{i+j}) - g(X_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} + f(X_{i-j}) I[x_0 \in (X_{i-j}, X_i)] \right] \\
&= \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[g(X_{i+j}) - g(X_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} \right] + \sum_{j=1}^{k} \frac{w_{ij} f(X_{i-j})}{X_{i+j} - X_{i-j}} I[x_0 \in (X_{i-j}, X_i)] \\
&= \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[g(X_{i+j}) - g(X_{i-j}) + \epsilon_{i+j} - \epsilon_{i-j} \right] \\
&\quad + \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i-j}) I[i - j \leq s < s + 1 \leq i] \\
&= \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[g(X_{i+j}) - g(X_{i-j}) \right] + \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} [\epsilon_{i+j} - \epsilon_{i-j}] \\
&\quad + \sum_{j=i-s}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i-j}). \qquad (5.36)
\end{aligned}
$$

From the equation above,

$$
\begin{aligned}
Y_i^{(1)} - \mu'(X_i) &= Y_i^{(1)} - g'(X_i) \\
&= E(Y_i^{(1)} | \boldsymbol{X}) - g'(X_i) + Y_i^{(1)} - E(Y_i^{(1)} | \boldsymbol{X}) \\
&= \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[g(X_{i+j}) - g(X_{i-j}) \right] - g'(X_i) \\
&\quad + \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} [\epsilon_{i+j} - \epsilon_{i-j}] \\
&\quad + \sum_{j=i-s}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i-j}) \qquad (5.37)
\end{aligned}
$$

Since $k = \Theta(n^\alpha)$ and $\frac{1}{4} < \alpha < \frac{1}{2}$, then from (3.38),

$$
\sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \left[g(X_{i+j}) - g(X_{i-j}) \right] - g'(X_i) = O_p(n^{\alpha - 1}) \qquad (5.38)
$$

If $i = s + 1$, then similar to (5.33), we'll have

$$\begin{aligned}
\sum_{j=i-s}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i-j}) &= \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i-j}) \\
&\geq C \sum_{j=1}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} \\
&\geq \sum_{j=1}^{k} \frac{C_1}{C_2} \frac{X_{i+j} - X_{i-j}}{\sum_{j=1}^{k} (X_{i+j} - X_{i-j})^2} C \\
&= \frac{C_1 C}{C_2} \frac{\sum_{j=1}^{k} (X_{i+j} - X_{i-j})}{\sum_{j=1}^{k} (X_{i+j} - X_{i-j})^2} \\
&= \frac{\Theta_p(n^{2\alpha-1})}{\Theta_p(n^{3\alpha-2})} \\
&= \Theta_p(n^{1-\alpha}). \qquad (5.39)
\end{aligned}$$

By (5.37), (5.38), (5.30) and (5.39) ⇒

$$\begin{aligned}
Y_i^{(1)} - \mu'(X_i) &= Y_i^{(1)} - g'(X_i) \\
&= O_p(n^{\alpha-1}) + \Theta_p(n^{1-\frac{3}{2}\alpha}) + \Theta_p(n^{1-\alpha}) \\
&= \Theta_p(n^{1-\alpha}). \qquad (5.40)
\end{aligned}$$

If $i = s + k$, then by (5.15)

$$\begin{aligned}
\sum_{j=i-s}^{k} \frac{w_{ij}}{X_{i+j} - X_{i-j}} f(X_{i-j}) &= \frac{w_{ik}}{X_{i+k} - X_{i-k}} f(X_{i-k}) \\
&\leq \frac{C_2}{C_1} \frac{X_{i+k} - X_{i-k}}{\sum_{j=1}^{k} (X_{i+j} - X_{i-j})^2} f(X_{i-k}) \qquad = \frac{\Theta_p(n^{\alpha-1})}{\Theta_p(n^{3\alpha-2})} \\
&= \Theta_p(n^{1-2\alpha}). \qquad (5.41)
\end{aligned}$$

Therefore, combine (5.37), (5.38), (5.30), (5.41)

$$\begin{aligned} Y_i^{(1)} - \mu'(X_i) &= Y_i^{(1)} - g'(X_i) \\ &= O_p(n^{\alpha-1}) + \Theta_p(n^{1-\frac{3}{2}\alpha}) + \Theta_p(n^{1-2\alpha}) \\ &= \Theta_p(n^{1-\frac{3}{2}\alpha}) \end{aligned} \tag{5.42}$$

If $i \geq s + k + 1$, then combine (5.38), (5.39) and (5.41),

$$\begin{aligned} Y_i^{(1)} - \mu'(X_i) &= Y_i^{(1)} - g'(X_i) \\ &= O_p(n^{\alpha-1}) + \Theta_p(n^{1-\frac{3}{2}\alpha}) \\ &= \Theta_p(n^{1-\frac{3}{2}\alpha}). \end{aligned} \tag{5.43}$$

5.4 Detection of the location of jump point

In section 5.2 and 5.3, we delved into the properties of "naive" compound estimation and the Empirical first derivative when random design point X_i is in the neighborhood of x_0 or not. Then we detect the jump point through the distance between compound estimation and the Empirical first derivative when X_i is near x_0 or not. This section is a modification of section 4.4.

Theorem 5.4.1 *Suppose we have model (5.1), the change point x_0 is in the interval (X_{s_n}, X_{s_n+1}), and the order of Empirical first derivative is $k_n = \Theta(n^\alpha)$. Let t_n be the integer such that*

$$|Y_{t_n}^{(1)} - \mu^\star(X_{t_n})'| = \max_{i \in \{k+1,\ldots,n-k\}} \left\{ |Y_i^{(1)} - \mu^\star(X_i)'| \right\}, \tag{5.44}$$

and $I_n = (X_{t_n-k_n}, X_{t_n+k_n})$. Then there exist α such that $P(x_0 \in I_n) \longrightarrow 1$.

The proof of theorem 5.4.1 is identical to the proof of theorem 4.4.1. The α is similar to the one specified in Theorem 4.4.1. The only difference is that we change the fixed design point x to the random design point X.

Corollary 5.4.2 *The length of interval I_n, $len(I_n) \to 0$ in probability as $n \to \infty$.*

Proof: It is easy to see that $len(I_n) = X_{t_n - k_n} - X_{t_n + k_n} \leq \Theta_p(n^\alpha)/n = \Theta_p(n^{\alpha-1}) \to 0$ in probability. ■

5.5 Simulation study

We investigated the method for a single nonparametric regression models as in (5.1). The model assume mean response function is $\mu(X_i) = X_i^2 + \cos(\pi X_i) + \frac{1}{2} I(X_i > 0)$, where X_i is distributed as uniform distribution on interval $(-1, 1)$. Also, error term ϵ_i is generated from $N(0, [|0.2\cos(2\pi X_i)| + 0.1]^2)$. Figure 5.3 displays the heteroskedasticity of the error terms. Notice that the variance is bounded above and below by 0.3 and 0.1 respectively, which satisfy condition (5.2). From Figure 5.4, the jump point is not very clear in the scatter plot and blurred by non-constant variance. However, Figure 5.5 shows obviously a jump at $x = 0$ for the mean response function.

We generated the nonparametric regression model with $n = 1000$ samples 20 times, then processed these 20 datasets with similar methodology as in chapter 4. The pointwise estimators are obtained from local regression with degree 2 by 50 grid points a's that are equally spaced on $[-0.98, 0.98]$.

At first, We let β_n to be 30 from our experience and pick bandwidth h by C_p criteria as (2.19) for the naive compound estimators of mean response and derivative. Figure 5.6 presents the compound estimator of mean response, which recovers $\mu(x)$ very well except in the neighborhood of change point $x = 0$.

Next step is to figure out the order k of Empirical derivative as (5.12) such that we could have a big signal of the jump point. Again, we let $k = n^\alpha$ as in section 4.5 and $n = 1000$ is the sample size, then α is obtained by maximizing the quantity Q_n as (4.64). We ignored the first and last k data points when calculating Q_n. In our simulation, we selected α from 100 candidate values equally spaced on $[0.45, 0.6]$. Once k is picked, the Empirical derivative could be calculated as (5.12), and the interval estimator of jump point is obtained by (5.35).

Figure 5.7 presents the relations among true derivative, Compound derivative estimation and the Empirical derivative estimation. From the plot, we know both Empirical derivative and the Compound estimation will behave abnormally when near the change point $x = 0$. However, Empirical derivative will be much more absurd. Therefore, the absolute difference of these two estimate will indicate where the jump is, as in Figure 5.8. We should notice that non-constant variance does have effects on the Empirical derivative. Figure 5.3 tells us the variance of error terms achieve its maximum when x equals to -0.5 and 0.5, which corresponding to the higher spikes on Figure 5.7 when x is around -0.5 or 0.5. However, condition (5.2) ensure us that the effects of non-constant variance will not exceed the jump point as n goes to infinity. In practice, the candidate value of α could be based on the plot of compound estimation and Empirical first derivative. If the plot presents an obvious big variance or bias of the Empirical derivative, then the α value may be too small or large.

Table 5.1 shows us the simulation results from the model. For all 20 trials, change point $x = 0$ was contained in the interval estimate. IndMax stands for the index of x_i that has the largest gap and LocMax means the location of the highest gap. We could see the data point which gives us the biggest gap is always around the jump point $x = 0$. Overall, our method seems to work well for moderately large data sets with a single change point and moderate heteroskedasticity. However, practically, the data with small samples and huge signal to noise ratio may lead to trouble with the method. Theoretically, our method needs to assume that the sample size n goes to infinity and variance is bounded by constants. Therefore, the assumption may be seriously violated with small samples and huge SNR. In that case, a parametric model for the change point may be a better choice (Reference?????).

Figure 5.3: Variance function of ϵ_i

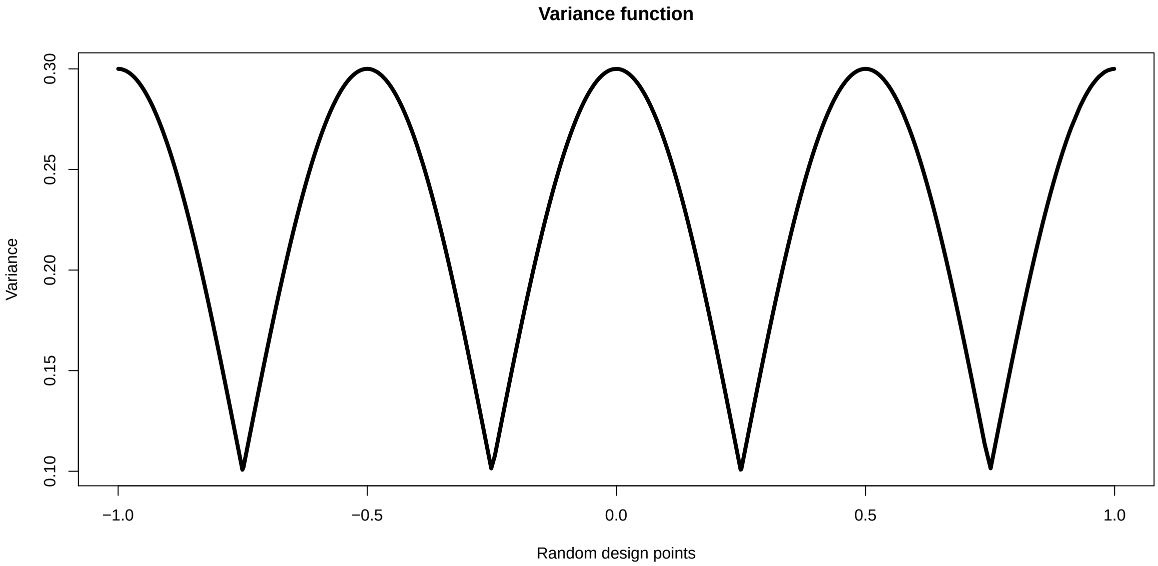


Figure 5.4: Scatter plot

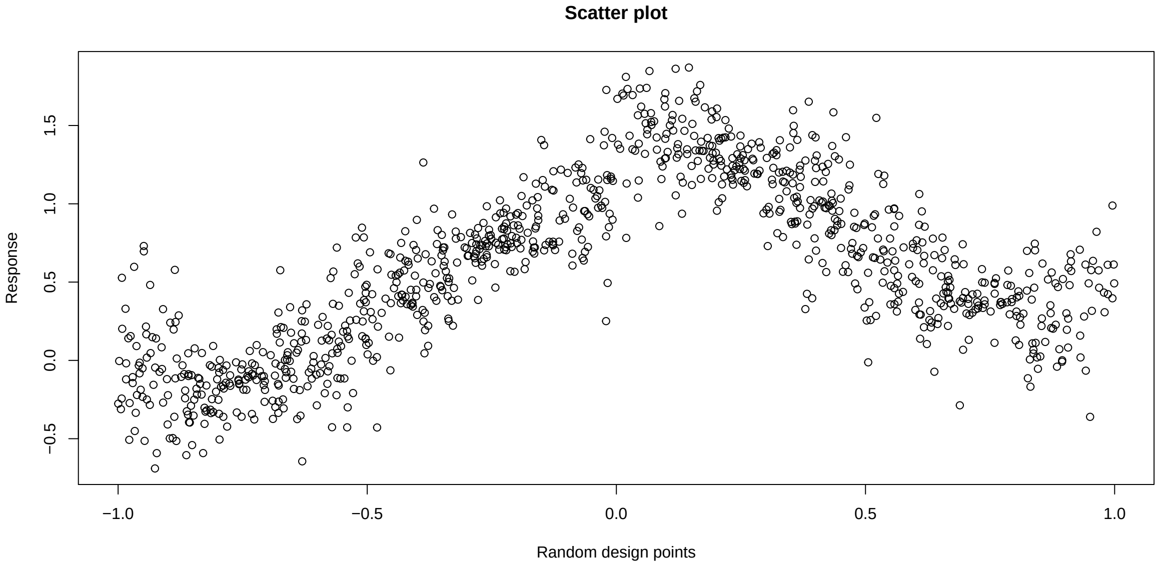


Figure 5.5: Mean response function

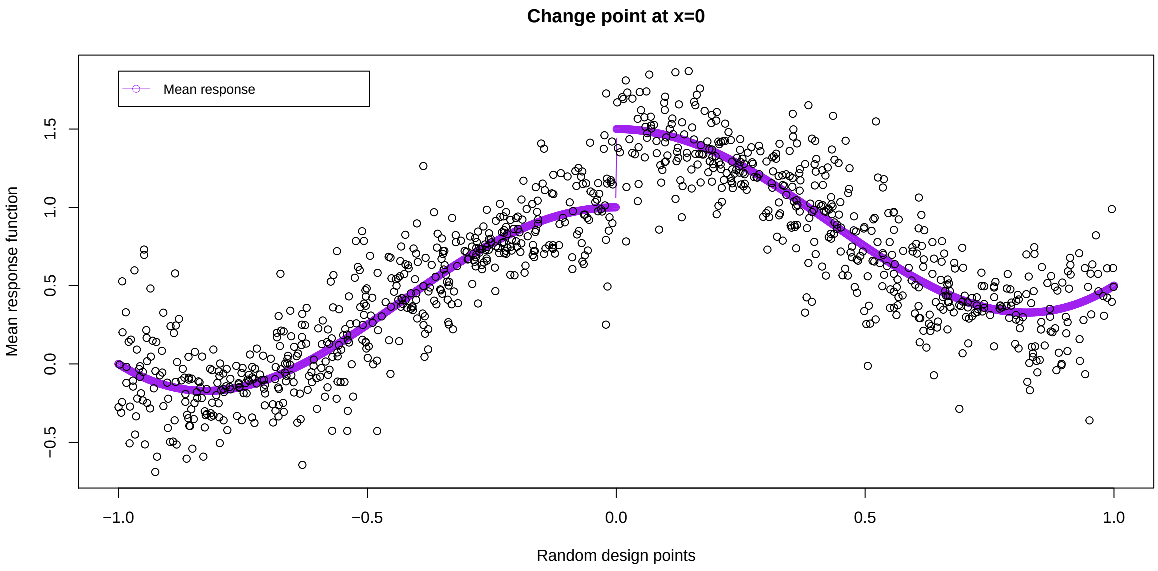


Figure 5.6: Compound Estimation

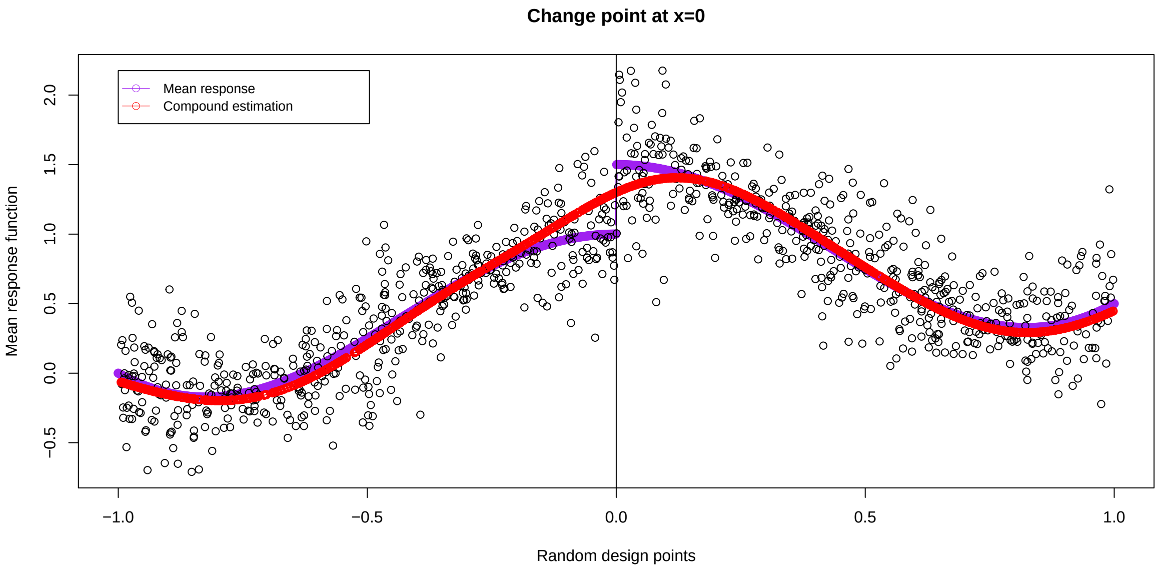


Figure 5.7: True Derivative, CPE and Empirical derivative

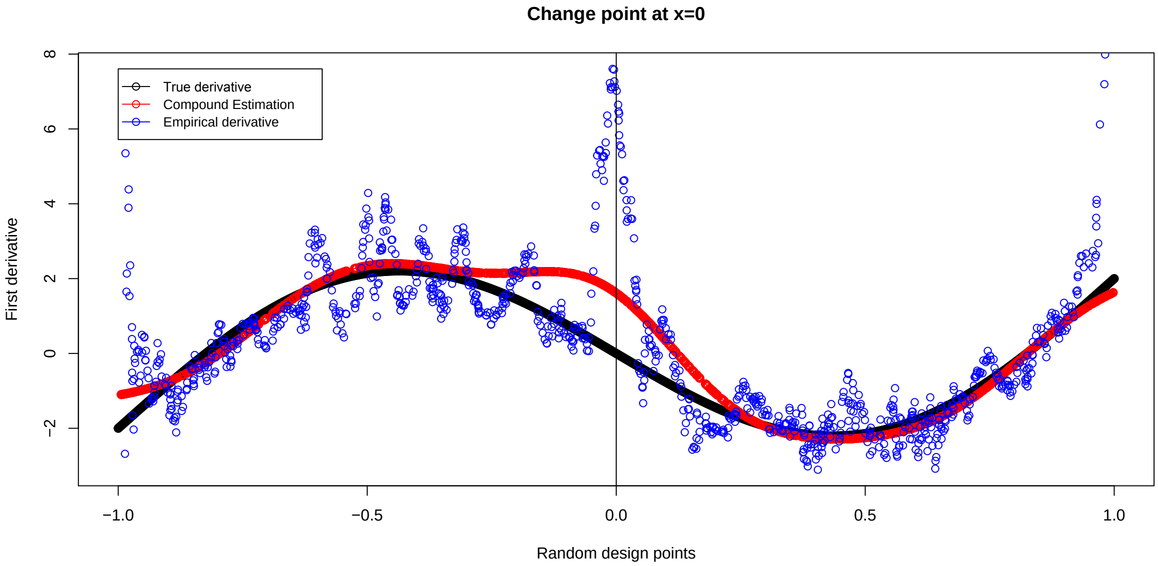


Figure 5.8: Distance between CPE and Empirical derivative

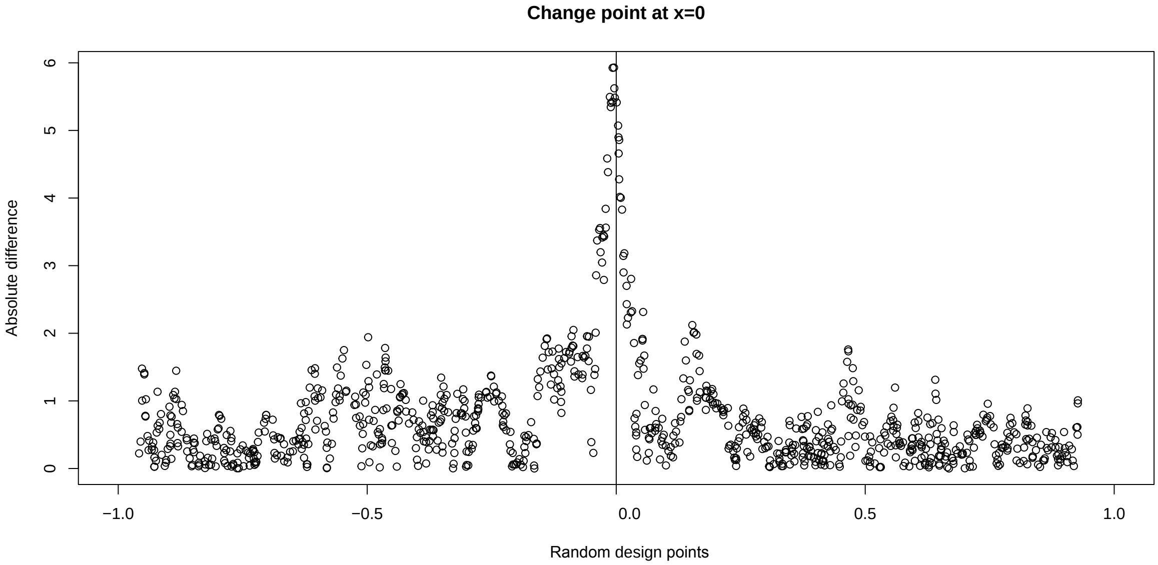


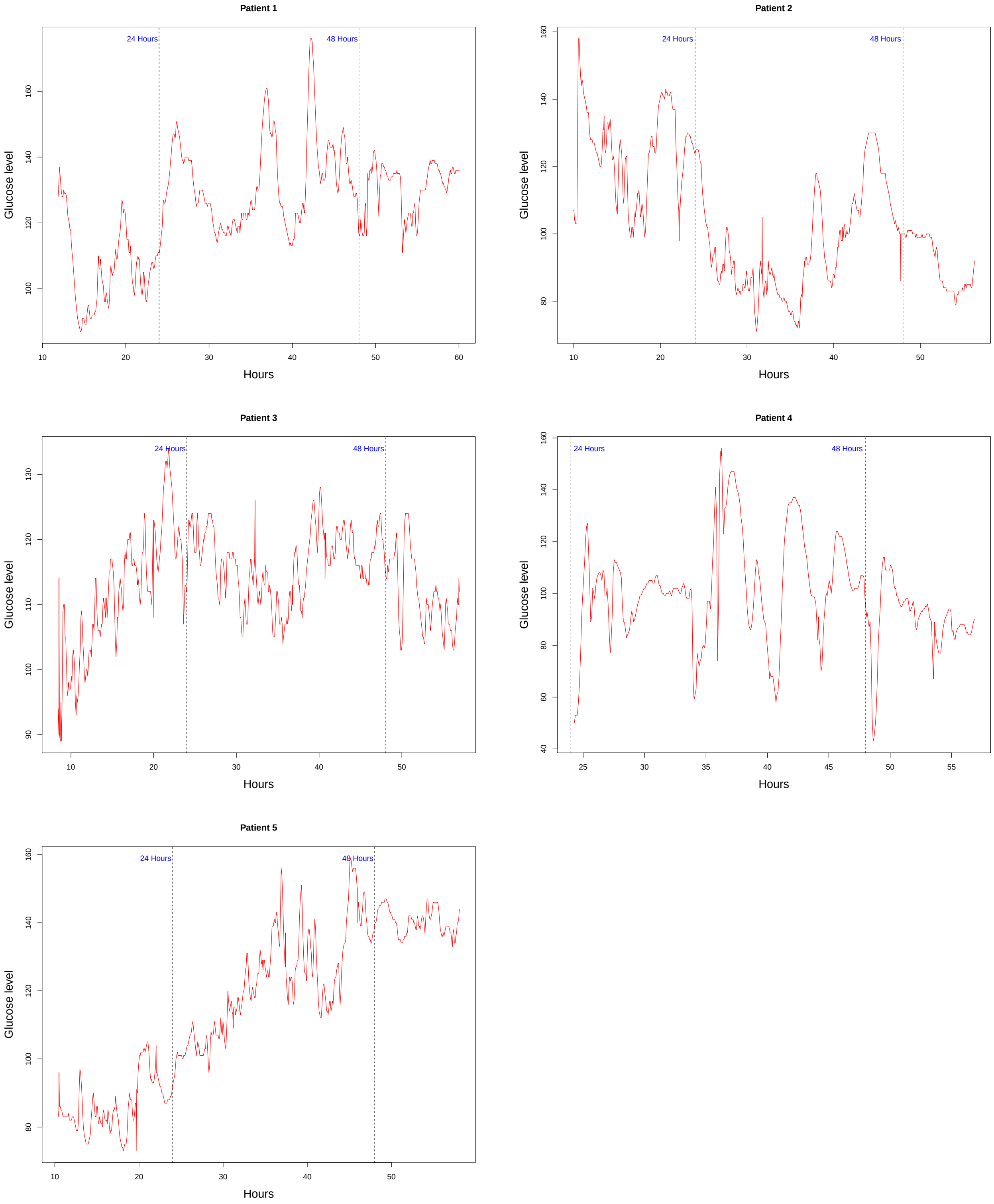
Table 5.1: Results from 20 trials for first model

Trials	h	α	k	IndMax	LocMax	Interval Estimator	percent
1	0.262	0.492	30	523	0.030	(-0.032, 0.100)	6.62%
2	0.378	0.515	35	480	-0.018	(-0.088, 0.042)	6.53%
3	0.312	0.488	29	492	0.010	(-0.059, 0.059)	5.92%
4	0.283	0.461	24	496	0.010	(-0.027, 0.065)	4.60%
5	0.156	0.483	28	519	0.013	(-0.052, 0.059)	5.57%
6	0.417	0.503	32	460	-0.049	(-0.100, 0.049)	7.44%
7	0.057	0.461	24	485	-0.015	(-0.047, 0.037)	4.20%
8	0.097	0.450	22	529	-0.003	(-0.043, 0.036)	3.93%
9	0.311	0.488	29	494	-0.004	(-0.058, 0.042)	5.03%
10	0.370	0.461	24	486	0.004	(-0.067, 0.042)	5.46%
11	0.281	0.503	32	514	-0.001	(-0.052, 0.060)	5.58%
12	0.118	0.450	22	495	0.031	(-0.021, 0.073)	4.70%
13	0.156	0.450	22	480	-0.004	(-0.054, 0.033)	4.35%
14	0.198	0.461	24	480	-0.013	(-0.058, 0.029)	4.36%
15	0.225	0.511	34	506	-0.005	(-0.089, 0.056)	7.23%
16	0.047	0.508	33	485	0.030	(-0.047, 0.076)	6.16%
17	0.153	0.567	50	521	0.028	(-0.074, 0.112)	9.28%
18	0.102	0.455	23	514	0.004	(-0.046, 0.059)	5.24%
19	0.435	0.461	24	473	-0.018	(-0.058, 0.021)	3.94%
20	0.282	0.527	38	515	-0.002	(-0.070, 0.072)	7.10%

5.6 Real data application:

In this section, we will apply the method established in Chapter 4 to analyze glucose data for several patients, some of whom may have diabetes. The patients' data are collected via a subcontract grant support from NIH 4P30DK020579-39 and a UK CCTS pilot grant. Five patients had their glucose levels measured over two days. The measurements were conducted every 5 minutes. If we number the patients from 1 to 5, the 4th patient had many missing measurements in the first day. All patients have reported their own meal times in two days. If a patient eats a meal, then his or her glucose level could have a big change, thus we may be able to locate an interval of the meal time of each patient by our method in Chapter 4. Our interests are focused on whether the intervals are consistent with the self-report times from the patients. Plots in Figure 5.9 show the relations between the glucose level and the timing of measurements for patients. Each plot exhibits a highly autocorrelation of

Figure 5.9: Patients' glucose data



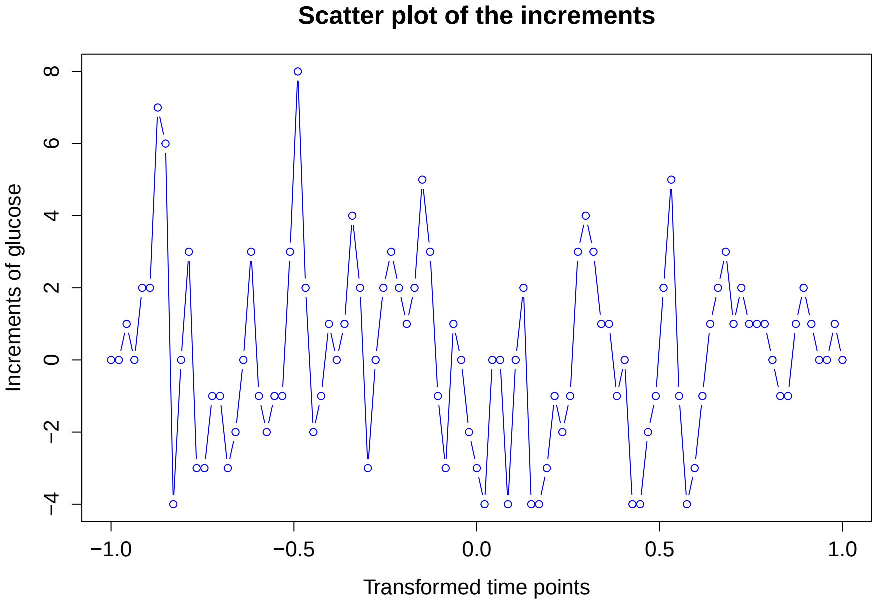
their glucose levels. However, the jump point detection method assumes the error terms are independent of each other. Therefore, we take differences between two consecutive glucose levels and assume the increments of these two glucose levels are independent. If each glucose level is denoted as G_i, then the model assumption is:

$$G_{i+1} - G_i = \mu(x_i) + \epsilon_i \tag{5.45}$$

for $i = 1, 2, 3...n$. x_i's are the times of measurements. For applying our method in Chapter 4, we looked at the time window between 7 hours before and 1 hour after the self-reported bedtime and assume they have only one major meal between these two times. Namely, there is only one jump point in this time window. These won't be strictly true, however, we could at least detect one of the major changes during that time and possible find other meals in this time window. We may not be able to give a theoretical justified interval estimate for other changes, however, practically, a plausible rough statement of other changes could be possible. Adhocly, for example, we could let the second interval estimate of another change to be the (x_{j-k}, x_{j+k}), with j such that $Y_j^{(1)} - \widehat{\mu(x_j)}$ has a the maximum peak outside the first interval estimate. Let $Y_i = G_{i+1} - G_i$, also we will transform the times into interval $[-1, 1]$, since measurements are recorded every 5 minutes, the timing points x_i's will be equally spaced on interval $[-1, 1]$. Figure 5.10 shows us the scale of 8 hours window after transformation of time points and the glucose increments for the first day of the first patient. It seems that the autocorrelations are reduced a lot after taking the subtraction.

In our glucose data, the 8 hours window only include no more than 100 time points for each patient in each day. Therefore, we let order $k = 3$ or 4 to avoid overfitting the data with empirical derivative. Whether k equals to 3 or 4 relies on the bandwidth of compound estimation. Moreover, since we are interested in the positive of glucose increments, a little modification was imposed on the method in section 4.5. We

Figure 5.10: Patient 1 Day 1

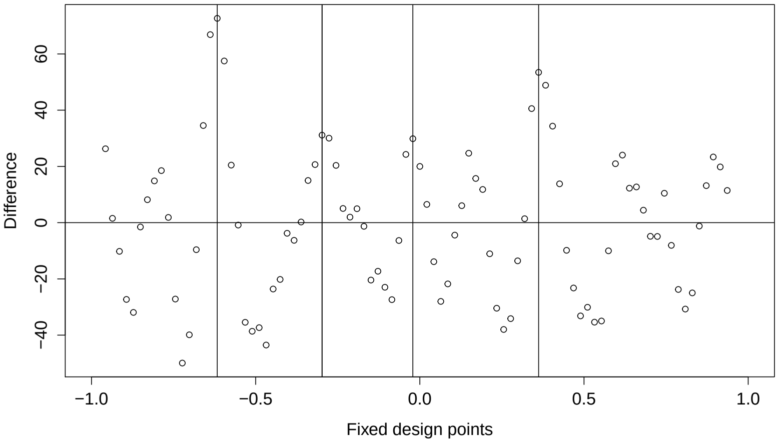


maximize the quantity

$$Q_n = \frac{Y_l^{(1)} - \mu^*(x_l)}{\max_{j \notin \{l-k,\ldots l+k\} \cup \{1,\ldots k\} \cup \{n-k+1,\ldots n\}} \left(Y_j^{(1)} - \mu^*(x_j) \right)} \tag{5.46}$$

instead of the Q_n defined as (4.64). The rest of the application will be the same as in section 4.5. We pick 10 grid points for the compound estimation with the convolution weight $\beta = 60$ and polynomial order 2 of local regression. Bandwidth h in local regression was chosen by Cp criteria from chapter 2. If we obtain a large bandwidth from the Cp criteria, let $k = 4$ would be a better choice, or else, let $k = 3$. Because if the bandwidth is small, compound estimation itself will be somehow wiggled, we may need a wiggler empirical derivative to produce a visible difference between them. First 4 peaks along with their interval estimates are computed as the ad hoc way we mentioned before. However, these do not mean all the peaks we found make sense biologically. Whether these point estimates or interval estimates make 100% sense will depend on the decisions of specialists from other area. We could describe the results numerically and trying to convert them into very limited knowledge about the

Figure 5.11: Patient 1, Day 2.



glucose level.

Table 5.2 display the 4 peaks and the ad hoc interval estimates of them from plot of distance between compound derivative estimate and Empirical derivative. Figure 5.11 is the example of plotting the difference from second day of the first patient. It is similar to Figure 5.8 in previous simulation except they present several peaks and negative differences are considered. We take these four design points and transform them back to hours. The meal times and choices of k and bandwidths are shown in the table.

Glucose level will be affected by many factors. It depends on a patient's health condition, whether they did exercise or consumed some food. Even if a patient had a meal, their glucose may increase corresponding to what kind of food they consumed. Their glucose level may increase immediately if they have some food called "fast carbs", like an ice-cream or coke. Also, if pizza or pasta was eaten, then the glucose may increase several hours after the meal time. This kind of food is called "slow carbs". Thus, each patient may have multiple changes of glucose level. We will pick the interval that is closest to the self-reported meal time as our interval estimate of the meal time. Figures 5.12 to 5.16 show us the locations of peaks and the interval estimates on the plots of time v.s. Glucose level. Four different colors (red, blue, green, pink) in that order correspond to four peaks respectively. The thicker the lines are, the higher the peaks. From Figures 5.12 to 5.16, the changes of increments of glucose level somehow exactly correspond to the changes of glucose level. For instance, the first three intervals we found in day 1 of Figure 5.12 contain three turning points. The last interval also indicates a change of the glucose increment since the glucose level was flat at first and increased suddenly. These may tell us that generally the changes of glucose increments are stable unless at the turning point of the glucose level, which may indicate the starting time of consumption. We also used a dashed purple vertical line to represent the self-reported meal time.

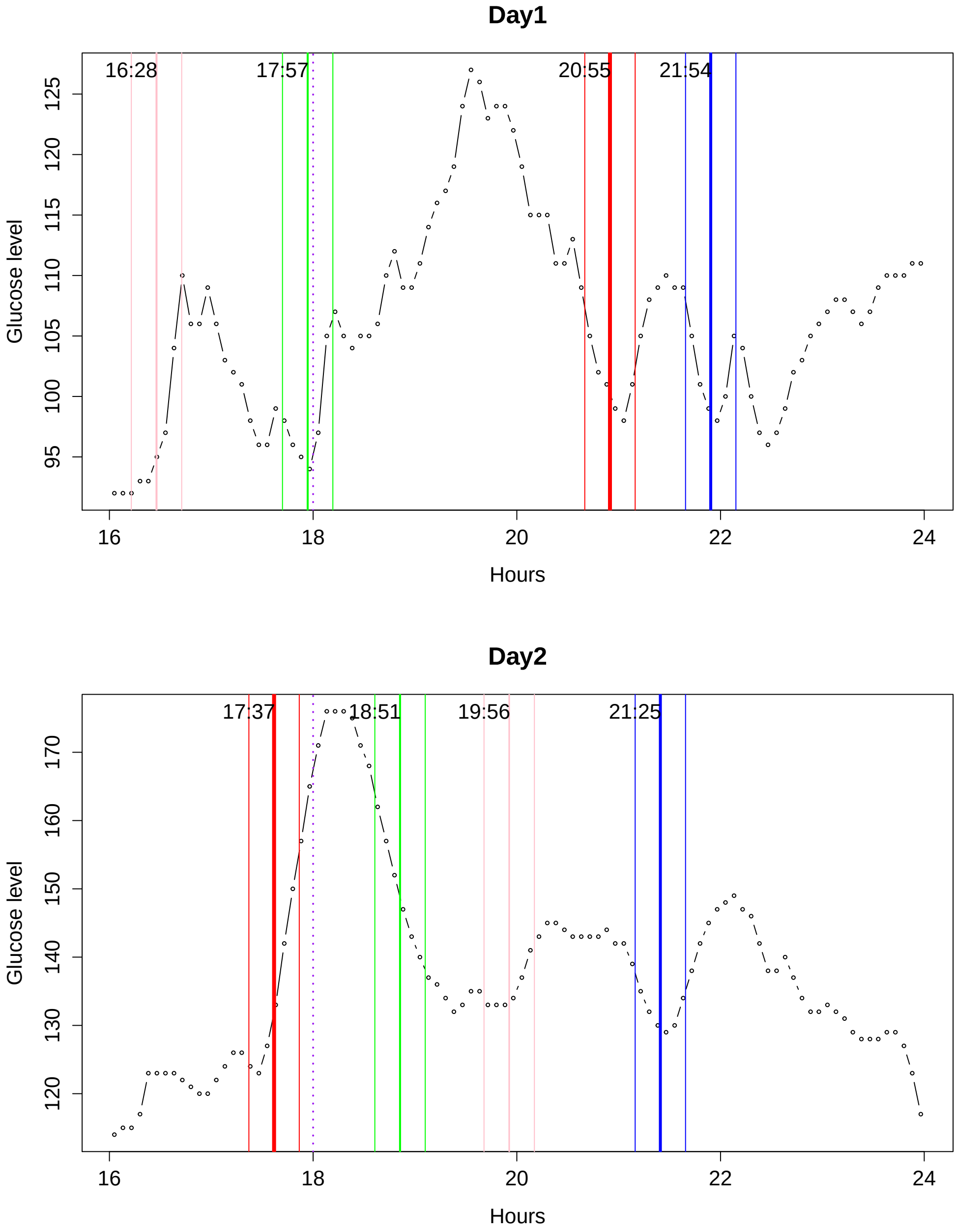
From Table 5.2, we could see the self-reported meal times are generally very close or inside one of the interval estimates except day 2 of third patient and fifth patient. Nevertheless, second picture in Figure 5.14 shows us that the glucose has already increased before the meal time and he may eat meal before 18:00 instead of the self-reported time. Table 5.2 tells us he may eat meal during 16:54 to 17:33. Also, the second picture on Figure 5.16 shows that the glucose level was decreasing after the self-reported meal time. The patient may have eaten the meal before that time, like during 17:16 and 17:45 from Table 5.2, or he may had some “fast carbs” before the meal time and had something during the meal which will cost time to start affecting the glucose.

Table 5.2: Possible changes for each patients

Patient	1	1	2	2	3	3	4	5	5
Day	1	2	1	2	1	2	2	1	2
meal	18:00	18:00	18:00	18:00	18:00	18:00	20:30	18:00	18:00
k	3	3	3	4	4	4	3	4	3
h	0.32	0.25	0.22	0.88	0.62	0.88	0.37	0.88	0.31
Peak1	20:55	**17 : 37**	**18 : 18**	19:06	19:32	15:46	**20 : 26**	19:21	16:37
I1	(20:40, 21:10)	(17:22, 17:52)	(18:03, 18:32)	(18:47, 19:25)	(19:12, 19:51)	(15:27, 16:05)	(20:12, 20:41)	(19:01, 19:40)	(16:22, 16:52)
Peak2	21:54	21:25	16:26	**17 : 44**	**18 : 28**	19:29	16:53	21:36	19:53
I2	(21:39, 22:09)	(21:10, 21:39)	(16:12, 16:41)	(17:24, 18:03)	(18:08, 18:47)	(19:09, 19:48)	(16:39, 17:08)	(21:17, 21:56)	(19:38, 20:07)
Peak3	**17 : 57**	18:51	21:56	21:31	15:37	17:13	21:20	**18 : 18**	15:48
I3	(17:42, 18:12)	(18:36, 19:06)	(21:41, 22:10)	(21:12, 21:51)	(15:17, 15:56)	(16:54, 17:33)	(21:05, 21:34)	(17:58, 18:37)	(15:33, 16:03)
Peak4	16:28	19:56	19:21	23:23	20:35	16:30	19:28	16:41	17:31
I4	(16:13, 16:43)	(19:41, 20:10)	(19:06, 19:35)	(23:03, 23:42)	(20:16, 20:55)	(16:10, 16:49)	(19:14, 19:43)	(16:21, 17:00)	(17:16, 17:45)

Here, meal represent the self-reported meal time for each patient, k and h correspond to the order of empirical derivative and the bandwidth we chosen for Local regression. Peak1-Peak4 and I1-I4 correspond to the locations of peaks and the interval estimates of consumption time.

Figure 5.12: Patient 1



Day1

Day2

Figure 5.13: Patient 2

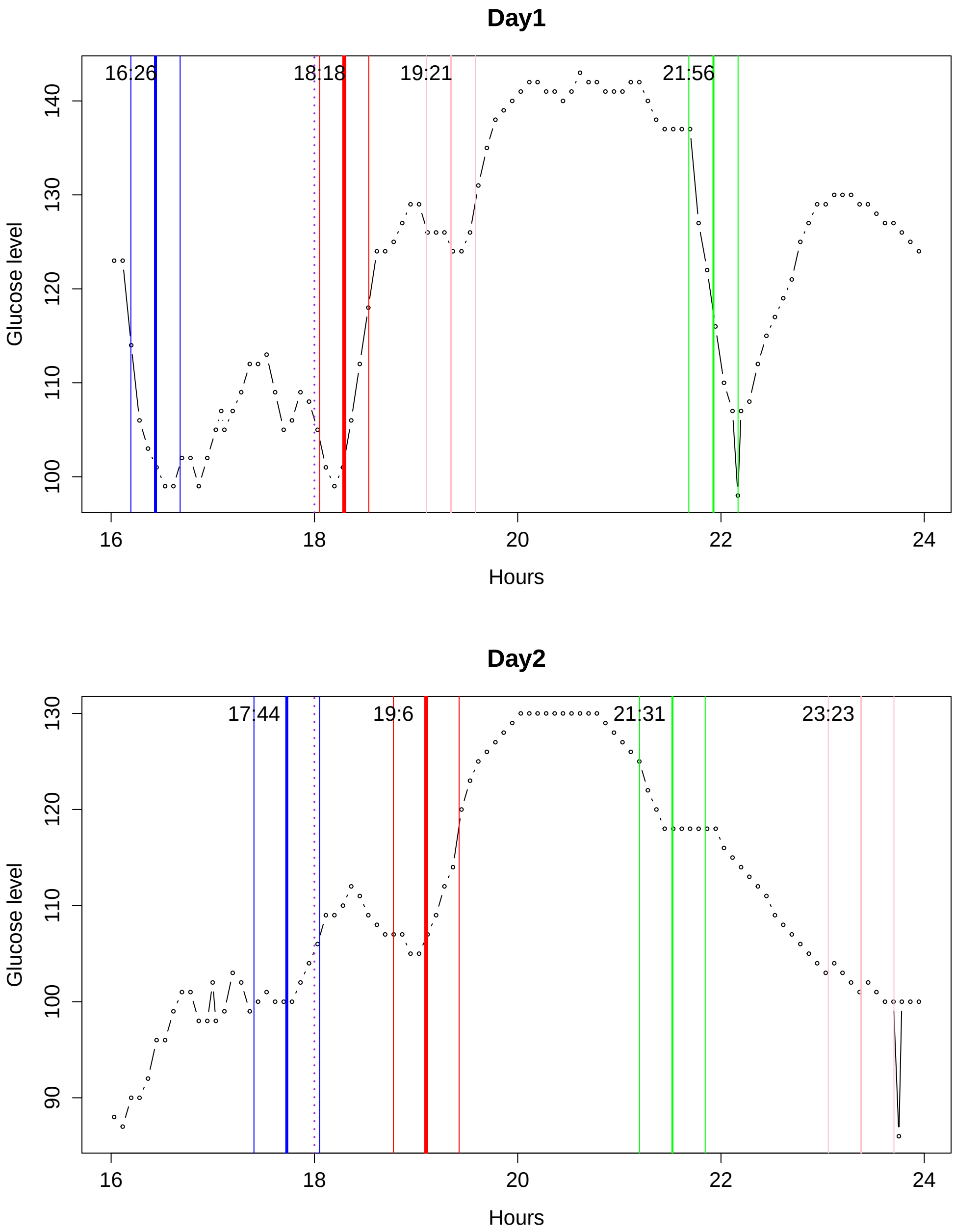


Figure 5.14: Patient 3

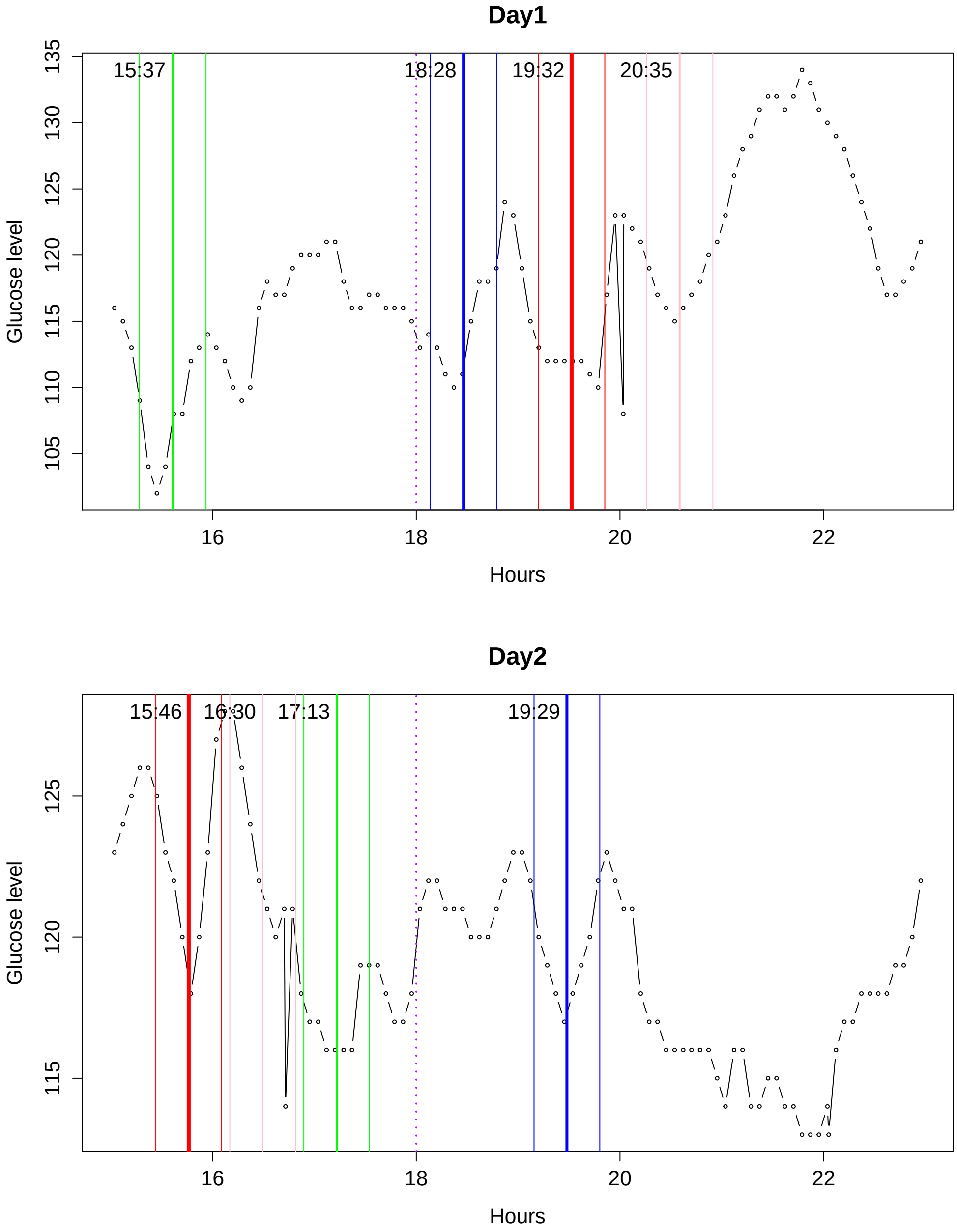


Figure 5.15: Patient 4

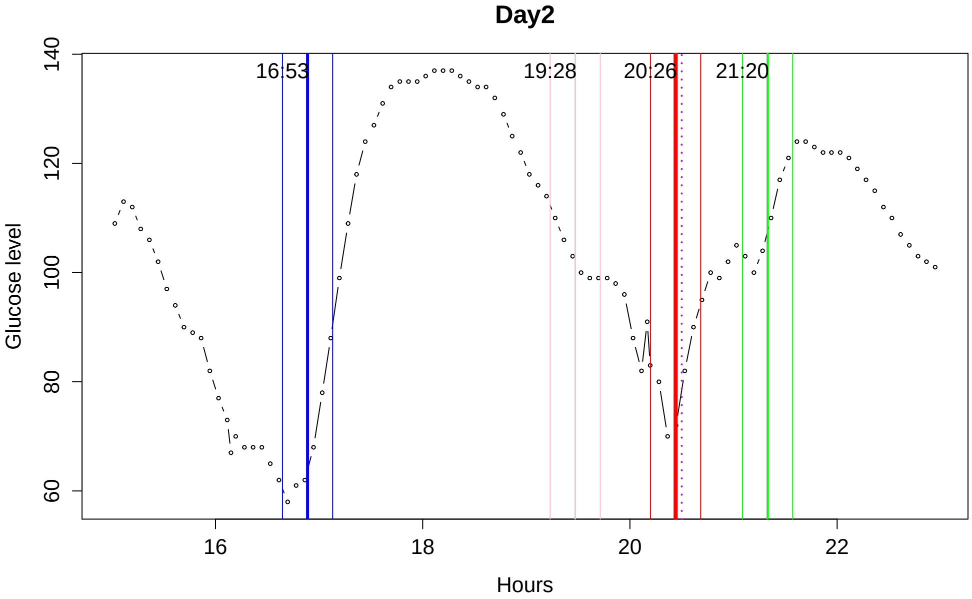
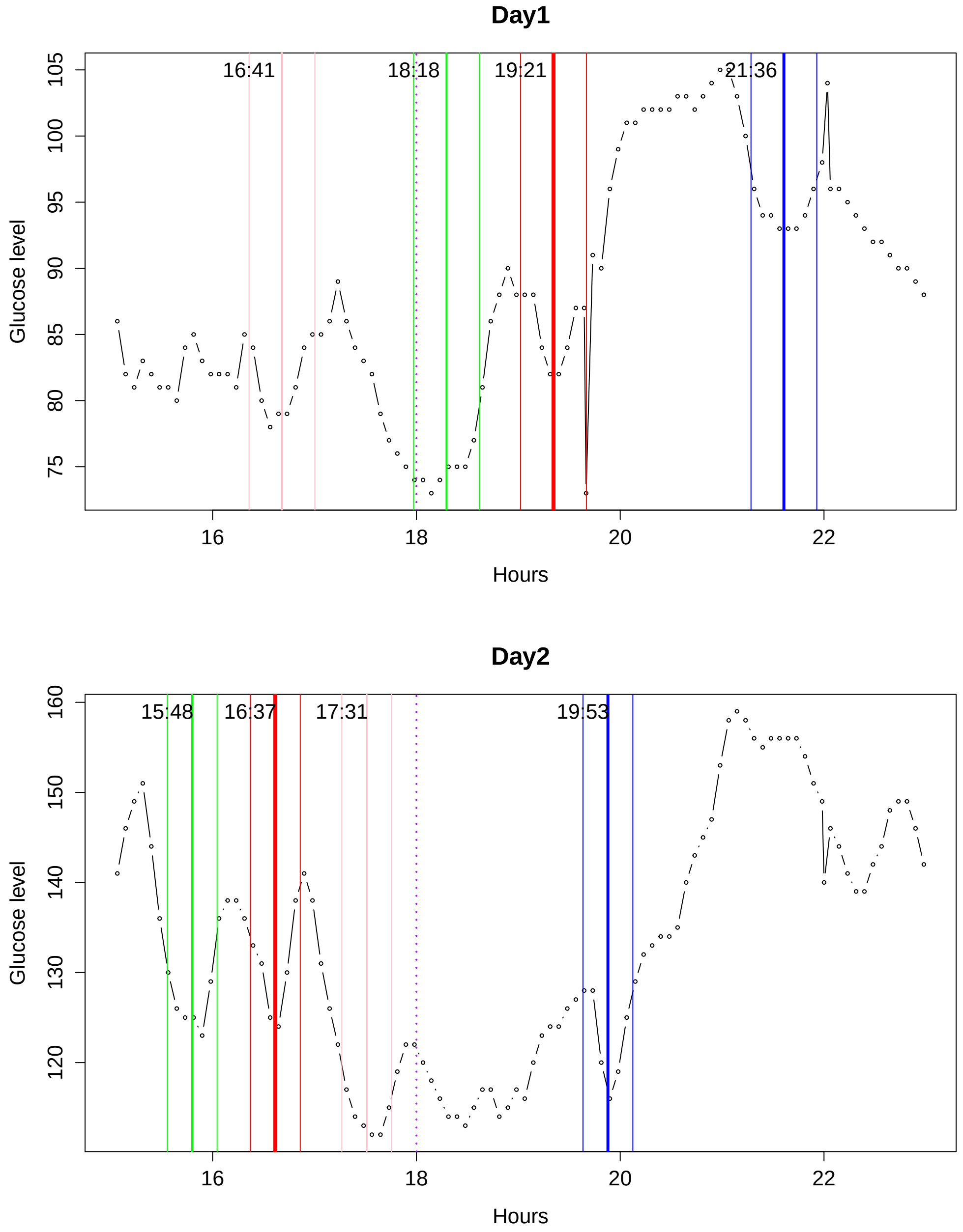


Figure 5.16: Patient 5



Day1

Day2

Bibliography

Clive Loader. *Local regression and likelihood*. Springer Science & Business Media, 2006.

Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press, 1996.

L. Breiman. *Classification and Regression Trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984. ISBN 9780534980542. URL `https://books.google.com/books?id=mlZgQgAACAAJ`.

Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

Theo Gasser and Hans-Georg Müller. Kernel estimation of regression functions. In *Smoothing techniques for curve estimation*, pages 23–68. Springer, 1979.

Jianqing Fan and Irene Gijbels. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, pages 2008–2036, 1992.

David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003.

Richard Charnigo and Cidambi Srinivasan. Self-consistent estimation of mean response functions and their derivatives. *Canadian Journal of Statistics*, 39(2):280–299, 2011.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN 9780387790527. URL `https://books.google.com/books?id=mwB8rUBsbqoC`.

Richard Charnigo, Benjamin Hall, and Cidambi Srinivasan. A generalized c p criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011.

Irène Gijbels. Smoothing and preservation of irregularities using local linear fitting. *Applications of Mathematics*, 53(3):177–194, 2008.

Hans-Georg Müller, Ulrich Stadtmüller, et al. Discontinuous versus smooth regression. *The Annals of Statistics*, 27(1):299–337, 1999.

Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

R Raj Bahadur. A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37(3):577–580, 1966.

Vita

Education

- Master of Science in Statistics, University of Kentucky, 2012-2014
- Bachelor of Science in Statistics, Hunan Normal University, 2007-2011

Employment

- Research assistant, 2015-2017
- Teaching assistant, 2013-2015