Asymptotic properties of Non-parametric Regression with Beta Kernels

by

Balasubramaniam Natarajan

B.E.(Honors), Birla Institute of Technology and Science, 1997

Ph.D., Colorado State University, 2002

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Abstract

Kernel based non-parametric regression is a popular statistical tool to identify the relationship between response and predictor variables when standard parametric regression models are not appropriate. The efficacy of kernel based methods depend both on the kernel choice and the smoothing parameter. With insufficient smoothing, the resulting regression estimate is too rough and with excessive smoothing, important features of the underlying relationship is lost. While the choice of the kernel has been shown to have less of an effect on the quality of regression estimate, it is important to choose kernels to best match the support set of the underlying predictor variables. In the past few decades, there have been multiple efforts to quantify the properties of asymmetric kernel density and regression estimators. Unlike classic symmetric kernel based estimators, asymmetric kernels do not suffer from boundary problems. For example, Beta kernel estimates are especially suitable for investigating the distribution structure of predictor variables with compact support. In this dissertation, two types of Beta kernel based non parametric regression estimators are proposed and analyzed. First, a Nadaraya-Watson type Beta kernel estimator is introduced within the regression setup followed by a local linear regression estimator based on Beta kernels. For both these regression estimators, a comprehensive analysis of its large sample properties is presented. Specifically, for the first time, the asymptotic normality and the uniform almost sure convergence results for the new estimators are established. Additionally, general guidelines for bandwidth selection is provided. The finite sample performance of the proposed estimator is evaluated via both a simulation study and a real data application. The results presented and validated in this dissertation help advance the understanding and use of Beta kernel based methods in other non-parametric regression applications.

Asymptotic properties of Non-parametric Regression with Beta Kernels

by

Balasubramaniam Natarajan

B.E.(Honors), Birla Institute of Technology and Science, 1997

Ph.D., Colorado State University, 2002

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Major Professor
Dr. Weixing Song

Copyright

Abstract

Kernel based non-parametric regression is a popular statistical tool to identify the relationship between response and predictor variables when standard parametric regression models are not appropriate. The efficacy of kernel based methods depend both on the kernel choice and the smoothing parameter. With insufficient smoothing, the resulting regression estimate is too rough and with excessive smoothing, important features of the underlying relationship is lost. While the choice of the kernel has been shown to have less of an effect on the quality of regression estimate, it is important to choose kernels to best match the support set of the underlying predictor variables. In the past few decades, there have been multiple efforts to quantify the properties of asymmetric kernel density and regression estimators. Unlike classic symmetric kernel based estimators, asymmetric kernels do not suffer from boundary problems. For example, Beta kernel estimates are especially suitable for investigating the distribution structure of predictor variables with compact support. In this dissertation, two types of Beta kernel based non parametric regression estimators are proposed and analyzed. First, a Nadaraya-Watson type Beta kernel estimator is introduced within the regression setup followed by a local linear regression estimator based on Beta kernels. For both these regression estimators, a comprehensive analysis of its large sample properties is presented. Specifically, for the first time, the asymptotic normality and the uniform almost sure convergence results for the new estimators are established. Additionally, general guidelines for bandwidth selection is provided. The finite sample performance of the proposed estimator is evaluated via both a simulation study and a real data application. The results presented and validated in this dissertation help advance the understanding and use of Beta kernel based methods in other non-parametric regression applications.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I was never blessed with a strong memory of people, places or events. However, there is one moment from 1999 that I vividly remember. It was a beautiful crisp spring morning in Fort Collins, Colorado. I was sitting in the office of my "Theory of Statistics" instructor Dr. Duane Boes, distinguished statistician and an excellent teacher. He looked me in the eye and told me that I should be doing my Ph.D. in Statistics and not in Engineering. "Maybe some day" I responded with a wry smile knowing very well that it would be impossible for me to do that. Growing up in India, I was programmed to complete whatever I started and doing something else midway would be considered a failure/crime! At the end of a successful semester with Dr. Boes, he brought in his friend and colleague Dr. Hari Iyer to convince me that I would make a really good statistician. Hari understood the Indian psyche and told me that at some point later in my life I should get a Ph.D. degree in statistics. That was when the seed was planted. With the help of many important people in my life, it has sprouted and borne fruit in the form of this dissertation.

My beautiful wife Rekha and my wonderful boys Advith (10) and Ayavanth (6) can blame Dr. Boes and Dr. Hari for kick starting this journey. As I slowly navigated through extensive coursework and research, they were always there to give me my space and time. The boys never complained about dad going to work on sundays. Rekha was supportive of me pursuing my dreams even as she was pursuing her Ph.D., working a full time job and taking care of the boys. Words cannot describe the sacrifices she has made to help me in this effort.

Many adventures in life are easier to start than to finish. This is definitely the case for research and a doctoral program. Luckily, I had the best guide possible for my adventure - Dr. Weixing Song. He is the kindest and most patient adviser who was very understanding of my time/family/work constraints throughout this prolonged multi-year process. I learned

a lot from him both in class and outside of it. His ability to develop elegant proofs is something that I will continue to try to emulate.

I also want to share my sincere gratitude to my doctoral committee members Dr. Weixin Yao, Dr. James Neill and Dr. Don Gruenbacher. Don, who is also my boss, never once questioned my commitment to my primary duties in the Electrical Engineering department as I was taking classes and doing research on the other side of campus. I greatly appreciate his unflinching support and friendship through the years. Jim was the department head when I went up to him and shared this idea of doing a part time PhD. He probably thought I was nuts but was very encouraging through the entire journey. Dr. Yao allowed me to attend his lectures on mixture models (without registering for the class) and has consistently provided useful feedback in the development of this research work.

The PhD dream would have ended prematurely had I not passed the three grueling 6 hour qualifying exams. With probably just a day to prepare for all of these exams, I am very grateful to my good friend Indu Seetharaman who shared her course notes to help me get through the exams. Last but not the least, my parents back in India were constantly praying for my success. Given my glacial pace of progress, they probably had doubts if I will ever get to the finish line. I think the doubts and concerns just made their prayers more intense and I cannot thank them enough for that. Without their blessings and the grace of my spiritual Guru, Sri Sathya Sai Baba, this dissertation will not be where it is today.

Dedication

..... To my beautiful family - Rekha, Advith and Ayavanth

Chapter 1

Introduction

A regression function or regression curve describes the general relationship between predictor variable X and response variable Y. It is very useful to know this relationship as it enables us to infer/predict trends and uncover special features such as monotonicity etc. If n i.i.d samples $\{(X_i, Y_i)\}_{i=1}^n$ are collected, the regression relationship can be captured via the model,

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n \tag{1.1}$$

Here, m is the unknown regresion function, and ε_i are the observation errors typically modeled as random with zero mean and finite variance $\sigma^2 > 0$. The aim of regression analysis is to produce a reasonable approximation of the function m by reducing the effect of observational errors. This curve approximation, commonly referred to as "smoothing" , allows us to focus on the mean dependence of Y on X.

The mean function approximation can be done essentially in two ways. The most popular *parametric* approach assumes that the mean curve m has a predefined functional form that is fully described via a finite set of parameters. A classic example is that of a polynomial regression model where the parameters are the coefficients of the predictor variables. In contrast, a *non parametric* approach does not make any prior assumptions on the specific functional form of m and lets the data drive the approximation of the regression function.

In many applications, a preselected parametric model might be too restrictive and one has to rely on a non parametric smoothing approach. As non parametric regression is the focus of this dissertation, we introduce the underlying concepts in the next section.

1.1 Non-Parametric Regression

Nonparametric regression is studied in both *fixed design* and *random design* contexts. For the univariate fixed design case, the design consists of $x_1, x_2, \cdots, x_n$ which are fixed inputs. The general heteroscedastic model in this case corresponds to,

$$Y_i = m(x_i) + \sigma(x_i)\varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{1.2}$$

where ε_i are independent random variables with mean 0 and variance 1. Here, $E(Y_i) = m(x_i)$ and $\mathrm{Var}(Y_i) = \sigma^2(x_i)$. In the case of random design, we are given random bivariate samples $(X_1, Y_1), \cdots, (X_n, Y_n)$ with the corresponding model as,

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \qquad i = 1, 2, \ldots, n. \tag{1.3}$$

Here, $(X_i, \varepsilon_i), i = 1, 2, \cdots, n$ are independent and identically distributed random variables with $E(\varepsilon|X) = 0$ and $\mathrm{Var}(\varepsilon|X) = 1$. In this case, $m(x) = E(Y|X = x)$ and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$.

There are several approaches to approximating the regression function m in the non parametric framework. The most popular methods are the ones based on kernel functions, spline functions and wavelets with each one having its unique strengths and drawbacks. In this dissertation, our analysis is focused on kernel regression as it offers both mathematical and intuitive simplicity as discussed next.

1.2 Kernel Regression

As stated in the previous section, kernel smoothing techniques are powerful tools to estimate $m(x)$ in a non parametric framework. The theoretical and empirical properties of kernel estimators are well-known in the classical fixed design and random design frameworks (see, e.g., [Wand and Jones(1995)], p. 115). In this section, we present a brief overview of the kernel regression approach.

For the regression model introduced in (1.3) , our goal is to find the regression function $m(x) = E(Y|X = x)$. If X takes on only a finite set of values, then a simple strategy would be to use the conditional sample means and by invoking the law of large numbers we can be confident that $\hat{m}(x) \to E(Y|X = x)$. This approach will not work if X is continuous as the probability of getting a sample at any particular value is 0! That is, the function to be estimated will always be undersampled and we need to fill in or "smooth" between the values that we observe. Linear smoothers [Buja(1989)] represent a class of popular smoothers that exploit a weighted linear combination of the responses taking the form

$$\hat{m}(x) = \sum_i Y_i \hat{w}(X_i, x) \tag{1.4}$$

The sample mean is a special case with $\hat{w}(X_i, x) = 1/n$ and ordinary least squares linear regression (without intercept) [Weisberg(2005)] is $\hat{w}(X_i, x) = \frac{X_i}{\sum_i X_i^2} x$. These simple smoothers ignore the distance of X_i from x. k-Nearest neighbor regression [Altman(1992)] involves setting the $\hat{w}(X_i, x) = 1/k$ if x_i is one of the k neighbors of x and $\hat{w}(X_i, x) = 0$ otherwise. Changing k in this setup allows us to change the amount of smoothing we are doing on the data. In 1964, [Nadaraya(1964)] proposed an elegant way to use the data in a location sensitive manner using a *Kernel function* $K_h(x - X_i)$ that is a function of both X_i, the location x and a smoothing parameter/bandwidth h. The Nadaraya Watson (NW)

estimator corresponds to [Nadaraya(1964)]

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}. \tag{1.5}$$

That is, the NW estimator is a linear smoother with $\hat{w}(X_i, x) = \frac{K_h(X_i - x)}{\sum_i K_h(X_i - x)}$. The Kernel function has to satisfy some basic properties that enable its use in this set up as discussed in [Wand and Jones(1995)]. Following the work of Nadaraya and Watson, Priestley and Chao proposed an alternative smoother in 1972 specifically for the fixed design case. The PC estimator [Priestley and Chao(1972)] corresponds to

$$\hat{m}_{PC}(x) = \sum_{i=1}^{n} (X_{(i)} - X_{(i-1)}) K_h(X_{(i)} - x) Y_{[i]} \tag{1.6}$$

where, $(X_{(i)}, Y_{[i]})$ denote the (X_i, Y_i) ordered with respect to X_i values. Another popular estimator for fixed design is the Gasser Müller Estimator that was proposed in 1979. This estimator corresponds to [Gasser and Müller(1979)]

$$\hat{m}_{PC}(x) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(u - x) du Y_{[i]} \tag{1.7}$$

where, $s_i = \frac{1}{2}(X_{(i)} + X_{(i+1)}), s_0 = 1, s_n = 1$. In addition to these classic smoothers, there have been many modified versions of regression estimators that researchers have introduced over the last few decades. The choice of kernel function and the bandwidth are critical in the kernel regression process as discussed in the next section. While a single universal choice of kernel and bandwidth may not be optimal for all practical problems, the quality of the resulting estimate is expected to improve with more data. The rate at which the estimate quality as measured in terms of bias, MSE and or other measures improve with increasing data is important to study. Therefore, asymptotic properties of these kernel regression estimators is a field of research that has attracted a lot of interest in the past few decades.

1.3 Kernel Density Estimation

Kernel regression is closely related to kernel based density estimation. Since $m(x) = E(Y|X = x) = \int \frac{y f_{X,Y}(x,y)}{f_X(x)} dy$, using a product kernel estimator for $f_{X,Y}(x,y)$ and a kernel density estimator for $f_X(x)$ yields the NW estimator $\hat{m}_{NW}(x)$ in (1.5). Kernel density estimate of $f(x)$ with Kernel K and bandwidth h is [Silverman(1986)],

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \tag{1.8}$$

where, K is typically chosen as a unimodal symmetric density (e.g., normal, uniform, Epanechnikov etc.) with h as the bandwidth parameter. A large value of bandwidth h leads to oversmoothing of the density function while a small value of h leads to under smoothing. As in regression function estimation, asymptotic mean squared error properties of these density estimators and its dependence on both the choice of the kernel function and bandwidth have been well investigated. Performance of kernels is usually measured in terms of mean integrated squared error (MISE) or asymptotic MISE (AMISE). Epanechnikov kernel has been shown to minimize AMISE and with the optimal bandwidth choice, the rate of convergence is of the order $n^{-\frac{4}{5}}$. A common measure used to compare different kernels is *efficiency*. Efficiency of kernel K relative to kernel K^* represents the ratio of sample sizes necessary to obtain the same minimum AMISE (for a given f) when using K^* as when using K. Table 1.1 illustrates the impact of kernel choices on the efficiency [Wand and Jones(1995)].

Epanechnikov	1.000
Biweight	0.994
Triangular	0.986
Normal	0.951
Uniform	0.930

Table 1.1: Comparison of efficiency of Kernels [Wand and Jones(1995)]

It has been widely acknowledged that the choice of kernel in both kernel density and re-

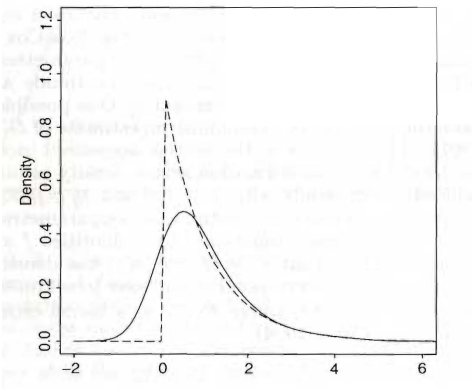


Figure 1.1: Kernel estimate of exponential density based on a sample of $n = 1000$. Solid curve is the estimate and dashed curve is the true density [Wand and Jones(1995)]

gression estimation problems is not as important as the choice of bandwidth [Turlach et al.(1993)]. However, the kernel choice has a significant impact on the quality of approximation and convergence rate at the boundary of the support set of X. Using symmetric kernels for approximating densities defined only on a subset of the real line leads to poor estimates around the boundaries as shown in figure 1.1. Typical AMISE analysis assumes that $f(x)$ satisfies some smoothness conditions over the entire real line. In the case of densities defined on a subset of the real line, the optimal choice of bandwidth at the boundaries and interior are different. One approach to address this boundary effect is to transform the data so that the support set is transformed to the entire real line or use a boundary kernel. A more effective approach to deal with this issue is to use asymmetric kernels for density and regression function estimation. For example, it is prudent to choose a kernel function that has the same support as that of the predictor variable X. Methods to overcome the boundary issues and the use of asymmetric kernels such as a Gamma and Beta kernels for both density and regression estimation has attracted significant research interest as discussed in the next section.

1.4 Asymmetric Kernel Estimation

As discussed in the previous section, one key drawback of classic kernel density estimators is the undesirable boundary issues. These boundary issues arise due to the fact that symmetric kernel functions assign positive weights outside the density support. This boundary problem is also carried over to the Nadaraya-Watson (N-W) estimators in a regression setup. Many approaches have been proposed to address the boundary problem. In the context of density estimation, see [Schuster(1985)] for a "tiedown" technique, [Marron and Ruppert(1994)] for a data transformation method, [Fan and Gijbels(1992)] for a variable bandwidth modification, [Cowling and Hall(1996)] for a pseudo-data method. In the regression set up, see [Gasser and Müller(1979)] for an asymptotic solution; [Müller(1991)] for some boundary kernels which are the solutions of a variational problem; [Müller and Wang(2007)] for a varying kernel and bandwidth method for estimating the hazard rate under random censoring; [John(1984)] for a boundary modification of the N-W kernel regression, and the references therein. However, there is continuing interest among statisticians to develop kernel-based methods for estimating the density functions or regression functions without data transformation and changing the density support.

As part of one such endeavor, there have been recent efforts on exploring the use of asymmetric kernels to estimate density functions and regression functions not supported on the entire real line. When density functions are supported on $(0, \infty)$, [Chen(2000b)] constructed a Gamma kernel density estimator and [Scaillet(2004)] proposed an inverse Gaussian kernel and a reciprocal inverse Gaussian kernel density estimator. [Mnatsakanov and Sarkisian(2012)] proposed an asymmetric kernel density estimator based on a Chen's Gamma kernel density estimator. Both estimators from Scaillet and Sarkisian suffer from inconsistency around $x = 0$ which is dealt by excluding that point from the support of the underlying distribution. [Chaubey(2012)] proposed a density estimator for non-negative random variables relying on two smoothing parameters based on a generalization of Hille's lemma [Hille and Phillips(1996)]. Recently, [Koul and Song(2013)][Shi and Song(2015)] established

asymptotic normality and uniform almost sure convergence results for the varying kernel density and regression functions estimators when the underlying random variable is positive.

While there are many discussions on the density and regression estimation procedures using asymmetrical kernels when data are supported on $(0, \infty)$, the work on asymmetric kernel density and regression estimation with compact support has been limited. When a density has a compact support, motivated by the Bernstein polynomial approximation theorem [Karlin and Studden(1966)] in functional analysis, Chen proposed Beta kernel density estimators [Chen(1999)] and Beta kernel local smoothers for regression curves [Chen(2002)]. Beta kernels, like other asymmetric kernels, offer some unique benefits with regard to estimating densities with compact support. First, the beta kernel shape varies naturally based on the position where the density estimation is performed leading to an adaptive change in the amount of smoothing. This adaptivity does not require any explicit change in the smoothing bandwidth. Secondly, the fact that the Beta kernels share the same support as the density to be estimated ensures that no weight is allocated outside the data range. However, the focus of these initial works in the density estimation context were limited to the analysis of the biases, variances and mean squared errors of these estimators. To the best of our knowledge, there has been no investigations related to the almost sure consistency and asymptotic distribution of Beta kernel density estimators. Similarly, while the asymptotic bias and variance of a Beta kernel based regression function estimator was presented in [Chen(2000a), Chen(2002)], there have been no published results related to their almost sure consistency and asymptotic distributions.

One might consider the possibility of developing a similar "relative efficiency" notion for the asymmetric kernel smoothing, similar to the one defined for the classical kernel case. However, this possibility is discouraged by the very different smoothing nature of symmetric kernels and asymmetric kernels. In symmetric kernel smoothing, the bandwidth can be viewed as the scale parameter of the kernel density, that is, the smoothness is controlled by the scale parameter. However, in asymmetric kernel smoothing, the bandwidth plays the

same role as the shape parameter of the kernel density. We cannot separate the bandwidth and the asymmetric kernel function in asymmetric smoothing. Therefore, comparison of different kernel smoothing can only be made by direct comparison of their MSEs or MISEs using their own optimal bandwidths, calculated based on a data-driven or a theoretical approach.

1.5 Objective and Contributions

The aim of this dissertation is to address two fundamental questions related to the use of asymmetric kernels in kernel based nonparametric methods.

Question 1: *What are the large sample properties (including consistency and asymptotic distribution) of Beta kernel methods for both density and regression function estimation? How do these properties compare with other symmetric and asymmetric kernels? How can we implement bandwidth selection within this methodology?*

Question 2: *What are the large sample properties (including consistency and asymptotic distribution) of local linear regression with Beta kernels? How do these properties compare with other kernel based local linear regression function estimates? How can we implement bandwidth selection within this methodology?*

In Chapter 2, we attempt to address Question 1 by exploring the large sample properties of the Beta kernel density and regression function estimators. For the first time, the asymptotic conditional bias and variance are derived, as well as its uniform almost sure consistency and asymptotic normality. Finally, some implementable bandwidth selection methodologies based on least square cross validation (LSCV) and generalized cross validation (GCV) are provided and tested using both simulation studies and a real data example. The usefulness of the Beta kernel estimation procedure is illustrated by comparing it with the Nadaraya-Watson (N-W)estimator and the local linear smoother. Results indicate that the Beta kernel estimator consistently outperforms the N-W estimator for all sample sizes

and in most cases, is comparable to the local linear estimator with normal kernel.

In Chapter 3, we address Question 2. Here, the asymptotic conditional bias and variance are derived for local linear regression with Beta kernels. Then, we prove the almost sure consistency and asymptotic normality of the Beta kernel based local linear smoother. The chapter concludes with numerical results based on a simulation study as well as real data. Once again, least squares cross validation and generalized cross validation methods are used to compute the optimal bandwidths. As expected, the N-W regression estimator shows more variability relative to the other schemes considered in this work. Both the Beta kernel and Normal kernel based local linear regression estimators behave similarly with the Beta kernel based local linear regression estimator able to better capture the data structure at the boundaries.

Chapter 2

Beta Kernel Regression

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population X with compact support. Without loss of generality, the compact support is assumed to be $[0, 1]$. Let $K_{p,q}$ be the density function of a Beta(p, q) random variable. For any fixed $x \in [0, 1]$, the Beta kernel estimator for the density function f of X proposed by [Chen(1999)] is defined as

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x/h_n+1,(1-x)/h_n+1}(X_i) = \frac{\sum_{i=1}^{n} X_i^{x/h_n}(1-X_i)^{(1-x)/h_n}}{nB(x/h_n+1,(1-x)/h_n+1)}, \tag{2.1}$$

where, h_n is a sequence smoothing parameters satisfying the conditions that $h_n \to 0$, $nh_n \to \infty$ as $n \to \infty$; $B(p, q)$ is the Beta function. For the sake of simplicity, the subscript n will be suppressed from h_n in the following discussion. [Chen(1999)] showed that

$$E\hat{f}_n(x) = f(x) + h[f'(x)(1-2x) + x(1-x)f''(x)/2] + o(h). \tag{2.2}$$

Here and in the sequel, for any function $g(x)$, $g'(x)$, $g''(x)$ denote the first and second derivatives of $g(x)$ with respect to x, respectively. So the bias of $\hat{f}_n(x)$ is $O(h)$ for all $x \in [0, 1]$, and moreover, the density estimator does not suffer from the boundary problem. A less desirable feature of the above defined $\hat{f}_n(x)$ is the presence of f' in the bias. This is

due to the fact that x is the mode and not the mean of the Beta$(x/h+1,(1-x)/h+1)$ distribution. [Chen(1999)] proposed the use of a modified Beta kernel density estimator that eliminates the f' term from the bias in the interior. While the f' term still forms a part of the bias in small areas near the boundaries, the f'' terms disappear at these areas. More properties of $\hat{f}_n(x)$ and its modified version, including the expressions of their variances, mean squared errors (MSE) and mean integrated squared errors, can also be found in [Chen(1999)]. In this paper, we focus our attention on the Beta kernel density estimator defined in (2.1). Although the presence of f' in the bias will be carried over to the regression estimation, the derivation of theoretical properties of the proposed estimation procedure will be much easier and concise. Moreover, the arguments we develop for the regression estimator based on $K_{x/h+1,(1-x)/h+1}$ can be easily adapted to accommodate the one constructed from the modified Beta kernel in [Chen(1999)].

It is very interesting to uncover the close relationship between the Beta kernel and the normal kernel used in the classical kernel density estimation as follows. For fixed $x \in (0,1)$, if we let R_h denote the random variable following a Beta distribution, i.e., Beta$(x/h+1,(1-x)/h+1)$, then we can show that $(hx(1-x))^{-1/2}(R_h-x) \Longrightarrow N(0,1)$ as $h \to 0$. This implies that asymptotically, the Beta kernel behaves like the normal kernel in which x-dependent bandwidth is used for each point x at which $f(x)$ is estimated.

To usher in the Beta kernel regression estimator, we assume a scalar response Y and a one-dimensional covariate X obeys the regression model $Y = m(X) + \varepsilon$, where ε accounts for the random error with usual assumptions $E(\varepsilon|X = x) = 0$ and $\sigma^2(x) := E(\varepsilon^2|X = x) > 0$, for almost all x. Analogous to the N-W kernel regression estimator, the Beta kernel regression estimator of $m(x)$ when the covariate X has compact support is defined as

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)Y_i}{\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)}. \tag{2.3}$$

From the definition of K_{h_n}, one can easily derive a much simpler expression for $\hat{m}_n(x)$ as

follows

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n X_i^{x/h_n}(1-X_i)^{(1-x)/h_n}Y_i}{\sum_{i=1}^n X_i^{x/h_n}(1-X_i)^{(1-x)/h_n}}.$$

This formula is mainly useful for the computation of $\hat{m}_n$ while (3.2) is more convenient for theoretical development.

2.1 Large Sample Results of $\hat{m}_n(x)$

We start with the asymptotic expansions of the conditional bias, variance, hence the MSE of $\hat{m}_n(x)$ defined in (3.2). Then a direct application of Lindeberg-Feller central limit theorem will lead to the asymptotic normality of $\hat{m}_n(x)$. The asymptotic normality of $\hat{f}_n(x)$ is also derived. Finally, uniform almost sure convergence results of $\hat{f}_n(x)$ and $\hat{m}_n(x)$ over $[0,1]$ are developed by using the Borel-Cantelli lemma after verifying the Cramér condition for the Beta kernel function. The following is a list of technical assumptions used for deriving these results.

(A1). The second order derivatives of $f(x)$ is continuous and bounded on $[0,1]$.

(A2). $E(\varepsilon|X) = 0$, and the second order derivatives of $f(x)m(x), f(x)m^2(x)$ are continuous and bounded on $[0,1]$.

(A3). The second order derivative of $\sigma^2(x) = E(\varepsilon^2|X=x)$ and $f(x)\sigma^2(x)$ with respect to $x \in [0,1]$ are continuous and bounded.

(A4). For some $\delta > 0$, the second order derivative of $E(|\varepsilon|^{2+\delta}|X=x)$ is continuous and bounded in $x \in [0,1]$.

(A5). $h \to 0$, $n\sqrt{h} \to \infty$ as $n \to \infty$.

Condition (A1) on $f(x)$ is also adopted by [Chen(1999)] implicitly when deriving MSE of $\hat{f}_n(x)$. Similar to (A1), condition (A2) is needed to control the higher order term in the asymptotic expansions of MSE for $\hat{m}_n(x)$. Condition (A3) is required for dealing with the

large sample argument pertaining to the random error, and is not needed if one is willing to assume the homoscedasticity. Condition (A4) is needed in proving the asymptotic normality of the proposed estimators, while (A5), similar to its classical kernel context, is a minimal condition needed for the smoothing parameter. Additional assumptions on h as needed are stated in various theorems presented below.

2.1.1 Bias and Variance

Let

$$b(x) = (1-2x)m'(x) + \frac{1}{2}x(1-x)m''(x) + \frac{x(1-x)m'(x)f'(x)}{f(x)},$$
$$v(x) = \frac{\sigma^2(x)}{2f(x)\sqrt{\pi x(1-x)}}, \tag{2.4}$$

and $\mathbf{X} := \{X_1, X_2, \ldots, X_n\}$. The following theorem presents the conditional biases and variances of $\hat{m}_n(x)$ defined in (3.2).

Theorem 2.1.1 *Suppose the assumptions (A1),(A2),(A3),and (A5) hold. Then, for any $x \in [0,1]$ with $f(x) > 0$,*

(i). For x/h and $(1-x)/h \to \infty$ (i.e., x is in the interior),

$$\text{Bias}(\hat{m}_n(x)|\mathbf{X}) = hb(x) + o_p(h) + O_p(\frac{1}{\sqrt{n\sqrt{h}}}), \tag{2.5}$$
$$\text{Var}(\hat{m}_n(x)|\mathbf{X}) = \frac{v(x)}{n\sqrt{h}} + o_p\Big(\frac{1}{n\sqrt{h}}\Big). \tag{2.6}$$

(ii). For x/h or $(1-x)/h \to K$, a positive constant (i.e., x is in the boundary),

$$\text{Bias}(\hat{m}_n(x)|\mathbf{X}) = hb(x) + o_p(h) + O_p(\frac{1}{\sqrt{nh}}), \tag{2.7}$$
$$\text{Var}(\hat{m}_n(x)|\mathbf{X}) = \frac{\Gamma(2K+1)\sigma^2(x)}{nh2^{1+2K}\Gamma^2(K+1)} + o_p(\frac{1}{nh}) \tag{2.8}$$

Thus the conditional MSE of $\hat{m}_n(x)$ *has the asymptotic expansion*

$$MSE(\hat{m}_n(x)|\mathbf{X}) = h^2 b^2(x) + \frac{v(x)}{n\sqrt{h}} + o_p(h^2) + o_p\Big(\frac{1}{n\sqrt{h}}\Big) + o_p\left(\frac{h^{-3/2}}{\sqrt{n}}\right)$$

when x/h *and* $(1-x)/h \to \infty$, *and*

$$MSE(\hat{m}_n(x)|\mathbf{X}) = h^2 b^2(x) + \frac{\Gamma(2K+1)\sigma^2(x)}{nh2^{1+2K}\Gamma^2(K+1)} + o_p(h^2) + o_p\left(\frac{1}{nh}\right)$$

when x/h *or* $(1-x)/h \to K$, *a positive constant.*

Theorem 2.1.1 indicates that both the conditional biases and variances of $\hat{m}_n(x)$ have a higher order at boundary points than at interior points. This is in contrast to Gamma kernel regression, see [Shi and Song(2015)], where the conditional biases are the same within the interior and at the boundary point $(x = 0)$. Similar to the N-W kernel regression case, one can choose the optimal smoothing parameter h by minimizing the leading term in the conditional MSE of $\hat{m}_n$ with respect to h.

2.1.2 Asymptotic Normality

In order to prove to the asymptotic normality of $\hat{m}_n(x)$, we first establish the asymptotic normality of $\hat{f}_n(x)$ along with its proof in Section 4.

Theorem 2.1.2 *Suppose the assumptions (A1) and (A5) hold. Then for any* $x \in (0,1)$ *with* $f(x) > 0$,

$$\Big(\frac{f(x)}{2n\sqrt{\pi x(1-x)h}}\Big)^{-1/2}\Big[\hat{f}_n(x) - f(x) - h[(1-2x)f'(x) + \frac{1}{2}x(1-x)f''(x)] + o(h)\Big] \to_d N(0,1).$$

The asymptotic normality of $\hat{f}_n(x)$ implies that $\hat{f}_n(x)$ converges to $f(x)$ in probability. Hence $1/\hat{f}_n(x)$ converges to $1/f(x)$ in probability, whenever $f(x) > 0$. This result is used in the proof of the asymptotic normality of $\hat{m}_n(x)$, which is stated in the next theorem.

Theorem 2.1.3 *Suppose the assumptions in Theorem 2.1.1 hold. Then, for any* $x \in (0,1)$ *with* $f(x) > 0$,

$$\Big(\frac{v(x)}{n\sqrt{h}}\Big)^{-1/2}\Big[\hat{m}_n(x) - m(x) - hb(x) + o_p(h)\Big] \to_d N(0,1),$$

where, $b(x)$ *and* $v(x)$ *are defined in (2.4).*

It is noted that there is a non-negligible asymptotic bias appearing in the above results, a characteristic shared with the N-W kernel density and regression estimators. These biases can be eliminated by under-smoothing which, in the current set up, is to select a proper h such that $nh^{5/2} \to 0$ for $0 < x < 1$, without violating conditions $h \to 0, n\sqrt{h} \to \infty$. The large sample confidence intervals for $m(x)$ thus can be constructed with the help of Theorem 2.1.3.

2.1.3 Uniform Almost Sure Consistency

In this section we develop a uniform almost sure convergence result for $\hat{m}_n(x)$ over an arbitrary closed sub-interval of $(0,1)$. To do this we apply Borel-Cantelli lemma and the Bernstein inequality, after verifying the Cramér condition: for some $k \geq 2$, $c > 0$, and h small enough,

$$E|K_{x/h+1,(1-x)/h+1}(X)|^k \leq k!\Big(\frac{c}{n\sqrt{h}}\Big)^{k-2} EK^2_{x/h+1,(1-x)/h+1}(X), \quad 0 < x < 1. \tag{2.9}$$

The following two theorems give the uniform almost sure convergence of $\hat{f}_n$ to f and $\hat{m}_n$ to m over bounded sub-intervals of $(0,1)$.

Theorem 2.1.4 *In addition to (A1) and (A5), assume that* $\log n/n\sqrt{h} \to 0$. *Then for any constants* a *and* b *such that* $0 < a < b < 1$,

$$\sup_{x\in[a,b]}\Big|\hat{f}_n(x) - f(x)\Big| = O(h) + o\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big), \quad a.s.$$

Theorem 2.1.5 *In addition to (A1) to (A5), assume that* $\log n/n\sqrt{h} \to 0$. *Then for any constants* a *and* b *such that* $0 < a < b < 1$,

$$\sup_{x\in[a,b]} \Big|\hat{m}_n(x) - m(x)\Big| = O(h) + o\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big), \quad a.s.$$

2.2 Selection of Smoothing Parameter

In this section, we propose several smoothing parameter selection procedures for implementing the Beta kernel technique. We seek to employ a data driven selection strategy which balances the bias and variance. We start with the least square cross validation (LSCV) procedure, and its extension, k-fold LSCV for the density estimation problem. Secondly, we propose the smoothing parameter selection procedures in the nonparametric regression setup. Finally, we present the generalized cross validation (GCV) method and its application to Beta kernel regression problem. These procedures are analogous to the commonly used data-driven procedures used in the N-W kernel regression estimation context. The theoretical properties, such as the consistency of these smoothing parameter selectors to some "optimal" smoothing parameter deserves an independent in-depth study.

2.2.1 Density Estimation: k-fold LSCV

The motivation of the LSCV comes from expanding the MISE of $\hat{f}$. Define

$$LSCV(h) = \int \hat{f}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i),$$

where $\hat{f}_{-i}(X_i)$ is the leave-one-out Gamma kernel density estimator for $f(X_i)$ without using the i-th observation. Then the LSCV smoothing parameter is defined by $\hat{h}_{\text{LSCV}} =$

$\text{argmin}_h LSCV(h)$. For the beta kernel density estimator (2.1),

$$LSCV(h) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int \frac{1}{B^2(x/h+1,(1-x)/h+1)}(X_iX_j)^{x/h}\Big((1-X_i)(1-X_j)\Big)^{(1-x)/h}dx$$
$$-\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{B(X_i/h+1,(1-X_i)/h+1)}X_j^{X_i/h}(1-X_j)^{(1-X_i)/h}.$$

The integration of the first term of LSCV is not trivial due to the presence of the square of the beta function in the denominator that depends on x. One approach is to use a brute force method where LSCV is calculated via numerical integration for a range of h values and the minimum is identified. However, by using Theorem 7 in Cerone(2007), we can upper bound the inverse of the square of the beta function by a constant K. This allows us to upper bound LSCV as follows:

$$LSCV(h) \leq \frac{K}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{h\Big[(X_iX_j)^{1/h}-((1-X_i)(1-X_j))^{1/h}\Big]}{\log\left(\frac{X_iX_j}{(1-X_i)(1-X_j)}\right)}$$
$$-\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{B(X_i/h+1,(1-X_i)/h+1)}X_j^{X_i/h}(1-X_j)^{(1-X_i)/h}.$$

In essence, instead of minimizing LSCV directly, one can determine the h that minimizes the upper bound of LSCV. This approach, while sub optimal, provides computational complexity reduction compared to a brute force search method.

An extension of the above leave-one-out LSCV is the k-fold LSCV procedure. Here, the data is first split into k roughly equal-sized parts; then for the l-th part, the Beta kernel density estimator and its associated error is calculated based on the other $k-1$ parts of the data. This procedure is repeated for $l = 1, 2, \ldots, k$ and the combination of the k estimation errors is minimized. In particular, for our current setup, the k-fold $LSCV$ has a similar

structure as the leave-one-out LSCV except for the second term now defined as

$$\frac{2}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-n_i)}\sum_{j\notin D(i)}\frac{1}{B(X_i/h+1,(1-X_i)/h+1)}X_j^{X_i/h}(1-X_j)^{(1-X_i)/h}\Big],$$

where, $D(i)$ is the set of indices of the data part including X_i, and n_i is the size of $D(i)$. The k-fold LSCV will reduce to the leave-one-out LSCV when $k = n$.

2.2.2 Beta Kernel Regression: k-fold LSCV

The basic idea of LSCV in regression setup is to select the smoothing parameter by minimizing prediction error. For this purpose, let $\hat{m}_{D/D(i)}(X_i)$ be the Beta kernel estimator of $m(x)$ at $x = X_i$ of the same type as $\hat{m}_n(x)$ except that it is computed without using the data parts including the i-th observation (X_i, Y_i), where $D = \{1, 2, \ldots, n\}$. The LSCV smoothing parameter $\hat{h}_{\text{LSCV}}$ is the value of h that minimizes the LSCV criterion

$$\text{LSCV}(h) = \sum_{i=1}^{n}[Y_i - \hat{m}_{D/D(i)}(X_i)]^2 = \sum_{i=1}^{n}\Big[Y_i - \frac{\sum_{j\notin D(i)} X_j^{X_i/h}(1-X_j)^{(1-X_i)/h}Y_j}{\sum_{j\notin D(i)} X_j^{X_i/h}(1-X_j)^{(1-X_i)/h}}\Big]^2.$$

The independence between (X_i, Y_i) and $\hat{m}_{D/D(i)}(X_i)$ indicates that $LSCV(h)$ will give an accurate assessment of how well the estimator $\hat{m}_n(x)$ will predict future observations.

2.2.3 Beta Kernel Regression: Generalized Cross Validation (GCV)

The GCV procedure from the N-W kernel regression can also be adapted to the Beta kernel regression setup. Define

$$w_{ij} = \frac{X_j^{X_i/h}(1-X_j)^{(1-X_i)/h}}{\sum_{k=1}^{n} X_k^{X_i/h}(1-X_k)^{(1-X_i)/h}}, \quad i, j = 1, 2, \ldots, n.$$

Then the GCV smoothing parameter $\hat{h}_{\text{GCV}}$ is the value of h that minimizes the GCV criterion defined as

$$\text{GCV}(h) = n\Big[n - \sum_{i=1}^{n} w_{ii}\Big]^{-2} \sum_{i=1}^{n} \Big[Y_i - \sum_{j=1}^{n} w_{ij} Y_j\Big]^2.$$

In addition to the methods described above, there are many other approaches (e.g., AIC or BIC type criteria-based) that can be used to select the smoothing parameter. It is important to remember that there is no one smoothing parameter selection approach that is uniformly superior to others in the sense of resulting in always the smallest MSE. A certain level of experimentation to identify the best smoothing parameter for a given problem is usually recommended. In the next section, we illustrate this point through a simulation study.

2.3 Numerical Study

In this section, we evaluate the finite sample performance of the proposed beta kernel regression estimator using both simulation and real data analysis.

2.3.1 Simulation Study

For the simulated data, the underlying density function of the design variable is chosen to be uniform between 0 and 1. A quadratic regression function is selected for generating the response variable. That is, we set $m(x) = 10(x-0.5)^2$ and the response $Y = m(x)+\varepsilon$ where ε values are drawn from a $N(0,1)$ distribution. For one set of random data, Figure 1 presents the estimated $\hat{m}(x)$ based on bandwidth selected using 5-fold LSCV. For comparison, along with beta kernel regression estimator, we present the N-W estimator, and the local linear estimator with the normal kernel. The true regression curve is also plotted for reference. Results based on sample sizes of 100 and 200 are presented in Figure 2.1, respectively. For the same data set, the regression estimate based on bandwidth derived from the GCV

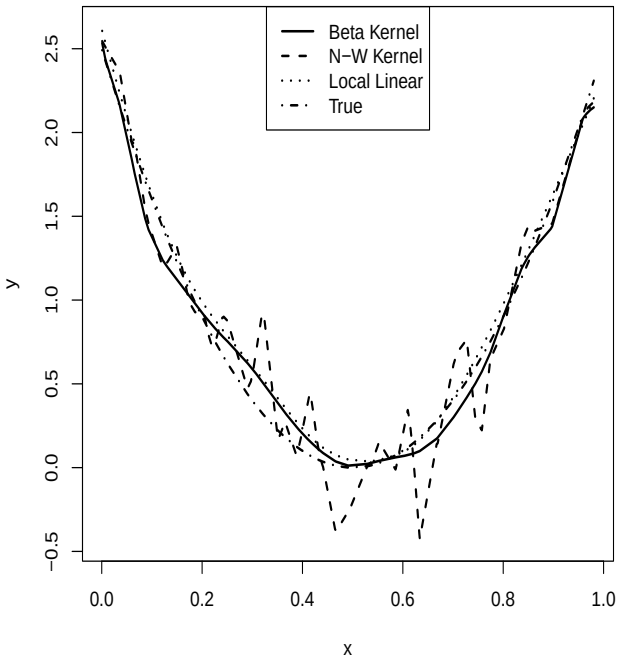
procedure is similar to the result using LSCV criterion and are presented in Figure 2.2. The optimal bandwidth values and the resulting LSCV and GCV metrics for the analyzed data set are provided in Table 1.

From Figures 2.1 and 2.2 and Table 2.1, it is evident that the Beta kernel estimator consistently outperforms the N-W estimator for all sample sizes and in most cases, is comparable to the local linear estimator with normal kernel. Of course, the performance gain offered by the local method is obtained at the cost of increased computational complexity. Due to the randomness of generated data, we also observe occasional cases where the Beta kernel regression outperforms the local linear method.

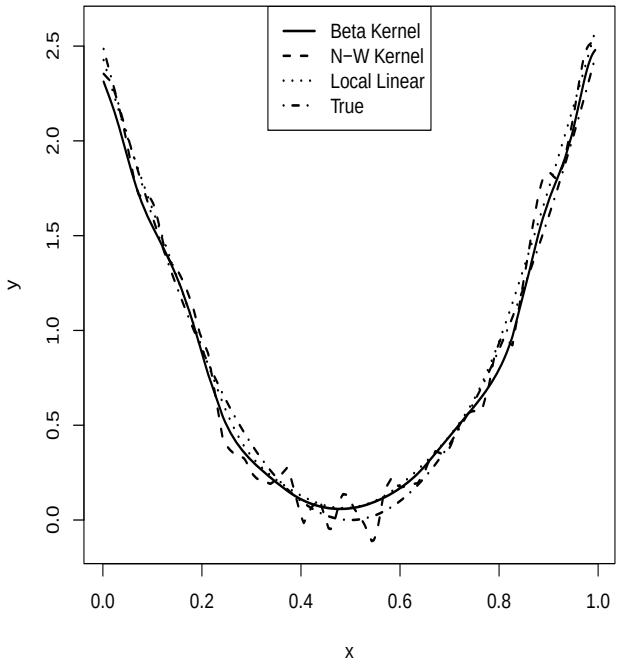
		LSCV(5-fold)		GCV	
	Kernel Type	h_{LSCV}	$LSCV(h_{LSCV})$	h_{GCV}	$GCV(h_{GCV})$
$n = 100$	Beta Kernel	0.06472	0.02944	0.05477	0.02600
	Nadaraya-Watson	0.11447	0.04032	0.10452	0.04231
	Normal Local Linear	0.14432	0.03775	0.12940	0.03260
$n = 200$	Beta Kernel	0.02990	0.03750	0.02990	0.03750
	Nadaraya-Watson	0.06970	0.05365	0.06970	0.05365
	Normal Local Linear	0.09955	0.03721	0.09457	0.03630

Table 2.1: LSCV and GCV values with optimal bandwidths

In Figure 2.3, we plot the mean squared errors (MSE) as a function of σ_ε^2, the variance of the noise term ε. From the figure it is clear that as the noise variance increases, the noisy measurements reduce the estimator performance. For low variance values (below 3 in this case), the N-W estimator has the worst performance relative to the beta kernel and local linear estimate. Beyond a certain noise variance level (above 3.5 in this case), the local linear estimate tends to perform worse. This could be attributed to the significant variability of measured response relative to the true regression function. The tendency of the local methods to accommodate all local variations in data even though they are noise induced, result in a poor overall performance at these high noise levels.

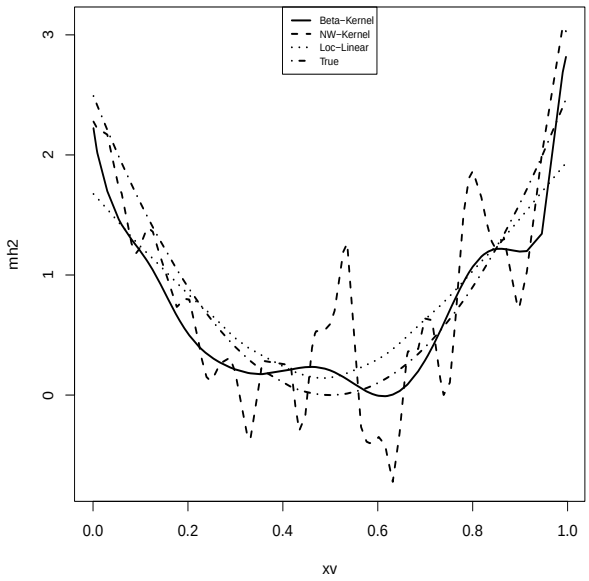


(a) $n = 100$

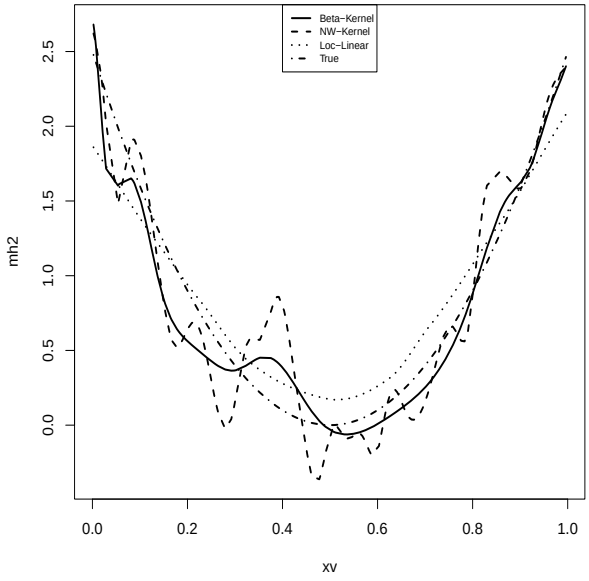


(b) $n = 200$

Figure 2.1: Comparison of Various Kernel Regression Estimators (5-fold LSCV)



(a) $n = 100$



(b) $n = 200$

Figure 2.2: Comparison of Various Kernel Regression Estimators (GCV)

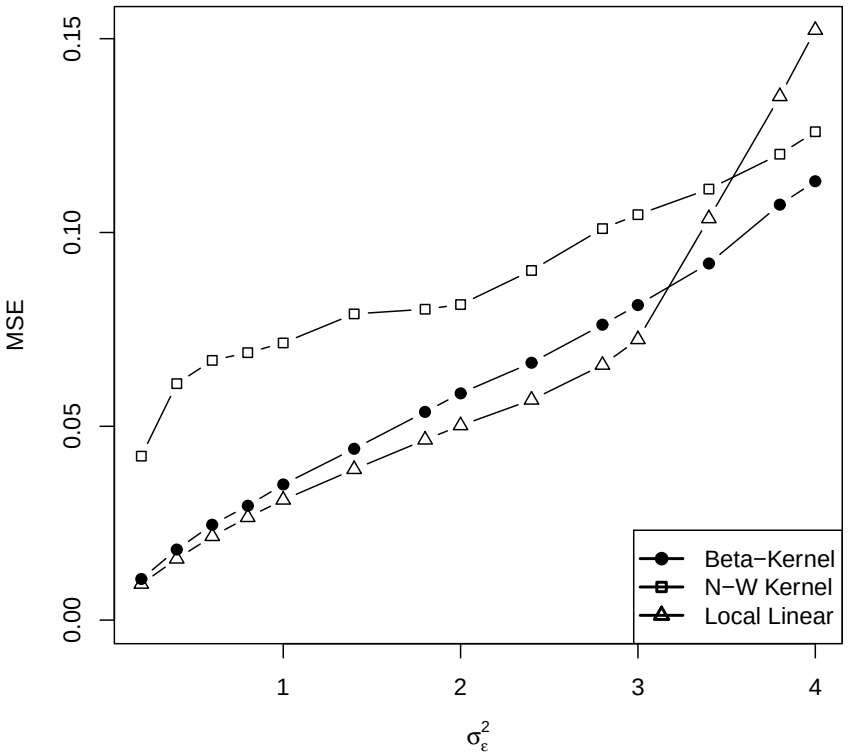


Figure 2.3: MSEs as a function of noise variance

2.3.2 Real Data Example

We now apply the Beta kernel regression estimation procedure for the data set from Azzalini and Bowman (1990) on the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. The data consist of 299 pairs of measurements on the waiting time X to the next eruption, and the eruption time Y in minutes, which were collected continuously from August 1st until August 15th, 1985. We shall use the nonparametric regression procedure developed in this paper to investigate the relationship between Y and X. To prevent the computation of the Beta kernel regression estimator from explosion, the waiting time X is transformed to $T = (X - \min(X) + 1)/(\max(X) - \min(X) + 2)$. This transformation also ensures that the range of T is between 0 and 1.

The scatter plot in Figures 2.4 and 2.5 shows the data structure of Y against T. The bandwidth h is chosen based on GCV and 5-fold LSCV criterion for all kernel regression

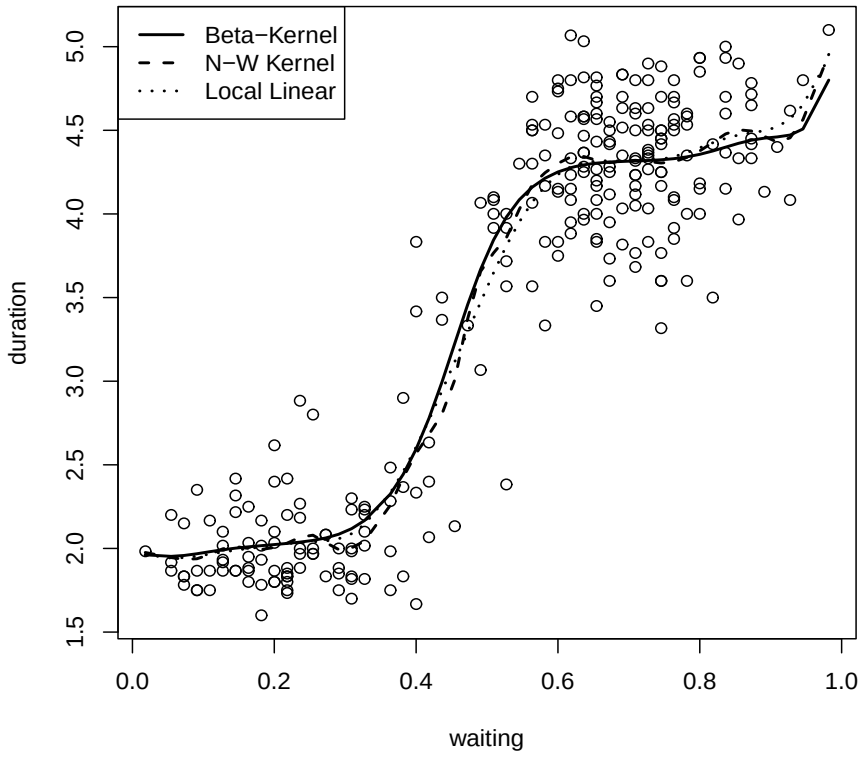


Figure 2.4: Regression for Geyser data based on GCV

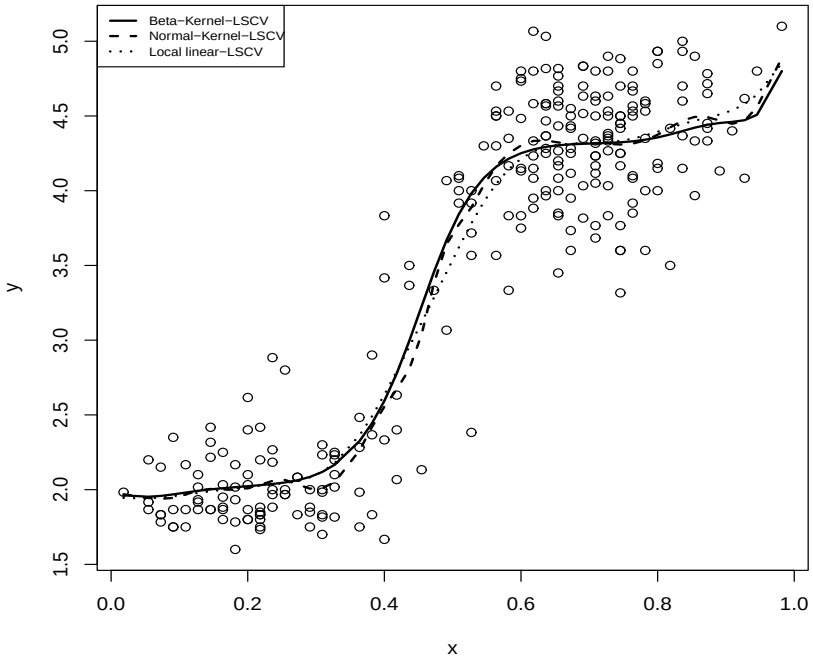


Figure 2.5: Regression for Geyser data based on 5-fold LSCV

estimation techniques. The Beta kernel regression curve is imposed on the scatter plot in Figures 2.4 and 2.5 with solid line. For comparison purpose, the N-W kernel and local linear regression curves are also plotted. All three estimators capture the main characteristic of the data structure, but the Beta kernel regression estimate appears less variable than the other two. Additionally the local linear method offers comparable regression fits as seen in Figures 2.4 and 2.5.

2.4 Proofs of the Main Results

This section contains the proofs of all the large sample results presented in Section 3. Beta density function and its moments will be repeatedly referred to in the following proofs. For convenience, we list all the needed results here. Density function of a Beta distribution with shape parameters p and q is

$$g(u;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} u^{p-1}(1-u)^{q-1} \quad u \in [0,1].$$

Its mean μ and variance τ^2 are $\mu = \frac{p}{p+q}$, and $\tau^2 = \frac{pq}{(p+q)^2(p+q+1)}$.

Let $p = x/h + 1, q = (1-x)/h + 1, x \in [0,1]$. The following lemma on the inverse beta distribution is crucial for the subsequent arguments.

Lemma 2.4.1 *Let $l(u)$ be a function such that the second order derivative of $l(u)$ is continuous and bounded on $[0,1]$. Then, for all $x \in [0,1]$,*

$$\int_0^1 g(u;p,q)l(u)du = l(x) + \left[(1-2x)l'(x) + \frac{1}{2}x(1-x)l''(x))\right] h + o(h). \tag{2.10}$$

PROOF. Fix an $x \in (0,1)$. Note that $\mu = (x/h+1)/(1/h+2)$. Taylor expansion of $l(\mu)$ around x up to second order yields

$$l(\mu) = l(x) + (1-2x)hl'(x) + \frac{1}{2}(1-2x)^2h^2l''(\xi), \tag{2.11}$$

where ξ is some value between $x+\mu$ and x. Recall μ is the mean of $g(u;p,q)$. The proof follows the classic procedure involving Taylor expansion of $l(u)$ around μ up to the second order followed by taking expectation. That is,

$$\begin{aligned}\int_0^1 l(u)g(u,p,q)du &= l(\mu)+\frac{1}{2}l''(\mu)\int_0^1(u-\mu)^2 g(u;p,q)du \\ &\quad +\frac{1}{2}\int_0^1(u-\mu)^2 g(u;p,q)[l''(\tilde{u})-l''(\mu)]du \end{aligned} \tag{2.12}$$

for some $\tilde{u}$ between u and μ. From (2.11) and the continuity of l'', we can verify that the two leading terms on the right hand side of (2.12) match the expansion in the lemma. Specifically, the first term follows from (2.11) and the second term corresponds to the variance of the beta distribution. So it suffices to show that the third term on the right hand side of (2.12) is of the order $o(h)$. This can be shown by using the boundedness of l'', replacing $g(u,p,q)$ by its modal value, and using Stirling approximation to bound the third term. □

The following decomposition of $\hat{m}_n(x)$ will be used repeatedly in the following proofs below.

$$\hat{m}_n(x)-m(x)=\frac{B_n(x)+V_n(x)}{f(x)}+\Big[\frac{1}{\hat{f}_n(x)}-\frac{1}{f(x)}\Big][B_n(x)+V_n(x)], \tag{2.13}$$

where

$$B_n(x)=\frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)[m(X_i)-m(x)],\quad V_n(x)=\frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i,$$

with $K_{x/h+1,(1-x)/h+1}(X_i)$ defined in (2.1). Now we are ready to prove Theorem 2.1.1.

Proof of Theorem 2.1.1 From the assumption (A2), the conditional bias of $\hat{m}_n(x)$ equals $E[\hat{m}_n(x)|\mathbf{X}]-m(x)=B_n(x)/\hat{f}_n(x)$. From [Chen(1999)], it is known that $\hat{f}_n(x)=f(x)+o_p(1)$. In the following, we shall develop asymptotic expansions of the expectation and variance of $B_n(x)$.

First, let's consider the expectation of $B_n(x)$ which can be written as

$$EB_n(x) = EK_{x/h+1,(1-x)/h+1}(X)m(X) - m(x)EK_{x/h+1,(1-x)/h+1}(X).$$

Applying Lemma 3.3.1 with $l = H = mf$ and $l = f$, by assumption (A1) and (A2),

$$EK_{x/h+1,(1-x)/h+1}(X)m(X) = H(x) + ((1-2x)H'(x) + \frac{1}{2}x(1-x)H''(x))h + o(h),$$
$$m(x)EK_{x/h+1,(1-x)/h+1}(X) = m(x)\left[f(x) + ((1-2x)f'(x) + \frac{1}{2}x(1-x)f''(x))h + o(h)\right].$$

After a few simplification steps, we have

$$EB_n(x) = \left[(1-2x)m'(x)f(x) + \frac{1}{2}x(1-x)(m''(x)f(x) + 2m'(x)f'(x))\right]h + o(h) \quad (2.14)$$

Direct calculation shows that $(1-2x)m'(x)f(x) + \frac{1}{2}x(1-x)(m''(x)f(x) + 2m'(x)f'(x)) = b(x)f(x)$, where $b(x)$ is defined in (2.4).

Now consider the variance of $B_n(x)$. Since the variance of a random variable is bounded above by its second moment, so

$$\mathrm{Var}(B_n(x)) \leq \frac{1}{n}E\left[K_{x/h+1,(1-x)/h+1}(X)[m(X) - m(x)]\right]^2 \quad (2.15)$$

The second moment can be broken down into three parts namely,

$$\begin{aligned} & E\left[K_{x/h+1,(1-x)/h+1}(X)[m(X) - m(x)]\right]^2 \\ = & \int_0^1 K^2_{x/h+1,(1-x)/h+1}(X)m^2(u)f(u)du + m^2(x)\int_0^1 K^2_{x/h+1,(1-x)/h+1}(X)f(u)du \\ & -2m(x)\int_0^1 K^2_{x/h+1,(1-x)/h+1}(X)m(u)f(u)du. \end{aligned} \quad (2.16)$$

By substituting the exact expression for the beta kernel and performing a simple algebraic

trick to isolate the constant terms, we can rewrite the integrals as

$$\begin{aligned}
&\int_0^1 K^2_{x/h+1,(1-x)/h+1}(X)m^2(u)f(u)du + m^2(x)\int_0^1 K^2_{x/h+1,(1-x)/h+1}(X)f(u)du \\
&-2m(x)\int_0^1 K^2_{x/h+1,(1-x)/h+1}(X)m(u)f(u)du \\
&= A_h(x)\left[E(m^2(\gamma_x)f(\gamma_x)) + m^2(x)E(f(\gamma_x)) - 2m(x)E(m(\gamma_x)f(\gamma_x))\right] \qquad (2.17)
\end{aligned}$$

where, γ_x is a Beta$(2x/h+1, 2(1-x)/h+1)$ random variable and

$$A_h(x) = \frac{B(2x/h+1, 2(1-x)/h+1)}{B^2(x/h+1,(1-x)/h+1} \qquad (2.18)$$

Applying a slightly modified version of Lemma 3.3.1 to reflect the difference in the underlying Beta distribution shape parameters and using $l = m^2 f$,$l = f$ and $l = mf$ for the three terms in the second moment, respectively, we can write down the complete expression for the right hand side of (2.17). The tedious algebra leads to the conclusion that the asymptotics of the second moment depends on $A_h(x)$. From [Chen(1999)], we know that

$$A_h(x) \approx \begin{cases} \frac{1}{2\sqrt{\pi}}(x(1-x))^{-1/2}h^{-1/2}, & \text{if } x/h \text{ and } (1-x)/h \to \infty \text{ (interior);} \\ \frac{\Gamma(2K+1)}{2^{1+2K}\Gamma^2(K+1)}h^{-1}, & \text{if } x/h \to K \text{ or } (1-x)/h \to K \text{ (boundary).} \end{cases} \qquad (2.19)$$

Therefore, the $Var(B_n(x)) = O(\frac{1}{n\sqrt{h}})$ if x is an interior point and is $O(\frac{1}{nh})$ if x is in the boundary. Since, $B_n(x) = EB_n(x) + O_p(\sqrt{Var(B_n(x))})$, we can write down $B_n(x) = b(x)f(x)h + O_p(\frac{1}{\sqrt{n\sqrt{h}}})$ if x is an interior point and $B_n(x) = b(x)f(x)h + O_p(\frac{1}{\sqrt{nh}})$ if x is in the boundary. Combining this with $\hat{f}_n(x) = f(x) + o_p(1)$ implies (3.5) and (3.7).

Next, consider the conditional variance $\text{Var}[\hat{m}_n(x)|\mathbf{X}]$. It is easily seen that

$$\text{Var}[\hat{m}_n(x)|\mathbf{X}] = \frac{1}{n^2\hat{f}_n^2(x)}\sum_{i=1}^n K^2_{x/h+1,(1-x)/h+1}\sigma^2(X_i).$$

Applying Lemma 3.3.1 with $l = f\sigma^2$ with a modification to reflect the change in underlying

Beta distribution and by (A3), one can show that, for $x \in [0,1]$

$$\frac{\Gamma^2(2+1/h)}{n\Gamma^2(x/h+1)\Gamma^2((1-x)/h+1)} E\left(X^{x/h}(1-X)^{(1-x)/h}\right)\sigma^2(X) = \frac{1}{n}A_h(x)\left(\sigma^2(x)f(x)+O(h)\right). \tag{2.20}$$

Therefore, from the approximation of $A_h(x)$ in (3.9) along with $\hat{f}_n(x) = f(x) + o_p(1)$, one can obtain (3.6) and (3.7). □

Proof of Theorem 2.1.2. Let $\xi_{in}(x) = n^{-1}[K_{x/h+1,(1-x)/h+1}(X_i) - EK_{x/h+1,(1-x)/h+1}(X)]$. Then

$$\hat{f}_n(x) = \sum_{i=1}^{n} \xi_{in}(x) + EK_{x/h+1,(1-x)/h+1}(X).$$

Since $EK_{x/h+1,(1-x)/h+1}(X) = f(x) + (1-2x)f'(x)h + \frac{1}{2}x(1-x)f''(x)h + o(h)$,

$$\hat{f}_n(x) - f(x) - (1-2x)f'(x)h - \frac{1}{2}x(1-x)f''(x)h + o(h) = \sum_{i=1}^{n}\xi_{in}(x).$$

Lindeberg-Feller CLT will be used to show the asymptotic normality of $\sum_{i=1}^{n}\xi_{in}(x)$. For any $a > 0, b > 0$ and $r > 1$, using the well known inequality $(a+b)^r \le 2^{r-1}(a^r + b^r)$, we have

$$E|\xi_{in}(x)|^{2+\delta} \le n^{-(2+\delta)}2^{1+\delta}[E(K_{x/h+1,(1-x)/h+1}(X))^{2+\delta} + (EK_{x/h+1,(1-x)/h+1}(X))^{2+\delta}].$$

Let $\theta = 2+\delta$. Then we can show that $E(K_{x/h+1,(1-x)/h+1}(X))^{\theta}$ equals

$$\frac{\Gamma^{\theta}(2+1/h)}{\Gamma^{\theta}(x/h+1)\Gamma^{\theta}((1-x)/h+1)}\int_0^1 u^{\theta x/h}(1-u)^{\theta(1-x)/h}f(u)du. \tag{2.21}$$

The integral in (2.21) is equal to $f(x) + o(1)$. The preceding term corresponds to,

$$\frac{\Gamma^{\theta}(2+1/h)}{\Gamma^{\theta}(x/h+1)\Gamma^{\theta}((1-x)/h+1)} = \frac{B(\theta x/h+1, \theta(1-x)/h+1)}{B^{\theta}(x/h+1,(1-x)/h+1)}$$

Using Stirling approximation for $B(a,b) \approx \frac{\sqrt{2\pi}a^{a-\frac{1}{2}}b^{b-\frac{1}{2}}}{(a+b)^{a+b-\frac{1}{2}}}$ followed by some routine calculations, we can show that

$$EK^{\theta}_{x/h+1,(1-x)/h+1}(X) = O\Big(h^{-(1-\theta)/2}\Big) = O\Big(h^{-(1+\delta)/2}\Big).$$

Let $v_n^2 = \mathrm{Var}\Big(\sum_{i=1}^n \xi_{in}(x)\Big) = \mathrm{Var}(\hat{f}_n(x))$. From [Chen(1999)], we can write down the expression for v_n^2 for any x in the interior of (0,1) as

$$v_n^2 = \frac{1}{2\sqrt{\pi n\sqrt{h}x(1-x)}} f(x) + o(\frac{1}{n\sqrt{h}}) \tag{2.22}$$

This fact together with $EK_{x/h+1,(1-x)/h+1}(X) = f(x) + o(1)$, imply

$$v_n^{-(2+\delta)} \sum_{i=1}^n E\xi_{in}^{2+\delta}(x) = nv_n^{-(2+\delta)} E\xi_{1n}^{2+\delta} = O\Big(\Big(\frac{1}{n\sqrt{h}}\Big)^{\delta/2}\Big),$$

which converges to 0, by assumption (A4). Hence the Lindeberg-Feller condition holds. This completes the proof of the Theorem 2.1.2. □

Proof of Theorem 2.1.3. Fix an $x \in (0,1)$. To show the asymptotic normality of $\hat{m}_n(x)$, again we use the decomposition (2.13). We shall first show that $V_n(x)$ is asymptotically normal. For this purpose, let $\eta_{in} = n^{-1}K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i$ so that $V_n(x) = \sum_{i=1}^n \eta_{in}$. Clearly, $E\eta_{in} = 0$. Assumption (A3) on $\sigma^2(x)$ and results from (2.20) leads to $E\eta_{in}^2 = (f(x)\sigma^2(x)/(2n^2\sqrt{x(1-x)h\pi})[1+o(1)]$. Therefore,

$$s_n^2 = \mathrm{Var}\Big(\sum_{i=1}^n \eta_{in}\Big) = nE\eta_{in}^2 = \frac{f(x)\sigma^2(x)}{2n\sqrt{x(1-x)h\pi}}[1+o(1)].$$

Using a similar argument as in dealing with $E|\xi_{in}(x)|^{2+\delta}$ in the proof of Theorem 2.1.2,

verify that for any $\delta > 0$,

$$E|\eta_{in}|^{2+\delta} = n^{-(2+\delta)} E K^{2+\delta}_{x/h+1,(1-x)/h+1}(x, X) E(|\varepsilon|^{2+\delta}|X = x) = O(n^{-(2+\delta)} h^{-(1+\delta)/2}).$$

Hence

$$s_n^{-(2+\delta)} \sum_{i=1}^{n} E|\eta_{in}|^{2+\delta} = O\Big(\Big(\frac{1}{n\sqrt{h}}\Big)^{\delta/2}\Big) = o(1).$$

Hence, by the Lindeberg-Feller CLT, $s_n^{-1} V_n(x) \to_d N(0, 1)$.

From the asymptotic results on $\hat{f}_n(x)$ and $V_n(x)$ in Theorem 2.1.2 and that of $B_n(x)$ discussed in proof of Theorem 2.1.1, we obtain

$$s_n^{-1}\Big[\frac{1}{\hat{f}_n(x)} - \frac{1}{f(x)}\Big][B_n(x) + V_n(x)] = o_p(1).$$

This, together with the result that $\sqrt{n\sqrt{h}} \cdot O_p\Big(1/\sqrt{\frac{n}{\sqrt{h}}}\Big) = o_p(1)$, implies

$$f(x) s_n^{-1}\Big(\hat{m}_n(x) - m(x) - b(x)h + o(h)\Big) = s_n^{-1} V_n(x) \to_d N(0, 1).$$

The proof is completed by noticing that $f(x) s_n^{-1} = \Big(v(x)/n\sqrt{h}\Big)^{-1/2}$. □

Proof of Theorem 2.1.4. Recall that $E\hat{f}_n(x) = \int_0^1 g(u; p, q) f(u) du$. By applying Lemma 3.3.1 with $l(u) = f(u)$, $k = 1$, and the boundedness of $xf''(x)$ on $[a, b]$, we obtain

$$E\hat{f}_n(x) - f(x) = O(h), \quad \text{for any } x \in [a, b].$$

Hence $\sup_{a \le x \le b} |E\hat{f}_n(x) - f(x)| = O(h)$. So, we only need to show that $\hat{f}_n(x) - E\hat{f}_n(x) = o(h^{-1/4}\sqrt{\log n}/\sqrt{n})$. For this purpose, let

$$\xi_{in}(x) = n^{-1}[K_{x/h+1,(1-x)/h+1}(X_i) - E K_{x/h+1,(1-x)/h+1}(X_i)].$$

Hence $\hat{f}_n(x) - E\hat{f}_n(x) = \sum_{i=1}^n \xi_{in}(x)$. In order to apply Bernstein inequality, we have to verify the Cramér condition for ξ_{in}, that is, we need to show that, for $k \geq 3$, $E|\xi_{1n}|^k \leq c_n^{k-2} k! E\xi_{1n}^2$ for some c_n only depending on n.

Note that $K_{x/h+1,(1-x)/h+1}(X)$ attains its maximum at the mode of the beta distribution, i.e., $X = \frac{x/h-1}{1/h-2}$. Therefore, $K_{x/h+1,(1-x)/h+1}(X)$ is bounded above by:

$$K_{x/h+1,(1-x)/h+1}(X) \leq \frac{\left(\frac{x/h-1}{1/h-2}\right)^{x/h}\left(1-\frac{x/h-1}{1/h-2}\right)^{(1-x)/h}}{B(x/h+1,(1-x)/h+1)} \tag{2.23}$$

Once again using the Stirling approximation for the Beta function and some simple algebraic computations, we can simplify the upper bound as follows:

$$\frac{(1/h+2)^{3/2}}{(x/h+1)^{1/2}((1-x)/h+1)^{1/2}} = C^* \frac{1}{\sqrt{hx(1-x)}} \tag{2.24}$$

for some positive constant C^*. Therefore, for any $k \geq 3$, and h small enough,

$$\begin{aligned} E|\xi_{in}|^k &= n^{-k} E|K_{x/h+1,(1-x)/h+1}(X_i) - EK_{x/h+1,(1-x)/h+1}(X_i)|^k \\ &\leq \left(\frac{C^*}{n\sqrt{hx(1-x)}}\right)^{k-2} n^{-2} E|K_{x/h+1,(1-x)/h+1}(X_i) - EK_{x/h+1,(1-x)/h+1}(X_i)|^2 \\ &= \left(\frac{C^*}{n\sqrt{hx(1-x)}}\right)^{k-2} E\xi_{in}^2. \end{aligned}$$

With $v_n := \left(\sum_{i=1}^n E\xi_{in}^2\right)^{1/2}$, this immediately implies,

$$E|\xi_{in}|^k \leq k!\Big(\frac{C^*}{n\sqrt{hx(1-x)}}\Big)^{k-2} E\xi_{in}^2, \quad \forall\ 1 \leq i \leq n,$$

or

$$E\Big(\frac{\xi_{in}}{v_n}\Big)^k \leq k!\Big(\frac{C^*}{n\sqrt{hx(1-x)}v_n}\Big)^{k-2} E\Big[\frac{\xi_{in}}{v_n}\Big]^2 \quad \forall\ 1 \leq i \leq n.$$

By (Å), $v_n^2 = \frac{1}{2\sqrt{\pi n\sqrt{h}x(1-x)}} f(x) + o(\frac{1}{n\sqrt{h}})$. This implies

$$E\Big[\frac{\xi_{in}}{v_n}\Big]^k \le k!\Big(\frac{C * h^{-1/4}}{\sqrt{n}}\Big)^{k-2} E\Big[\frac{\xi_{in}}{v_n}\Big]^2. \tag{2.25}$$

Then, by (2.25) and the Bernstein inequality, for any positive number c,

$$P\bigg(\Big|\frac{\sum_{i=1}^n \xi_{in}}{v_n}\Big| \ge c\sqrt{\log n}\bigg) \le 2\exp\Big(-\frac{c^2\log n}{4\Big(1 + ch^{-1/4}\sqrt{\log n}/\sqrt{n}\Big)}\Big).$$

Since $h^{-1/2}\log n/n \to 0$, so for n large enough,

$$P\bigg(\Big|\frac{\sum_{i=1}^n \xi_{in}}{v_n}\Big| \ge c\sqrt{\log n}\bigg) \le 2\exp\Big(-\frac{c^2\log n}{8}\Big).$$

Upon taking $c = 8$, we have

$$P\Big(\Big|\sum_{i=1}^n \xi_{in}\Big| \ge c\sqrt{\log n}v_n\Big) \le \frac{2}{n^8}.$$

Since $\sum_{n=1}^\infty n^{-8} < \infty$, so by the Borel-Cantelli Lemma and by the fact $v_n^2 = O(\frac{1}{\sqrt{hn}})$, we obtain

$$\hat{f}_n(x) - E\hat{f}_n(x) = \sum_{i=1}^n \xi_{in} = o\Big(\frac{h^{-1/4}\sqrt{\log n}}{\sqrt{n}}\Big).$$

To bound the $\sum_{i=1}^n \xi_{in}$ uniformly for all $x \in [a,b]$, we partition the interval $[a,b]$ by the equally spaced points x_i, $i = 0,1,2,\ldots,N_n$, such that $a = x_0 < x_1 < x_2 < \cdots < x_{N_n} = b$, $N_n = n^3$. It is easily seen that

$$P\Big(\max_{0\le j\le N_n}\Big|\sum_{i=1}^n \xi_{in}(x_j)\Big| > c\frac{h^{-1/4}\sqrt{\log n}}{\sqrt{n}}\Big) \le \frac{2N_n}{n^8} = \frac{2}{n^5}.$$

Borel-Cantelli Lemma implies that

$$\max_{0\le j\le N_n}\Big|\sum_{i=1}^{n}\xi_{in}(x_j)\Big| = o\Big(\frac{h^{-1/4}\sqrt{\log n}}{\sqrt{n}}\Big). \tag{2.26}$$

For any $x \in [x_j, x_{j+1}]$,

$$\begin{aligned}\xi_{in}(x)-\xi_{in}(x_j) &= n^{-1}[K_{x/h+1,(1-x)/h+1}(X_i)-EK_{x/h+1,(1-x)/h+1}(X_i)]\\ &\quad -n^{-1}[K_{x_j/h+1,(1-x_j)/h+1}(X_i)-EK_{x_j/h+1,(1-x_j)/h+1}(X_i)].\end{aligned}$$

For ease in discussion, we will denote $K_{x/h+1,(1-x)/h+1}(X)$ as $K(x,X)$ understanding the dependance on h is implicit. A Taylor expansion of $K(x,X_i)$ at $x=x_j$ up to the first order leads to the following expression for the difference $K_{x_j/h+1,(1-x_j)/h+1}(X_i)-K_{x_j/h+1,(1-x_j)/h+1}(X_i)$:

$$|K_{x_j/h+1,(1-x_j)/h+1}(X_i)-K_{x_j/h+1,(1-x_j)/h+1}(X_i)| \approx (x-x_j)K'(\tilde{x},X) \tag{2.27}$$

for $\tilde{x}\in[x_j,x_{j+1}]$. We will bound the difference in (3.32) by evaluating $K'(x,X)$ as follows:

$$K'(x,X) = \frac{\mathcal{A}}{B(x/h+1,(1-x)/h+1)} + X^{x/h}(1-X)^{(1-x)/h}\mathcal{B}$$

where, $\mathcal{A}$ and $\mathcal{B}$ correspond to the derivative of $X^{x/h}(1-X)^{(1-x)/h}$ and $1/B(x/h+1,(1-x)/h+1)$, respectively. We can show that,

$$\mathcal{A} = \frac{1}{h}\Big[\log X-\log(1-X)\Big]X^{x/h}(1-X)^{(1-x)/h} \tag{2.28}$$

and

$$\mathcal{B} = \frac{\psi^0((1-x)/h+1)-\psi^0(x/h+1)}{hB(x/h+1,(1-x)/h+1)} \tag{2.29}$$

where, $\psi^0(x)$ represents the digamma function. Exploiting the properties of the digamma

function and some straightforward algebra, we can upper bound the numerator of (3.34) as

$$\psi^0((1-x)/h+1) - \psi^0(x/h+1) \leq \frac{1-2x+h}{x}.$$

Substituting the expressions for $\mathcal{A}$ and $\mathcal{B}$ in (3.32) we have,

$$\begin{aligned} K'(x,X) \quad \leq \quad & \frac{X^{x/h}(1-X)^{(1-x)/h}}{B(x/h+1,(1-x)/h+1)}\Big[\log X - \log(1-X)\Big] \\ & + \frac{X^{x/h}(1-X)^{(1-x)/h}}{B(x/h+1,(1-x)/h+1)}\frac{1-2x+h}{x}. \end{aligned}$$

Observing that $\frac{X^{x/h}(1-X)^{(1-x)/h}}{B(x/h+1,(1-x)/h+1)}$ corresponds to the beta density, we can bound it based on its value at its mode similar to the derivation in (2.23). Returning to the notation in (3.32), this bound can be written as,

$$K'(\tilde{x},X) \leq \frac{ph^{-3/2}}{\sqrt{x(1-x)}} \tag{2.30}$$

for some positive constant p. Since $0 \leq x - x_j \leq (b-a)/N_n$, and $\tilde{x} > 1/a$,

$$|K_{x/h+1,(1-x)/h+1}(X_i) - K_{x_j/h+1,(1-x_j)/h+1}(X_i)| \leq \frac{ph^{-3/2}}{N_n},$$

which implies that when n is large enough, for some constant p,

$$|\xi_{in}(x) - \xi_{in}(x_j)| \leq \frac{ph^{-3/2}}{nN_n}, \quad 1 \leq i \leq n. \tag{2.31}$$

These bounds imply that for all $x \in [x_j, x_{j+1}]$ and $0 \leq j \leq N_n - 1$,

$$\Big|\sum_{i=1}^n \xi_{in}(x) - \sum_{i=1}^n \xi_{in}(x_j)\Big| \leq \frac{ph^{-3/2}}{n^3} = o\Big(\frac{h^{-1/4}\sqrt{\log n}}{\sqrt{n}}\Big). \tag{2.32}$$

Finally, from (2.26) and (2.32), we obtain

$$\begin{aligned}
&\sup_{a\le x\le b}|\hat f_n(x)-E\hat f_n(x)| = \sup_{a\le x\le b}\Big|\sum_{i=1}^n \xi_{in}(x)\Big| \\
\le\ & \max_{0\le j\le N_n}\Big|\sum_{i=1}^n \xi_{in}(x_j)\Big| + \max_{0\le j\le N_n-1}\sup_{x\in[x_j,x_{j+1}]}\Big|\sum_{i=1}^n \xi_{in}(x)-\sum_{i=1}^n \xi_{in}(x_j)\Big| \\
=\ & o\Big(\frac{h^{-1/4}\sqrt{\log n}}{\sqrt{n}}\Big).
\end{aligned}$$

This, together with the result $\sup_{a\le x\le b}|E\hat f_n(x)-f(x)| = O(h)$, completes the proof of Theorem 2.1.4. □

Proof of Theorem 2.1.5. Given that $\hat m_n(x)-m(x) = \Big(B_n(x)+V_n(x)\Big)/\hat f_n(x)$, it suffices to prove the following two facts:

$$\sup_{x\in[a,b]}\Big|\frac{B_n(x)}{\hat f_n(x)}\Big| = O(h)+o\Big(\frac{h^{-1/4}\sqrt{\log n}}{\sqrt{n}}\Big), \tag{2.33}$$

$$\sup_{x\in[a,b]}\Big|\frac{V_n(x)}{\hat f_n(x)}\Big| = O(h)+o\Big(\frac{h^{-1/4}\sqrt{\log n}}{\sqrt{n}}\Big). \tag{2.34}$$

We shall prove (2.34) only, the proof of (2.33) being similar.

Let β,η be such that $\beta<2/5$, $\beta(2+\eta)>1$ and $\beta(1+\eta)>2/5$ and define $d_n=n^\beta$. For each i, write $\varepsilon_i = \varepsilon_{i1}^{d_n}+\varepsilon_{i2}^{d_n}+\mu_i^{d_n}$, with

$$\varepsilon_{i1}^{d_n} = \varepsilon_i I(|\varepsilon_i|>d_n),\ \varepsilon_{i2}^{d_n} = \varepsilon_i I(|\varepsilon_i|\le d_n)-\mu_i^{d_n},\ \mu_i^{d_n} = E[\varepsilon_i I(|\varepsilon_i|\le d_n)|X_i].$$

Denoting $K_{x/h+1,(1-x)/h+1}(X)$ as $K_{x,h}(X)$, we can express

$$\frac{V_n(x)}{\hat f_n(x)} = \frac{\sum_{i=1}^n K_{x,h}(X_i)\varepsilon_{i1}^{d_n}}{\sum_{i=1}^n K_{x,h}(X_i)} + \frac{\sum_{i=1}^n K_{x,h}(X_i)\varepsilon_{i2}^{d_n}}{\sum_{i=1}^n K_{x,h}(X_i)} + \frac{\sum_{i=1}^n K_{x,h}(X_i)\mu_i^{d_n}}{\sum_{i=1}^n K_{x,h}(X_i)}. \tag{2.35}$$

We will prove (2.34) by considering the three terms on the right hand side of (2.35). The

first term involving $\varepsilon_{i1}^{d_n}$ can be bounded via Markov Inequality as

$$\sum_{n=1}^{\infty} P(|\varepsilon_n| > d_n) \leq E|\varepsilon|^{2+\eta} \sum_{n=1}^{n} \frac{1}{d_n^{2+\eta}} < \infty.$$

Borel-Cantelli Lemma implies that

$$\begin{aligned} & P\{\exists N, |\varepsilon_n| \leq d_n \text{ for } n > N\} = 1 \\ \Rightarrow\ & P\{\exists N, |\varepsilon_i| \leq d_n, i = 1, 2, \ldots, n, \text{ for } n > N\} = 1 \\ \Rightarrow\ & P\{\exists N, \varepsilon_{i,1}^{d_n} = 0, i = 1, 2, \ldots, n, \text{ for } n > N\} = 1. \end{aligned}$$

Hence,

$$\sup_{x\in[a,b]} \Big| \frac{\sum_{i=1}^n K_{x,h}(X_i)\varepsilon_{i,1}^{d_n}}{\sum_{i=1}^n K_{x,h}(X_i)} \Big| = O(n^{-k}), \quad \forall k > 0.$$

Now consider the third term in (2.35). Since $E(\varepsilon_i|X_i) = 0$, so $\mu_i^{d_n} = -E[\varepsilon_i I(|\varepsilon_i| > d_n)|X_i]$, then from assumption (A4), we have $|\mu_i^{d_n}| \leq c d_n^{-(1+\eta)}$. Hence

$$\sup_{x\in[a,b]} \Big| \frac{\sum_{i=1}^n K_{x,h}(X_i)\mu_i^{d_n}}{\sum_{i=1}^n K_{x,h}(X_i)} \Big| \leq c d_n^{-(1+\eta)} = o\Big(\frac{h^{-1/4}}{\sqrt{n}}\Big).$$

For the second term $\varepsilon_{i,2}^{d_n}$ in (2.35), we have $E[\varepsilon_{i,2}^{d_n}|X_i] = 0$, and it is easy to show that

$$\mathrm{Var}(\varepsilon_{i,2}^{d_n}|X_i) = \sigma^2(X_i) + O[d_n^{-\eta} + d_n^{-2(1+\eta)}]$$

and for $k \geq 2$, $E(|\varepsilon_{i,2}^{d_n}|^k|X_i) \leq 2^{k-2} d_n^{k-2} E(|\varepsilon_{i,2}^{d_n}|^2|X_i)$. Then from (2.24) and the boundedness of $\sigma^2(x)$ over $(0,1)$, we have

$$\begin{aligned} E|n^{-1}K_{x,h}(X_i)\varepsilon_{i,2}^{d_n}|^k &\leq n^{-k} E[K_{x,h}^k(X) E(|\varepsilon_{i,2}^{d_n}|^k|X_i)] \\ &\leq c n^{-k} 2^{k-2} d_n^{k-2} E K_{x,h}^k(X)\sigma^2(X) \\ &\leq \Big(c d_n/n\sqrt{h}\Big)^{k-2} E|n^{-1}K_{\alpha_n}(x, X_i)\varepsilon_{i,2}^{d_n}|^2. \end{aligned}$$

Since

$$\begin{aligned} E|n^{-1}K_{x,h}(X_i)\varepsilon_{i,2}^{d_n}|^2 &= \frac{1}{n^2}E[K_{x,h}^2(X)\sigma^2(X)][1+o(1)] \\ &= \frac{1}{n^2}A_h(x)f(x)\sigma^2(X)][1+o(1)] \\ &= \frac{h^{-1/2}f(x)\sigma^2(x)}{2n^2\sqrt{\pi x(1-x)}}[1+o(1)], \end{aligned}$$

the random variable $n^{-1}K_{x,h}(X_i)\varepsilon_{i,2}^{d_n}$ satisfies the Cramér condition. So, using the Bernstein inequality as in proving Theorem 2.1.4, one establishes the fact that for all $c>0$,

$$\begin{aligned} P\Big(\Big|\sum_{i=1}^n K_{x,h}(X_i)\varepsilon_{i,2}^{d_n}\Big| \geq c\sqrt{\log n}\sqrt{\sum_{i=1}^n E\Big[K_{x,h}(X_i)\varepsilon_{i,2}^{d_n}\Big]^2}\Big) \\ \leq 2\exp(-c^2\log n/8). \end{aligned}$$

Take $c=4$ and $C(x)=c\sqrt{f(x)\sigma^2(x)/(2\sqrt{\pi x(1-x)})}$ in the above inequality to obtain

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^n K_{x,h}(X_i)\varepsilon_{i,2}^{d_n}\Big| \geq C(x)\sqrt{h^{-1/2}\log n/n}\Big) \leq \frac{2}{n^2},$$

by Borel-Cantelli Lemma and the boundedness of $f(x)\sigma^2(x)/\sqrt{x(1-x)}$ over $x\in[a,b]$, this implies, for each $x\in[a,b]$,

$$\Big|\frac{1}{n}\sum_{i=1}^n K_{x,h}(X_i)\varepsilon_{i,2}^{d_n}\Big| = o\Big(\frac{h^{-1/4}\sqrt{\log n}}{\sqrt{n}}\Big). \tag{2.36}$$

To show the above bound is indeed uniform, we can use the same technique that was used to demonstrate the uniform convergence of $\hat{f}_n(x)$ in the proof of Theorem 2.1.4. In fact, the only major difference is that, instead of using (2.31), we should use the inequality

$$\Big|K_{x,h}(X_i)\varepsilon_{i,2}^{d_n} - K_{x_j,h}(X_i)\varepsilon_{i,2}^{d_n}\Big| \leq \frac{ch^{-3/2}d_n}{N_n}, \quad x\in[x_j,x_{j+1}],\ 1\leq i\leq n.$$

The above result, together with the facts that $f(x)$ is bounded below from 0 on $[a, b]$, and $\sup_{x\in[a,b]} |\hat{f}_n(x) - f(x)| = o(1)$, implies

$$\sup_{x\in[a,b]} \Big| \frac{\sum_{i=1}^n K_{x,h}(X_i)\varepsilon_{i,2}^{d_n}}{\sum_{i=1}^n K_{x,h}(X_i)} \Big| = o\Big(\frac{h^{-1/4}\sqrt{\log n}}{\sqrt{n}}\Big), \quad \text{a.s.}$$

This concludes the proof of Theorem 2.1.5. □

Chapter 3

Beta Kernel based Local Linear Regression

As seen in the previous chapter, kernel regression estimators could suffer from design bias (a bias that depends on the distribution of X) and/or boundary bias (a bias near the end points of X). These biases can be reduced by using local polynomial regression estimators which is the focus of this chapter.

The relationship between a scalar response Y and a one-dimensional covariate X is often investigated through the regression model $Y = m(X) + \varepsilon$, where ε accounts for the random error with usual assumptions $E(\varepsilon|X = x) = 0$ and $\sigma^2(x) := E(\varepsilon^2|X = x) > 0$, for almost all x. Local polynomial regression is based on the idea that we can improve the estimator of m by using a higher order polynomial as a local approximation to m. Taylor's theorem suggests that this is a reasonable approach as $m(x)$ can be represented based on some z in the neighborhood of x as,

$$\begin{aligned} m(x) &\approx m(z) + m^{(1)}(z)(z-x) + \frac{m^{(2)}(z)}{2!}(z-x)^2 + \cdots + \frac{m^{(p)}(z)}{p!}(z-x)^p \\ &= \beta_0 + \beta_1(z-x) + \beta_2(z-x)^2 + \cdots + \beta_p(z-x)^p \\ &\equiv M_x(z,\beta) \end{aligned}$$

Here $m^{(p)}$ denotes the pth derivative of m. The underlying idea in local polynomial kernel regression is to estimate the regression function at a point x by "locally" fitting a pth degree polynomial to the data such that a weighted least squares error metric is minimized. That is, local polynomial regression of order p minimizes

$$\sum_{i=1}^{n} w_i(x)(Y_i - M_x(x_i, \beta))^2 \tag{3.1}$$

The weights $w_i(x)$ are chosen according to the height of the kernel function centered about the point x. Given the usual symmetric kernel function choices, observations close to x have more influence on the regression estimate at x than those that are far away. The actual amount of influence of distant observations is controlled by the bandwidth parameter h that goes along with the Kernel choice (i.e., $K_h(x_i - x)$). As discussed before, if h is small the local fitting depends strongly on observations in the vicinity of x leading to wiggly estimates. In the extreme case, when h gets closer to 0, we end up with interpolation of data. When h gets larger, the estimate tends towards the weighted least squares polynomial that fits the data.

It is straightforward to see that the traditional Nadaraya Watson estimator corresponds to the case of $p = 0$ as it corresponds to fitting local constants. In this chapter, we are particularly interested in local linear kernel regression estimators corresponding to $p = 1$. This is motivated by the fact that local linear kernel estimators while providing much more favorable asymptotic properties and boundary behaviors (relative to traditional kernel estimators), also allow for mathematical analysis with reasonable tractability. [Fan(1992), Fan(1993)] proved that the leading bias term in the case of a local linear regression estimate with a symmetric kernel depends on x only through $m''(x)$ (hereon, we will use $m''(x)$ to indicate $m^{(2)}(x)$) which effectively captures the error in the linear approximation of $m(x)$. That if m is close to being linear, the local linear fit is more accurate resulting in lower bias. If m has larger curvature at x, then a linear fit will result in a more biased estimate.

[Fan(1992)] also demonstrated that the bias increases with more smoothing. When $p = 1$, these properties were shown to hold for both interior and boundary points with minor differences in constants associated with the leading bias terms. In general, [Fan(1992), Hastie and Loader(1993), Fan(1993)] have demonstrated the favourable properties of local polynomial kernel estimators of odd degree. For even p, there is discrepancy between the orders of magnitude of bias in the interior and boundary (usually referred to as the boundary bias). For example, it has been shown that for even p, the minimum MSE is of the order of $n^{-\frac{(2p+4)}{(2p+5)}}$ at the interior and $n^{-\frac{(2p+2)}{(2p+3)}}$ at the boundaries. For the NW estimator with $p = 0$, we already know that the optimal interior MSE order of $n^{-4/5}$ is inflated to $n^{-2/3}$ near the boundaries.

In this chapter, our objective is to investigate the impact of transitioning from using symmetric kernel functions in local linear regression estimators to asymmetric kernel functions. Specifically, our interest is in quantifying the performance of Beta kernel based local linear regression and comparing its performance to a traditional normal kernel based local linear regression estimator.

Analogous to the local linear estimator with classical symmetric kernel functions, for a sample (X_i, Y_i) from the regression model, at any fixed $x \geq 0$, we define a function $L(\beta)$ of $\beta = (\beta_0, \beta_1)'$ as

$$L(\beta) = \sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i)[Y_i - \beta_0 - \beta_1(X_i - x)]^2.$$

Then the local linear estimator of $m(x)$ and $m'(x)$ with Beta kernel function is simply the minimizer of $L(\beta)$, where $m'(x)$ is the first derivative of $m(x)$ with respect to x. In matrix form, the minimizer has the form of

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y, \tag{3.2}$$

where

$$\mathrm{X} = \begin{pmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{pmatrix}, \ \mathrm{W} = \begin{pmatrix} \ddots & & \\ & \frac{1}{n}K_{x/h+1,(1-x)/h+1}(X_i) & \\ & & \ddots \end{pmatrix}, \ \mathrm{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

Under some regularity conditions on the underlying density function $f(x)$ and the regression function $m(x)$, our goal is to prove the asymptotic normality of $\hat{\beta}$, as well as its uniform consistency. As a first step towards that goal, we aim to uncover the large sample properties of the conditional bias and variance in the next section.

3.1 Large Sample Results of $\hat{m}(x)$ and $\hat{m}'(x)$

In this section, we discuss the large sample properties of the proposed estimator $\hat{\beta}$ defined in (3.2). To be specific, the asymptotic expressions of the conditional bias, variance, hence the MSE of $\hat{m}(x)$ will be established.

We start with a list of technical assumptions needed for developing the asymptotic theories.

(C1). The third order derivatives of f is continuous and bounded on $[0, 1]$.

(C2). $E(\varepsilon|X) = 0$, and the second order derivative of m is continuous and bounded on $[0, 1]$.

(C3). The second order derivative of $\sigma^2(x) = E(\varepsilon^2|X = x)$, $f(x)\sigma^2(x)$ and $f(x)\sigma^4(x)$ are continuous and bounded on $[0, 1]$.

(C4). For some $\delta > 0$, the second order derivative of $E(|\varepsilon|^{2+\delta}|X = x)$ is continuous and bounded in $x \in (0, \infty)$.

(C5). $h \to 0$, $n\sqrt{h} \to \infty$ as $n \to \infty$.

These conditions are similar to those assumed in deriving the MSE properties of Beta kernel regression estimates in the previous chapter.

3.1.1 Conditional Bias and Variance

For the sake of simplicity, denote $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$,

$$S = \begin{pmatrix} f(x) + (1-2x)f'(x)h + \frac{1}{2}x(1-x)f''(x)h & (1-2x)f(x)h + x(1-x)f'(x)h \\ (1-2x)f(x)h + x(1-x)f'(x)h & x(1-x)f(x)h \end{pmatrix},$$

$$S^* = \begin{pmatrix} f(x)\sigma^2(x) + \frac{(1-2x)k_1(x)h}{2} + \frac{x(1-x)k_2(x)h}{4} & k_3(x) \\ k_3(x) & \frac{x(1-x)f(x)\sigma^2(x)h}{2} \end{pmatrix},$$

where $k_1(x) = f'(x)\sigma^2(x) + f(x)\sigma^{2'}(x)$,
$k_2(x) = f''(x)\sigma^2(x) + 2f'(x)\sigma^{2'}(x) + f(x)\sigma^{2''}(x)$,
$k_3(x) = \frac{(1-2x)f(x)\sigma^2(x)h}{2} + \frac{x(1-x)(f'(x)\sigma^2(x)+f(x)\sigma^{2'}(x))h}{2}$.

The following theorem presents the conditional biases and variances of $\hat{m}(x)$ and $\hat{m}'(x)$.

Theorem 1 Suppose the assumptions (C1), (C2), (C3), and (C5) hold. Then, for any $x \in (0,1)$ with $f(x) > 0$,

$$\begin{aligned}
\text{bias}(\hat{m}(x)|X) &= \frac{x(1-x)m''(x)h}{2} + o_p(h), \\
\text{bias}(\hat{m}'(x)|X) &= \frac{(2x-1)m''(x)h}{2} - \frac{x(1-x)m''(x)f'(x)h}{2f(x)} + o_p(h), \\
\text{Var}(\hat{m}(x)|X) &= \frac{1}{2n\sqrt{\pi x(1-x)h}} e_0^T S^{-1} S^* S^{-1} e_0 + o_p\left(\frac{1}{n\sqrt{h}}\right), \\
&= \frac{\sigma^2(x)}{2nf(x)\sqrt{\pi x(1-x)h}} + o_p\left(\frac{1}{n\sqrt{h}}\right), \\
\text{Var}(\hat{m}'(x)|X) &= \frac{1}{2n\sqrt{\pi x(1-x)h}} e_1^T S^{-1} S^* S^{-1} e_1 + o_p\left(\frac{1}{nh\sqrt{h}}\right) \\
&= \frac{\sigma^2(x)}{2n\sqrt{\pi} f(x)(x(1-x)h)^{3/2}} + o_p\left(\frac{1}{nh\sqrt{h}}\right).
\end{aligned}$$

where, $e_0 = (1,0)^T, e_1 = (0,1)^T$.

Thus, when $x \in (0,1)$, the conditional MSE of $\hat{m}(x)$ has the asymptotic expansion

$$\text{MSE}(\hat{m}(x)|\mathbf{X}) = \frac{(x(1-x)m''(x))^2 h^2}{4} + \frac{\sigma^2(x)}{2nf(x)\sqrt{\pi x(1-x)h}} + o_p(h^2) + o_p\left(\frac{1}{n\sqrt{h}}\right).$$

It is important to note that unlike the Beta kernel regression estimator discussed in Chapter 2, the Beta kernel based local linear regression estimator has a conditional bias that is not dependent on $f(x)$. This "design adaptivity" is a desirable property that does not translate to the conditional bias of the derivative of the regression function.

Similar to the N-W kernel regression case, one can choose the optimal smoothing parameter h by minimizing the leading term in the conditional MSE of $\hat{m}_n$ with respect to h. For $x \in (0,1)$, we can verify that h has the order of $n^{-2/5}$, with the corresponding MSE having the order of $n^{-4/5}$. Recall the same order is obtained for the Beta kernel regression estimate based on the same criterion. Additionally, the bias of $\hat{m}(x)$ is free of the term $f'(x)/f(x)$, indicating the proposed estimator has design adaptivity, a desired property inherited directly from the local linear smoothing. Understanding the behavior around the boundaries (i.e., $x = 0$ and $x = 1$) will be part of our future work.

To conclude this section, we would like to discuss the notion of an "equivalent kernel". In standard local linear estimation, it is finally the "equivalent kernel" that matters rather than the kernel which is used in the minimization process, see Fan and Gijbels (1992). We might wonder if this notion of "equivalent kernel" can be extended to local linear estimation with the Beta kernel. Unfortunately, due to the inherently different nature of the Beta kernel in contrast to the traditional kernel (which is a member of location family), the form of this "equivalent kernel" is relatively complicated. While we hardly see the usefulness of an "equivalent kernel" in developing any of the following theoretical results, we hope to pursue this as part of our future work.

To see this, let e_v be a 2-dimensional vector whose $v+1$-th element is 1, and the other

one is 0. Define $S_n = X'WX$, and

$$w_v^n(X,x) = e_v' S_n^{-1}(1, X-x)' K_{x/h+1,(1-x)/h+1}(X).$$

Then we can rewrite $\hat{m}^{(v)}(x)$ as a weighted sum of Y_i's,

$$\hat{m}^{(v)}(x) = e_v'(X'WX)^{-1}X'WY = \sum_{i=1}^{n} w_v^n(X_i,x)Y_i.$$

It is easy to see that the weights $w_v^n(X,x)$ also satisfies the discrete moment conditions

$$\sum_{i=1}^{n}(X_i-x)^q w_v^n(X_i,x) = \delta_{v,q}, \quad v,q = 0,1.$$

Define

$$H = \begin{pmatrix} 1 & 0 \\ 0 & h \end{pmatrix}, \quad S_h = \begin{pmatrix} 1 & 1-2x+\frac{x(1-x)f'(x)}{f(x)} \\ 1-2x+\frac{x(1-x)f'(x)}{f(x)} & x(1-x)/h \end{pmatrix},$$

then from the proof of Theorem 1 in the paper, we have

$$S_n = nf(x)HS_hH(1+o_p(1)).$$

Thus, we can rewrite $w_v^n(X,x)$ as the following

$$w_v^n(X,x) = \frac{1}{nf(x)h^v} e_v{}' S_h^{-1}(X-x)^v K_{x/h+1,(1-x)/h+1}(X)(1+o_p(1)),$$

and, accordingly,

$$\hat{m}^{(v)}(x) = \frac{1}{nf(x)h^v}\sum_{i=1}^{n} L_v^n(X_i,x)Y_i(1+o_p(1)),$$

where

$$L_v^n(X,x) = e_v{}' S_h^{-1}(X-x)^v K_{x/h+1,(1-x)/h+1}(X).$$

Different from the equivalent kernel in Fan and Gijbels, this "equivalent kernel" $L_v^n(X,x)$ can not be written as some function evaluated at $(X-x)/h$, and such a function is free from data and sample size!

3.1.2 Asymptotic Normality

The asymptotic normality $\hat{m}(x)$ is summarized in the following theorem, which can be used for constructing confidence intervals and hypothesis testing.

Theorem 2 Suppose the assumptions C1-C5 hold. Then for any $x \in (0,1)$ with $f(x) > 0$, as $nh^{3/2} \to \infty$ and $nh^{9/2} \to 0$,

$$\begin{aligned}
&\mathrm{diag}(1, \sqrt{h}) \cdot S_0 \cdot \sqrt{\frac{n\sqrt{h}}{f(x)\sigma^2(x)}} \left[\begin{pmatrix} \hat{m}(x) - m(x) \\ \hat{m}'(x) - m'(x) \end{pmatrix} - \frac{m''(x)}{2} \begin{pmatrix} x(1-x)f(x)h \\ 0 \end{pmatrix} \right] \\
&\xrightarrow{d} N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{2\sqrt{\pi x(1-x)}} & 0 \\ 0 & \frac{\sqrt{x(1-x)}}{4\sqrt{\pi}} \end{pmatrix} \right],
\end{aligned}$$

where

$$S_0 = \begin{pmatrix} f(x) + (1-2x)f'(x)h + \frac{x(1-x)f''(x)h}{2} & (1-2x)f(x)\sqrt{h} + x(1-x)f'(x)\sqrt{h} \\ (1-2x)f(x)\sqrt{h} + x(1-x)f'(x)\sqrt{h} & x(1-x)f(x) + f(x)h \end{pmatrix},$$

It is noted that there is a non-negligible asymptotic bias appearing in the above results, a characteristic shared with the Beta kernel regression estimates. These biases can be eliminated by under-smoothing which, in the current set up, is to select a proper h such that $nh^{9/2} \to 0$ for $x \in (0,1)$ without violating conditions $h \to 0, nh^{3/2} \to \infty$.

3.1.3 Uniform Almost Sure Consistency

In this section, we develop an almost sure uniform convergence result for $\hat{m}(x)$ over an arbitrary closed sub-interval in $(0,1)$. As was done for the Beta kernel regression estimator, we will apply the Borel-Cantelli lemma and the Bernstein inequality, after verifying the Cramér condition: for some $k \geq 2$, $c > 0$, and h small enough,

$$E|K_{x/h+1,(1-x)/h+1}(X)|^k \leq k!\Big(\frac{c}{n\sqrt{h}}\Big)^{k-2} EK^2_{x/h+1,(1-x)/h+1}(X), \quad x > 0. \tag{3.3}$$

The following theorem gives the almost sure uniform convergence of $\hat{m}(x)$ to $m(x)$ over bounded sub-intervals of $(0,1)$.

Theorem 3 In addition to (C1)-(C5), further assume that $\log n/n\sqrt{h} \to 0$. Then for any constants a and b such that $0 < a < b < 1$,

$$\sup_{x\in[a,b]} \Big|\hat{m}(x) - m(x)\Big| = O(h) + o\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big), \quad \text{a.s.}$$

By assuming some stronger conditions on the tails of f and m at the boundaries, the above uniform almost sure convergence results can be extended to be over some suitable intervals increasing to $(0,1)$. However, we do not pursue it here simply because of the involved technical details and lack of a useful application.

3.2 Numerical Results

Once again, to evaluate the finite sample performance of the proposed local linear regression estimator based on the Beta kernel, a simulation and comparison study will be conducted in this section, together with an application to a real data set. The selection of smoothing parameters will follow the approach discussed in the previous chapter, i.e., least squares cross validation (LSCV) and generalized cross validation (GCV).

Simulation Study: We retain the simulation setup used to evaluate the Beta kernel regression estimator discussed in the previous chapter. Specifically, the underlying density function of the design variable is chosen to be uniform between 0 and 1. A quadratic regression function is selected for generating the response variable. That is, we set $m(x) = 10(x-0.5)^2$ and the response $Y = m(x) + \varepsilon$ where ε values are drawn from a $N(0,1)$ distribution. For one set of random data, Figure 3.1 presents the estimated $\hat{m}(x)$ based on bandwidth selected using 5-fold GCV. For comparison, along with beta kernel based local linear regression estimator, we present the N-W estimator, Beta kernel estimator from the previous chapter and the local linear estimator with the normal kernel. The true regression curve is also plotted for reference. Results based on sample sizes of 100 and 300 are presented in Figure 3.1 and Figure 3.2, respectively. For the same data set, the regression estimate based on bandwidth derived from the LSCV procedure is similar to the result using GCV criterion and is presented in Figures 3.3 and 3.4. The optimal bandwidth values and the resulting mean squared error (MSE) values (defined as the average of squared differences between the estimated and true values of the regression functions at n equally spaced x-values) for the analyzed data set are provided in Table 3.1. From these simulation results, it is interesting to notice that the two local linear estimators perform almost equally well. As discussed in the previous chapter, the Beta kernel estimator offers a better MSE performance relative to the NW estimator. However, the Beta and Normal local linear regression estimators perform the best as expected. These observations, together with the relative ease in deriving the bias and other properties of Beta local linear regression estimator, make it a strong candidate in the field of kernel regression estimators.

Real Data Example: Finally, we apply the Beta kernel based local linear regression estimation procedure to the data set from Azzalini and Bowman (1990) on the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. The data consist of 299 pairs of measurements on the waiting time X to the next eruption, and the eruption time Y in minutes, which were collected continuously from August 1st until August 15th, 1985. The nonpara-

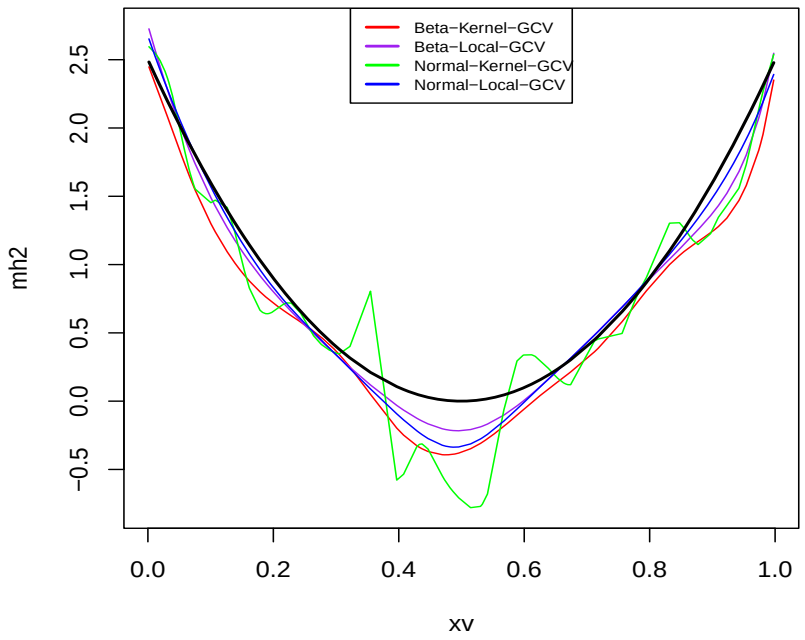


Figure 3.1: Comparison of Local linear Kernel Regression Estimators - $f(x) \sim U[0,1]$, $m(x) = 10(x-0.5)^2$, ε values are drawn from a $N(0,1)$, GCV, sample size = 100

		LSCV(5-fold)		GCV	
	Kernel Type	h_{LSCV}	$MSE(h_{LSCV})$	h_{GCV}	$MSE(h_{GCV})$
$n = 300$	Beta Kernel	0.0249	0.0310	0.0199	0.0314
	Nadaraya-Watson	0.0547	0.0578	0.0448	0.0682
	Beta Local Linear	0.0497	0.0125	0.0398	0.0146
	Normal Local Linear	0.0945	0.0122	0.0796	0.0146
$n = 100$	Beta Kernel	0.0398	0.0617	0.0298	0.0604
	Nadaraya-Watson	0.0796	0.0847	0.0647	0.1030
	Beta Local Linear	0.0647	0.0178	0.0647	0.0178
	Normal Local Linear	0.1045	0.0161	0.0945	0.0212

Table 3.1: MSE associated with different regression estimators with optimal bandwidths

metric regression procedure developed in this dissertation will be used to investigate the relationship between Y and X. Similar to the setup in the previous chapter, the waiting time X will be transformed to $T = (X-43)/30$ first. As a result, the range of T is between 0 and 2.17. Here 43 is the minimum value of X.

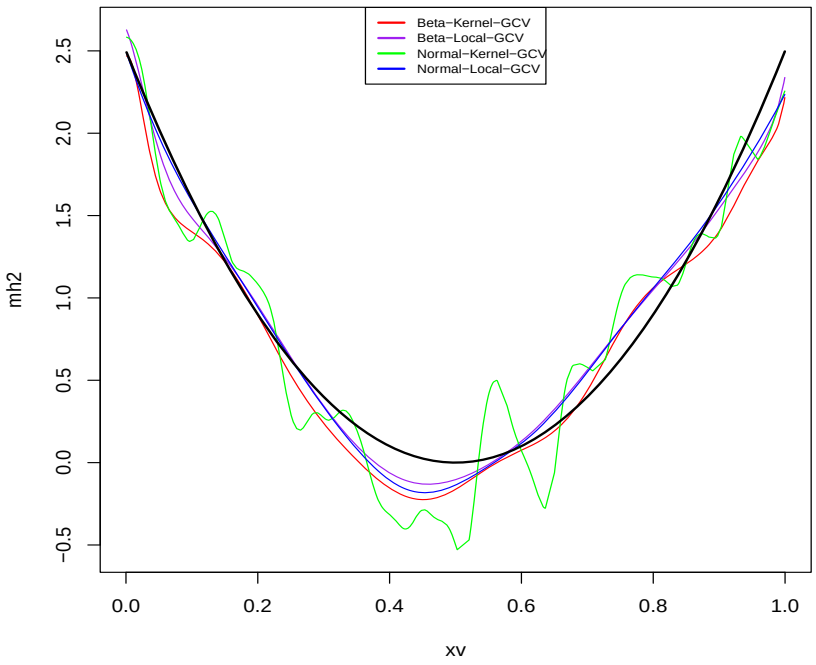


Figure 3.2: Comparison of Local linear Kernel Regression Estimators - $f(x) \sim U[0,1]$, $m(x) = 10(x-0.5)^2$, ε values are drawn from a $N(0,1)$, GCV, sample size = 300

The scatter plot in Figures 3.5 and 3.6 shows the data structure with Y against T. The estimated regression curves are imposed on the scatter plot in Figures 3.5 and 3.6 with solid lines. For comparison purposes, the normal kernel, the local linear with normal kernel and the Beta kernel regression curves are also provided. All four estimation procedures use bandwidths chosen based on 5-fold LSCV criteria in Figure 3.5. It is important to note that using the GCV criterion also results in similar performance as presented in Figure 3.6. For the proposed local linear estimator with Beta kernel, the optimal bandwidth with 5-fold LSCV is 0.01497, for local linear with normal kernel, the optimal bandwidth equals 0.05974, and for the Normal and Beta kernel estimator, it is 0.07467 and 0.01994, respectively. Clearly, all four estimation procedures capture the structure of the data set. However, the normal kernel based regression estimator shows more variability relative to the other schemes. Both the local linear regression estimators behave similarly with the Beta kernel

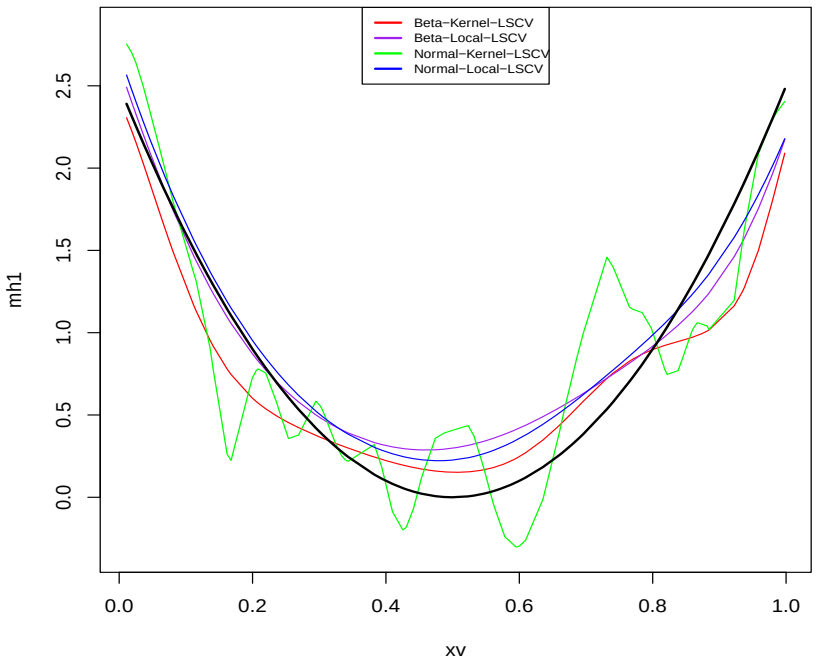


Figure 3.3: Comparison of Local linear Kernel Regression Estimators - $f(x) \sim U[0,1]$, $m(x) = 10(x-0.5)^2$, ε values are drawn from a $N(0,1)$, 5 fold LSCV, sample size = 100

based local linear regression estimator able to better capture the data structure at the boundaries.

3.3 Proof of the Main Results

This section contains the proof of the large sample results presented in previous sections. Since Beta density function and its moments will be repeatedly referred to in the following proofs, so for the sake of convenience, we list all the needed results here. Density function of a Beta distribution with shape parameters p and q is

$$g(u;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} u^{p-1}(1-u)^{q-1} \quad u \in [0,1].$$

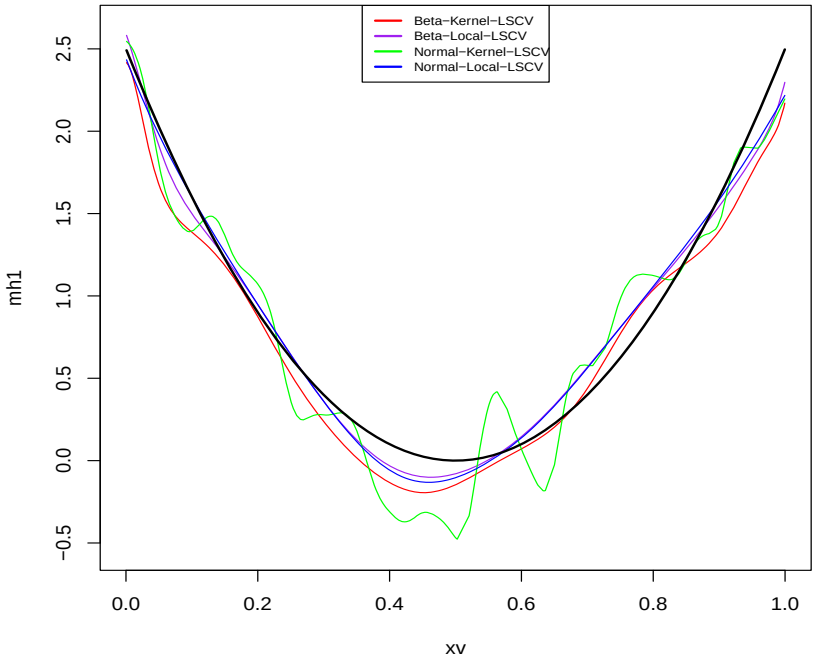


Figure 3.4: Comparison of Local linear Kernel Regression Estimators - $f(x) \sim U[0,1]$, $m(x) = 10(x-0.5)^2$, ε values are drawn from a $N(0,1)$, 5-fold LSCV, sample size = 300

Its mean μ and variance τ^2 are $\mu = \frac{p}{p+q}$, and $\tau^2 = \frac{pq}{(p+q)^2(p+q+1)}$.

Let $p = x/h + 1, q = (1-x)/h + 1, x \in [0,1]$. The following lemma on the inverse beta distribution is crucial for the subsequent arguments.

Lemma 3.3.1 *Let $l(u)$ be a function such that the second order derivative of $l(u)$ is continuous and bounded on $[0,1]$. Then, for all $x \in [0,1]$,*

$$\int_0^1 g(u;p,q)l(u)du = l(x) + \left[(1-2x)l'(x) + \frac{1}{2}x(1-x)l''(x))\right]h + o(h). \qquad (3.4)$$

Lemma 3.2.1 is the same as Lemma 2.4.1 in Chapter 2. It is reproduced here for the sake of completeness and ease of referencing.

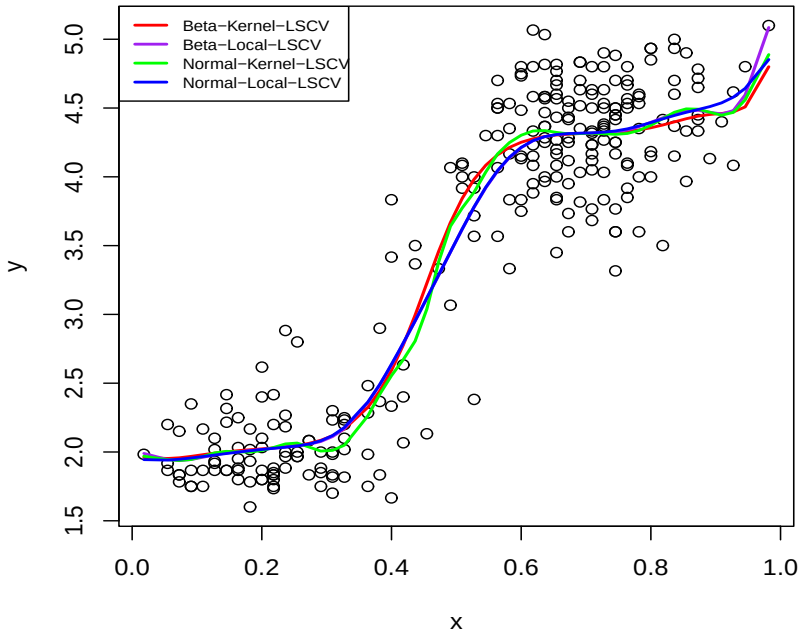


Figure 3.5: Comparison of Local linear Kernel Regression Estimators on Old Faithful geyser data (5-fold LSCV)

The following decomposition of $\hat{\beta}$ will be also used in the proofs below.

$$\begin{aligned} E(\hat{\beta}|X) &= (X^T W X)^{-1} X^T W E(Y|X) \\ &= (X^T W X)^{-1} X^T W(\mathbf{m} - X\beta + X\beta) = \beta + (X^T W X)^{-1} X^T W r, \end{aligned} \tag{3.5}$$

$$\begin{aligned} \mathrm{Var}(\hat{\beta}|X) &= (X^T W X)^{-1} X^T W \mathrm{Var}(Y|X) W X (X^T W X)^{-1} \\ &= (X^T W X)^{-1} X^T \Sigma X (X^T W X)^{-1}, \end{aligned} \tag{3.6}$$

where $\mathbf{m} = (m(X_1), \cdots, m(X_n))$, $r = \mathbf{m} - X\beta$, $\mathrm{Var}(Y|X) = \mathrm{diag}\{\sigma^2(X_i)\}$, and $\Sigma = W\mathrm{Var}(Y|X)W = \mathrm{diag}\{K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i)\}$, for $i = 1, \cdots, n$, $K_{x/h+1,(1-x)/h+1}(X_i) =$

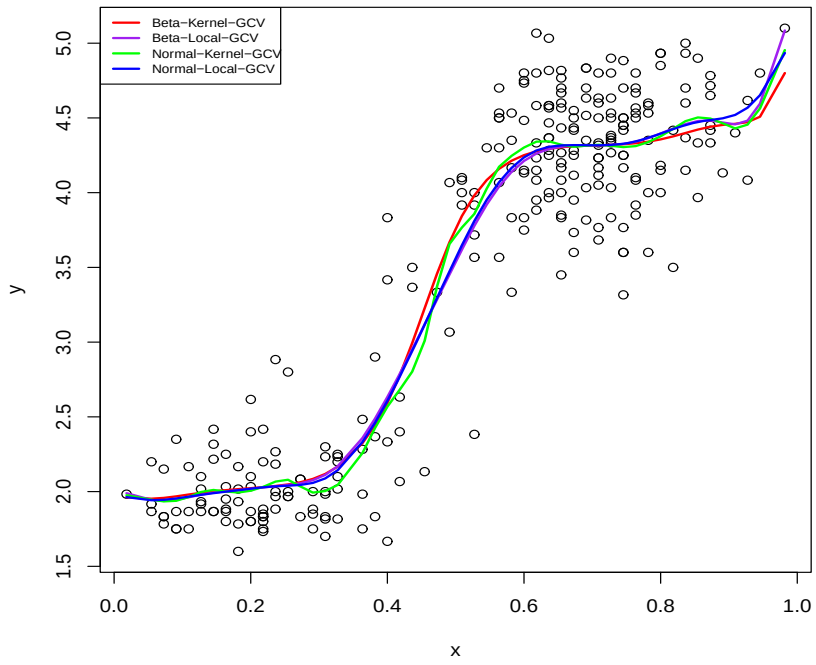


Figure 3.6: Comparison of Local linear Kernel Regression Estimators on Old Faithful geyser data (GCV)

$\frac{X_i^{x/h}(1-X_i)^{(1-x)/h}}{B(x/h+1,(1-x)/h+1)}$, with X, W defined in the previous section. Now we are ready to prove Theorem 1

Proof of Theorem 1: Consider the conditional variance $\mathrm{Var}(\hat{\beta}|X)$. Note that

$$X^TWX = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i) & \frac{1}{n}\sum_{i=1}^n (X_i-x)K_{x/h+1,(1-x)/h+1}(X_i) \\ \frac{1}{n}\sum_{i=1}^n (X_i-x)K_{x/h+1,(1-x)/h+1}(X_i) & \frac{1}{n}\sum_{i=1}^n (X_i-x)^2K_{x/h+1,(1-x)/h+1}(X_i) \end{pmatrix}.$$

Denote $X^TWX = \left(\frac{1}{n}\sum_{i=1}^n (X_i-x)^j K_{x/h+1,(1-x)/h+1}(X_i)\right),\ j=0,1,2.$ Applying Lemma 3.3.1 with $l(u) = (u-x)^j f(u)$, $u = X_i$, and $k = 1$, by assumption (C.1),

$$E\left(\frac{1}{n}\sum_{i=1}^n (X_i-x)^j K_{x/h+1,h}(X_i)\right) = l(x) + \left[(1-2x)l'(x) + \frac{1}{2}x(1-x)l''(x))\right]h + o(h),$$

then

$$E(X^TWX) =$$

$$\begin{pmatrix} f(x) + [(1-2x)f'(x) + \frac{1}{2}x(1-x)f''(x)]h + o(h) & (1-2x)f(x)h + x(1-x)f'(x)h + o(h) \\ (1-2x)f(x)h + x(1-x)f'(x)h + o(h) & x(1-x)f(x)h + o(h) \end{pmatrix} =$$

$$\begin{pmatrix} f(x) + (1-2x)f'(x)h + \frac{1}{2}x(1-x)f''(x)h & (1-2x)f(x)h + x(1-x)f'(x)h \\ (1-2x)f(x)h + x(1-x)f'(x)h & x(1-x)f(x)h \end{pmatrix} + o(h).$$

Denote the matrix on the right as S.

Now consider the variance of X^TWX.

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - x)^j K_{x/h+1,(1-x)/h+1}(X_i)\right) \leq \frac{1}{n}E\left((X_i - x)^j K_{x/h+1,(1-x)/h+1}(X_i)\right)^2, \ j = 0, 1, 2$$

which can be written as

$$\begin{aligned} &\frac{1}{n}E\left((X_i - x)^j K_{x/h+1,(1-x)/h+1}(X_i)\right)^2 \\ &= \frac{1}{n}\int_0^1 \left[\frac{\Gamma(2+1/h)}{\Gamma(x/h+1)\Gamma((1-x)/h+1)}u^{x/h}(1-u)^{(1-x)/h}\right]^2 (u-x)^{2j}f(u)du \\ &= \frac{\Gamma^2(2+1/h)(h/2)}{n\Gamma^2(x/h+1)\Gamma^2((1-x)/h+1)}\int_0^1 u^{2x/h}(1-u)^{2(1-x)/h}(u-x)^{2j}f(u)du \\ &= \frac{A_h(x)}{nB(2x/h+1, 2(1-x)/h+1)}\int_0^1 u^{2x/h}(1-u)^{2(1-x)/h}(u-x)^{2j}f(u)du \\ &= \frac{A_h(x)}{n}E[l(\gamma_x)]. \end{aligned}$$

Here, γ_x is a Beta$(2x/h+1, 2(1-x)/h+1)$ random variable, $l(u) = (u-x)^{2j}f(u)$ and

$$A_h(x) = \frac{B(2x/h+1, 2(1-x)/h+1)}{B^2(x/h+1, (1-x)/h+1} \tag{3.7}$$

.

By the continuity of f, f', f'', one can show that $E[l(\gamma_x)] = l(x) + \left[\frac{(1-2x)}{2} l'(x) + \frac{x(1-x)}{4} l''(x)\right] h + o(h)$. Then the variance of $X^T W X$

$$\begin{aligned} \mathrm{Var}(X^T W X) &= \frac{1}{n} A_h(x) \begin{pmatrix} f(x) + \left[\frac{(1-2x)}{2f'(x)} + \frac{x(1-x)}{4f''(x)}\right] h + o(h) & \frac{x(1-x)f(x)h}{2} + o(h) \\ \frac{x(1-x)f(x)h}{2} + o(h) & o(h) \end{pmatrix} \\ &= O(1/n\sqrt{h}). \end{aligned} \tag{3.8}$$

The last result directly follows after substituting the approximation of $A_h(x)$ derived in Chen(1999). That is,

$$A_h(x) \approx \begin{cases} \frac{1}{2\sqrt{\pi}} (x(1-x))^{-1/2} h^{-1/2}, & \text{if } x/h \text{ and } (1-x)/h \to \infty \text{ (interior);} \\ \frac{\Gamma(2K+1)}{2^{1+2K}\Gamma^2(K+1)} h^{-1}, & \text{if } x/h \to K \text{ or } (1-x)/h \to K \text{ (boundary).} \end{cases} \tag{3.9}$$

Therefore, when $x \in (0,1)$, together with (3.7) and (3.8), we get

$$X^T W X = S[1 + o(1)] + \frac{1}{\sqrt{n\sqrt{h}}} O(1) = S + \frac{1}{\sqrt{n\sqrt{h}}} O(1) = S[1 + o(1)]. \tag{3.10}$$

For the matrix $X^T \Sigma X$, a similar approach can be followed. That is,

$$\begin{aligned} X^T \Sigma X = & \begin{pmatrix} \frac{1}{n^2} \sum_{i=1}^n K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i) & \frac{1}{n^2} \sum_{i=1}^n (X_i - x) K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i) \\ \frac{1}{n^2} \sum_{i=1}^n (X_i - x) K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i) & \frac{1}{n^2} \sum_{i=1}^n (X_i - x)^2 K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i) \end{pmatrix} \\ & \triangleq \begin{pmatrix} \frac{1}{n^2} \sum_{i=1}^n (X_i - x)^j K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i) \end{pmatrix} \end{aligned} \tag{3.11}$$

where $j = 0, 1, 2$. Using Lemma 3.1 with $l(u) = (u-x)^j f(u)\sigma^2(u)$, $u = X_i$, and $k = 2$,

$$\begin{aligned} & E\left(\frac{1}{n^2}\sum_{i=1}^{n}(X_i - x)^j K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i)\right) \\ &= \frac{1}{n}\int_0^1 g^2(u,p,q)\sigma^2(u)(u-x)^j f(u)du \\ &= \frac{1}{n}A_h(x)E[l(\gamma_x)]. \end{aligned}$$

where, γ_x is a Beta$(2x/h + 1, 2(1-x)/h + 1)$ random variable. Using the approximation for $A_h(x)$, we can further simplify the previous result as,

$$\begin{aligned} & E\left(\frac{1}{n^2}\sum_{i=1}^{n}(X_i - x)^j K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i)\right) \\ &= \frac{1}{2n\sqrt{\pi x(1-x)h}}[1+o(1)](l(x) + [\frac{(1-2x)}{2}l'(x) + \frac{x(1-x)}{4}l''(x)]h + o(h)). \end{aligned}$$

Therefore,

$$\begin{aligned} & E(X^T\Sigma X) \\ &= \frac{1}{2n\sqrt{\pi x(1-x)h}}\begin{pmatrix} p_1(x) & p_2(x) \\ p_2(x) & \frac{x(1-x)}{2}f(x)\sigma^2(x)h \end{pmatrix} \\ &+ \frac{1}{2n\sqrt{\pi x(1-x)h}}o(1) = \frac{1}{2n\sqrt{\pi x(1-x)h}}S^*[1+o(1)], \end{aligned} \tag{3.12}$$

where, $p_1(x) = f(x)\sigma^2(x) + \frac{1-2x}{2}[f(x)\sigma^{2'}(x) + \sigma^2 f'(x)]h + \frac{x(1-x)}{4}[f(x)\sigma^{2''}(x) + \sigma^{2'}(x)f'(x) + \sigma^2(x)f''(x) + f'(x)\sigma^{2'}(x)]h$ and $p_2(x) = \frac{1-2x}{2}f(x)\sigma^2(x)h + \frac{x(1-x)}{2}(f'(x)\sigma^2(x) + f(x)\sigma^{2'}(x))h$. Next, the variance of $X^T\Sigma X$ is evaluated. Var $\left(\frac{1}{n^2}\sum_{i=1}^{n}(X_i - x)^j K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i)\right)$ corresponds to

$E\left(\frac{1}{n^2}\sum_{i=1}^{n}(X_i - x)^j K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i)\right)^2 -$

$\left[E\left(\frac{1}{n^2}\sum_{i=1}^{n}(X_i - x)^j K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i)\right)\right]^2$, with $j = 0,\ 1,\ 2$, for $x \in (0,1)$, as

$h \to 0$, the leading term in the previous equation can be written as

$$\begin{aligned}
&\frac{1}{n^3} E\left((X_i - x)^j K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i)\right)^2 \\
&= \frac{1}{n^3}\int_0^1 g^4(u,p,q)(u-x)^{2j} f(u)\sigma^4(u)du \\
&= \frac{1}{n^3} A_h^3(x) E[l(\gamma_1(x)]
\end{aligned}$$

where, $\gamma_1 x$ is a Beta$(4x/h+1, 4(1-x)/h+1)$ random variable, $l(u) = (u-x)^{2j} f(u)\sigma^4(u)$. Using the approximation for $A_h(x)$ and the continuity of $f\sigma^4, f\sigma^2\sigma^{2'}, f'\sigma^4$, and (C3), we can show that

$$\begin{aligned}
\mathrm{Var}(X^T\Sigma X) &= \frac{1}{8n^3\sqrt{\pi^3 x^3(1-x)^3 h^3}} \begin{pmatrix} p_3(x) + o(h) & \frac{x(1-x)}{4} f(x)\sigma^4 h + o(h) \\ \frac{x(1-x)}{4} f(x)\sigma^4 h + o(h) & o(h) \end{pmatrix} \\
&= O(\frac{1}{(n\sqrt{h})^3}). \qquad (3.13)
\end{aligned}$$

where, $p_3(x) = f(x)\sigma^4(x) + \frac{1-2x}{4}[f'(x)\sigma^4(x) + f(x)\sigma^{4'}(x)]h + \frac{x(1-x)}{8}[f(x)\sigma^{4''}(x) + 2\sigma^{4'}(x)f'(x) + \sigma^4(x)f''(x)]h$. Therefore, together with (3.12) and (3.13), $X^T\Sigma X$ corresponds to

$$\begin{aligned}
X^T\Sigma X &= \frac{1}{2n\sqrt{\pi x(1-x)h}} S^*[1+o(1)] + \frac{1}{\sqrt{(n\sqrt{h})^3}} O(1) \\
&= \frac{1}{2n\sqrt{\pi x(1-x)h}} S^*[1+o(1)]. \qquad (3.14)
\end{aligned}$$

Combined with (3.10) and (3.14), we have

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}|X) &= (X^T W X)^{-1} X^T \Sigma X (X^T W X)^{-1} \\
&= S^{-1} \frac{1}{2n\sqrt{\pi x(1-x)h}} S^* S^{-1}[1+o_p(1)] = \frac{1}{2n\sqrt{\pi x(1-x)h}} S^{-1} S^* S^{-1}[1+o_p(1)]
\end{aligned}$$

Therefore,

$$\text{Var}(\hat{m}(x)|X) = \frac{1}{2n\sqrt{\pi x(1-x)h}} e_0^T S^{-1} S^* S^{-1} e_0 + o_p(\frac{1}{n\sqrt{h}}), \tag{3.15}$$

$$\text{Var}(\hat{m}'(x)|X) = \frac{1}{2n\sqrt{\pi x(1-x)h}} e_1^T S^{-1} S^* S^{-1} e_1 + o_p(\frac{1}{nh\sqrt{h}}). \tag{3.16}$$

where $e_0 = (1,0)^T, e_1 = (0,1)^T$.

The expressions for conditional variance can be further simplified by carefully considering the matrix products in (3.15). We know that,

$$S = \begin{pmatrix} f(x) + ah & bh \\ bh & ch \end{pmatrix} \tag{3.17}$$

where, $a = (1-2x)f'(x) + \frac{1}{2}x(1-x)f''(x)$, $b = (1-2x)f(x) + x(1-x)f'(x)$, and $c = x(1-x)f(x)$. Similarly, S^* can be written as,

$$S^* = \begin{pmatrix} f(x)\sigma^2(x) + dh & eh \\ eh & gh \end{pmatrix} \tag{3.18}$$

where, $d = \frac{(1-2x)k_1(x)}{2} + \frac{x(1-x)k_2(x)}{4}$, $e = \frac{(1-2x)f(x)\sigma^2(x)}{2} + \frac{x(1-x)(f'(x)\sigma^2(x)+f(x)\sigma^{2'}(x))}{2}$ and $g = \frac{x(1-x)f(x)\sigma^2(x)}{2}$. With the modified representation of S, S^{-1} corresponds to

$$S^{-1} = \frac{1}{cf(x)h + ach^2 - b^2h^2} \begin{pmatrix} ch & -bh \\ -bh & f(x) + ah \end{pmatrix} \tag{3.19}$$

Therefore,

$$S^{-1}S^*S^{-1} = \frac{1}{(cf(x)h + (ac-b^2)h^2)^2} \begin{pmatrix} ch & -bh \\ -bh & f(x)+ah \end{pmatrix} \begin{pmatrix} f(x)\sigma^2(x)+dh & eh \\ eh & gh \end{pmatrix} \begin{pmatrix} ch & -bh \\ -bh & f(x)+ah \end{pmatrix} \tag{3.20}$$

Using this result, we can simplify,

$$e_0^T S^{-1}S^*S^{-1}e_0 = \frac{c^2 f(c)\sigma^(x) + o(h)}{[cf(X)+o(h)]^2} \to \frac{\sigma^2(x)}{f(x)} \tag{3.21}$$

and

$$\begin{aligned} e_1^T S^{-1}S^*S^{-1}e_1 &= \frac{f^2(x)g/h + 2agf(x) - f(x)be + O(h)}{c^2 f^2(x)} = \frac{f^2(x)g}{c^2 f^2(x)h} + O(1) = \frac{g}{hc^2} \\ &= \frac{x(1-x)f(x)\sigma^2(x)}{2(x(1-x)f(x))^2 h} = \frac{\sigma^2(x)}{2f(x)x(1-x)h} \end{aligned} \tag{3.22}$$

Substituting back in the conditional variance results provides the expressions in Theorem 1. Next, consider the conditional bias in (3.2), that is $(X^TWX)^{-1}X^TWr$. It is straightforward

to arrive at the following result,

$$
\begin{aligned}
X^T W r &= X^T W(\mathbf{m} - X\beta) \\
&= X^T W \begin{pmatrix} m(X_1) - (m(x) + m'(x)(X_1 - x)) \\ \vdots \\ m(X_n) - (m(x) + m'(x)(X_n - x)) \end{pmatrix} \\
&= X^T W \begin{pmatrix} \frac{m''(x)}{2}(X_1 - x)^2 + o\left((X_1 - x)^2\right) \\ \vdots \\ \frac{m''(x)}{2}(X_n - x)^2 + o\left((X_n - x)^2\right) \end{pmatrix} \\
&= \frac{m''(x)}{2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n (X_i - x)^2 K_{x/h+1,(1-x)/h+1}(X_i) \\ \frac{1}{n}\sum_{i=1}^n (X_i - x)^3 K_{x/h+1,(1-x)/h+1}(X_i) \end{pmatrix} \\
&\triangleq \frac{m''(x)}{2} \left(\frac{1}{n}\sum_{i=1}^n (X_i - x)^j K_{x/h+1,(1-x)/h+1}(X_i) \right),
\end{aligned}
$$

with $j = 2, 3$, the the expectation of the right hand side of $X^T W r$ can be showed that,

$$
\begin{aligned}
&\frac{m''(x)}{2} E\left((X_i - x)^j K_{x/h+1,(1-x)/h+1}(X_i)\right) \\
&= \frac{m''(x)}{2} \int_0^1 g(u, p, q)(u - x)^j f(u) du \\
&= \frac{m''(x)}{2} \left(l(x) + \left[(1 - 2x) l'(x) + \frac{1}{2} x(1 - x) l''(x) \right] h + o(h) \right)
\end{aligned}
$$

Based on assumptions (C.1), (C.2),

$$
E\left(X^T W r\right) = \frac{m''(x)}{2} \begin{pmatrix} x(1 - x) f(x) h + o(h) \\ o(h) \end{pmatrix} \tag{3.23}
$$

The variance of $x^T W r$ can be written as,

$$\begin{aligned}
&\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{m''(x)}{2}(X_i-x)^j K_{x/h+1,(1-x)/h+1}(X_i)\right)\\
&=\frac{1}{n}\left(\frac{m''(x)}{2}\right)^2 E\left((X_i-x)^j K_{x/h+1,(1-x)/h+1}(X_i)\right)^2\\
&-\frac{1}{n}\left(\frac{m''(x)}{2}\right)^2 \left[E\left((X_i-x)^j K_{x/h+1,(1-x)/h+1}(X_i)\right)\right]^2,
\end{aligned}$$

The leading term can be rewritten as

$$\begin{aligned}
&\frac{1}{n}\left(\frac{m''(x)}{2}\right)^2 E\left((X_i-x)^j K_{x/h+1,(1-x)/h+1}(X_i)\right)^2\\
&=\frac{1}{n}\left(\frac{m''(x)}{2}\right)^2 A_h(x)\left(l(x)+\left[(1-2x)l'(x)+\frac{1}{2}x(1-x)l''(x)\right]h+o(h)\right)
\end{aligned}$$

where, $l(u)=(u-x)^{2j}f(u)$. Based on the approximation of $A_h(x)$, by the continuity of f, f' and (C.1), (C.2),

$$\text{Var}\left(X^T W r\right)=\frac{\left(m''(x)\right)^2}{8n\sqrt{\pi x(1-x)h}}\begin{pmatrix}O(h^2)\\ O(h^2)\end{pmatrix}=O(h^{\frac{3}{2}}/n). \tag{3.24}$$

Together with (3.23) and (3.24), we have

$$\begin{aligned}
X^T W r&=\frac{x(1-x)m''(x)f(x)h}{2}\begin{pmatrix}1+o(h)\\ o(h)\end{pmatrix}+\begin{pmatrix}O(h^{\frac{3}{4}}/\sqrt{n})\\ O(h^{\frac{3}{4}}/\sqrt{n})\end{pmatrix}\\
&=\frac{xm''(x)f(x)h}{2}\left[\begin{pmatrix}1+o(h)\\ o(h)\end{pmatrix}+\begin{pmatrix}O(\frac{1}{\sqrt{n\sqrt{h}}})\\ O(\frac{1}{\sqrt{n\sqrt{h}}})\end{pmatrix}\right]\\
&=\frac{xm''(x)f(x)h}{2}\begin{pmatrix}1+o(1)\\ o(1)\end{pmatrix}
\end{aligned} \tag{3.25}$$

Combined with (3.10) and (3.25), we have

$$\begin{aligned}\text{bias}(\hat{\beta}|X) &= (X^TWX)^{-1}X^TWr \\ &= \frac{x(1-x)m''(x)f(x)h}{2}S^{-1}\begin{pmatrix}1+o(1)\\ 2h+o(1)\end{pmatrix} \triangleq \frac{xm''(x)f(x)h}{2}S^{-1}c,\end{aligned}$$

Therefore,

$$\text{bias}(\hat{m}(x)|X) = \frac{x(1-x)m''(x)h}{2} + o_p(h), \tag{3.26}$$

$$\text{bias}(\hat{m}'(x)|X) = \frac{(2x-1)m''(x)h}{2} - \frac{x(1-x)m''(x)f'(x)h}{2f(x)} + o_p(h). \tag{3.27}$$

Proof of Theorem 2

Fix an $x \in (0,1)$. To show the asymptotic normality of $\hat{m}(x)$, we use a convenient decomposition corresponding to,

$$\begin{aligned}\hat{m}(x) - m(x) &= e_0^T(X^TWX)^{-1}X^TW\left[Y - m(x)Xe_0^T - m'(x)Xe_1\right], \\ &= e_0^T(X^TWX)^{-1}X^TW\left[\mathbf{m} - m(x)Xe_0^T - m'(x)Xe_1\right] + e_0^T(X^TWX)^{-1}X^TW\varepsilon\end{aligned}$$

Similarly,

$$\begin{aligned}\hat{m}'(x) - m'(x) &= e_1^T(X^TWX)^{-1}X^TW\left[Y - m(x)Xe_0^T - m'(x)Xe_1\right], \\ &= e_1^T(X^TWX)^{-1}X^TW\left[\mathbf{m} - m(x)Xe_0^T - m'(x)Xe_1\right] + e_1^T(X^TWX)^{-1}X^TW\varepsilon\end{aligned}$$

So, to consider the asymptotics of above differences, we need to study the behavior of the following three matrices

$$X^TWX, \quad X^TW\left[Y - m(x)Xe_0^T - m'(x)Xe_1\right], \quad X^TW\varepsilon$$

First, we analyze a slightly modified version of the first matrix

$$X^T W X$$

. We can easily show that $\mathrm{diag}(1, 1/\sqrt{h}) \cdot X^T W X \cdot \mathrm{diag}(1, 1/\sqrt{h})$ by (3.7) is equal to

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i) & \frac{1}{\sqrt{h}} \cdot \frac{1}{n}\sum_{i=1}^n (X_i - x) K_{x/h+1,(1-x)/h+1}(X_i) \\ \frac{1}{\sqrt{h}} \cdot \frac{1}{n}\sum_{i=1}^n (X_i - x) K_{x/h+1,(1-x)/h+1}(X_i) & \frac{1}{h} \cdot \frac{1}{n}\sum_{i=1}^n (X_i - x)^2 K_{x/h+1,(1-x)/h+1}(X_i) \end{pmatrix},$$

Similar to the discussion in the proof of Theorem 1, one can obtain that under condition (C.1),

$$\begin{aligned} &\mathrm{diag}(1, 1/\sqrt{h}) \cdot X^T W X \cdot \mathrm{diag}(1, 1/\sqrt{h}) \\ &= \begin{pmatrix} f(x) + (1-2x)f'(x)h + \frac{x(1-x)f''(x)h}{2} & (1-2x)f(x)\sqrt{h} + x(1-x)f'(x)\sqrt{h} \\ (1-2x)f(x)\sqrt{h} + x(1-x)f'(x)\sqrt{h} & x(1-x)f(x) + f(x)h \end{pmatrix} + o_p(1) \triangleq S_0 + o_p(1). \end{aligned}$$

Hence,

$$(X^T W X)^{-1} = \mathrm{diag}(1, 1/\sqrt{h}) \cdot \big(S_0^{-1} + o_p(1)\big) \cdot \mathrm{diag}(1, 1/\sqrt{h}). \tag{3.28}$$

Next, let us look at the second matrix of interest,

$$X^T W \varepsilon$$

. Once again, considering a slightly modified form of this matrix, we obtain

$$\mathrm{diag}(1, 1/\sqrt{h}) \cdot X^T W \varepsilon = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i \\ \frac{1}{\sqrt{h}} \cdot \frac{1}{n}\sum_{i=1}^n (X_i - x) K_{x/h+1,h}(X_i)\varepsilon_i \end{pmatrix} = O_p(\frac{1}{\sqrt{n\sqrt{h}}}),$$

Combining this result with (3.28), we have

$$
\begin{aligned}
&(X^TWX)^{-1}X^TW\varepsilon = \mathrm{diag}(1, 1/\sqrt{h}) \cdot \big(S_0^{-1} + o_p(1)\big) \cdot \mathrm{diag}(1, 1/\sqrt{h}) X^TW\varepsilon \\
&= \mathrm{diag}(1, 1/\sqrt{h}) \cdot S_0^{-1} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i + o_p(\frac{1}{\sqrt{n\sqrt{h}}}) \\ \frac{1}{\sqrt{h}} \cdot \frac{1}{n}\sum_{i=1}^n (X_i - x) K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i + o_p(\frac{1}{\sqrt{n\sqrt{h}}}) \end{pmatrix} \\
&= \begin{pmatrix} O_p(\frac{1}{\sqrt{n\sqrt{h}}}) \\ O_p(\frac{1}{\sqrt{nh^{3/2}}}) \end{pmatrix},
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathrm{diag}(1, \sqrt{h}) \cdot S_0 \cdot \sqrt{\frac{n\sqrt{h}}{f(x)\sigma^2(x)}} (X^TWX)^{-1}X^TW\varepsilon \\
&= \begin{pmatrix} \sqrt{\frac{n\sqrt{h}}{f(x)\sigma^2(x)}} \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i + o_p(1) \\ \sqrt{\frac{n\sqrt{h}}{f(x)\sigma^2(x)}} \frac{1}{\sqrt{h}} \cdot \frac{1}{n}\sum_{i=1}^n (X_i - x) K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i + o_p(1) \end{pmatrix} \triangleq \begin{pmatrix} V_1(x) + o_p(1) \\ V_2(x) + o_p(1) \end{pmatrix}.
\end{aligned}
$$

We will first show that $V_1(x)$ is asymptotically normal. For this purpose, let $\eta_{in} = \sqrt{\frac{n\sqrt{h}}{f(x)\sigma^2(x)}} \cdot n^{-1} K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i$ so that $V_1(x) = \sum_{i=1}^n \eta_{in}$. Clearly, $E\eta_{in} = 0$. By assumption (C.3) on $\sigma^2(x)$, a routine argument leads to $E\eta_{in}^2 = [1/(2n\sqrt{\pi x(1-x)})][1+o(1)]$. Therefore,

$$
s_n^2 = \mathrm{Var}\Big(\sum_{i=1}^n \eta_{in}\Big) = nE\eta_{in}^2 = \frac{1}{2\sqrt{\pi x(1-x)}}[1 + o(1)].
$$

for any $\delta > 0$,

$$
E|\eta_{in}|^{2+\delta} = (n\sqrt{h})^{(2+\delta)/2} n^{-(2+\delta)} E K^{2+\delta}_{x/h+1,(1-x)/h+1}(X) E(|\varepsilon|^{2+\delta}|X = x) = O(n^{-(2+\delta)/2} h^{-\delta/4}).
$$

Then

$$
s_n^{-(2+\delta)} \sum_{i=1}^n E|\eta_{in}|^{2+\delta} = O\Big(\Big(\frac{1}{n\sqrt{h}}\Big)^{\delta/2}\Big) = o(1).
$$

Hence, by the Lyapunov central limit theorem, $s_n^{-1}V_1(x) \xrightarrow{d} N(0,1)$, that is $V_1(x) \xrightarrow{d} N(0, \frac{1}{2\sqrt{\pi x(1-x)}})$.

Also, as $nh^{\frac{3}{2}} \to \infty$, using a similar argument as in dealing with $V_1(x)$, we have $V_2(x) \xrightarrow{d} N(0, \frac{\sqrt{x(1-x)}}{4\sqrt{\pi}})$, and the covariance $\mathrm{Cov}\,(V_1(x), V_2(x)) \to 0$

By Cramér-Wald's device,

$$\begin{aligned}
&\mathrm{diag}(1, \sqrt{h}) \cdot S_0 \cdot \sqrt{\frac{n\sqrt{h}}{f(x)\sigma^2(x)}}(X^TWX)^{-1}X^TW\varepsilon \\
&\xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{2\sqrt{\pi x(1-x)}} & 0 \\ 0 & \frac{\sqrt{x(1-x)}}{4\sqrt{\pi}} \end{pmatrix}\right].
\end{aligned}$$

Next, together with (3.28), it is easy to see that our third matrix of interest, namely, $(X^TWX)^{-1}X^TW\left[\mathbf{m} - m(x)Xe_0^T - m'(x)Xe_1\right]$ equals to

$$\begin{aligned}
&\mathrm{diag}(1, 1/\sqrt{h}) \cdot \left(S_0^{-1} + o_p(1)\right) \cdot \mathrm{diag}(1, 1/\sqrt{h})X^TW\left[\mathbf{m} - m(x)Xe_0^T - m'(x)Xe_1\right] \\
&= \mathrm{diag}(1, 1/\sqrt{h}) \cdot \left(S_0^{-1} + o_p(1)\right) \cdot \\
&\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)\left[\frac{m''(x)}{2}(X_i - x)^2 + o_p\left((X_i - x)^2\right)\right] \\ \frac{1}{\sqrt{h}} \cdot \frac{1}{n}\sum_{i=1}^n (X_i - x)K_{x/h+1,(1-x)/h+1}(X_i)\left[\frac{m''(x)}{2}(X_i - x)^2 + o_p\left((X_i - x)^2\right)\right] \end{pmatrix} \\
&= \mathrm{diag}(1, 1/\sqrt{h}) \cdot \left(S_0^{-1} + o_p(1)\right) \cdot \frac{m''(x)}{2} \cdot \begin{pmatrix} x(1-x)f(x)h + o_p(h) \\ 2(1-2x)x(1-x)f(x)h^2 + o_p(h^2) \end{pmatrix} \\
&= \mathrm{diag}(1, 1/\sqrt{h}) \cdot S_0^{-1} \cdot \frac{m''(x)}{2} \begin{pmatrix} x(1-x)f(x)h + o_p(h) \\ 2(1-2x)x(1-x)f(x)h^2 + o_p(h^2) \end{pmatrix}
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\operatorname{diag}(1,\sqrt{h})\cdot S_0\cdot\sqrt{\frac{n\sqrt{h}}{f(x)\sigma^2(x)}}(X^TWX)^{-1}X^TW\left[\mathbf{m}-m(x)Xe_0^T-m'(x)Xe_1\right]\\
&=\sqrt{\frac{n\sqrt{h}}{f(x)\sigma^2(x)}}\cdot\frac{m''(x)}{2}\begin{pmatrix}x(1-x)f(x)h+o_p(h)\\2(1-2x)x(1-x)f(x)h^2+o_p(h^2)\end{pmatrix}\\
&=\frac{m''(x)}{2}\begin{pmatrix}x(1-x)\frac{\sqrt{nh^{5/2}f(x)}}{\sigma(x)}+o_p(\sqrt{nh^{5/2}})\\2(1-2x)x(1-x)\frac{\sqrt{nh^{9/2}f(x)}}{\sigma(x)}+o_p(\sqrt{nh^{9/2}})\end{pmatrix}
\end{aligned}$$

Putting all the results together along with the assumption $nh^{9/2}\to 0$, we can show that

$$\begin{aligned}
&\operatorname{diag}(1,\sqrt{h})\cdot S_0\cdot\sqrt{\frac{n\sqrt{h}}{f(x)\sigma^2(x)}}\left[\begin{pmatrix}\hat{m}(x)-m(x)\\\hat{m}'(x)-m'(x)\end{pmatrix}-\frac{m''(x)}{2}\begin{pmatrix}x(1-x)f(x)h\\0\end{pmatrix}\right]\\
&\xrightarrow{d}N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\frac{1}{2\sqrt{\pi x(1-x)}}&0\\0&\frac{x(1-x)}{4\sqrt{\pi}}\end{pmatrix}\right].
\end{aligned}$$

Proof of Theorem 3

Recall that $\hat{m}(x)-m(x)=e_0^T(X^TWX)^{-1}X^TW\left[\mathbf{m}-m(x)Xe_0^T-m'(x)Xe_1\right]+e_0^T(X^TWX)^{-1}X^TW\varepsilon\triangleq B(x)+V(x)$. To show the uniform convergence of $\hat{m}(x)$ over $[a,b]$, it suffices to show that

$$\sup_{x\in[a,b]}|B(x)|=O(h)+O\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big),\tag{3.29}$$

$$\sup_{x\in[a,b]}|V(x)|=O\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big).\tag{3.30}$$

We shall prove (3.30) as (3.29) can be proved using a similar approach.

Let α,η be such that $\alpha<2/5$, $\alpha(2+\eta)>1$ and $\alpha(1+\eta)>2/5$ and define $d_n=n^\alpha$. For

each i, write $\varepsilon_i = \varepsilon_{i,1}^{d_n} + \varepsilon_{i,2}^{d_n} + \mu_i^{d_n}$, with

$$\varepsilon_{i,1}^{d_n} = \varepsilon_i I(|\varepsilon_i| > d_n),\ \varepsilon_{i,2}^{d_n} = \varepsilon_i I(|\varepsilon_i| \le d_n) - \mu_i^{d_n},\ \mu_i^{d_n} = E[\varepsilon_i I(|\varepsilon_i| \le d_n)|X_i].$$

Note that,

$$\begin{aligned}(X^TWX)^{-1}X^TW\varepsilon &= (X^TWX)^{-1}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i \\ \frac{1}{n}\sum_{i=1}^n (X_i - x)K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_i \end{pmatrix} \\ &\triangleq (X^TWX)^{-1}\begin{pmatrix} V_{10}(x) \\ V_{20}(x) \end{pmatrix}.\end{aligned}$$

For $V_{10}(x)$, it can be decomposed as

$$V_{10}(x) = \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,1}^{d_n} + \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n} + \frac{1}{n}\sum_{i=1}^n K_{x/h+1,(1-x)/h+1}(X_i)\mu_i^{d_n}.$$

Let us first look at the third term in $V_{10}(x)$. Since $E(\varepsilon_i|X_i) = 0$, so $\mu_i^{d_n} = -E[\varepsilon_i I(|\varepsilon_i| > d_n)|X_i]$, then from assumption (C.4), the result in Chapter 2 with $EK_{x/h+1,(1-x)/h+1}(X) = f(x) + o(1)$, we have $|\mu_i^{d_n}| \le cd_n^{-(1+\eta)}$. Hence

$$\sup_{x\in[a,b]} \Big|\frac{1}{n}\sum_{i=1}^n K_{x/h+1,h}(X_i)\mu_i^{d_n}\Big| \le cd_n^{-(1+\eta)} = o\Big(\frac{1}{\sqrt{n\sqrt{h}}}\Big).$$

Next, consider the first term in $V_{10}(x)$ involving $\varepsilon_{i,1}^{d_n}$. Using Markov inequality we can write,

$$\sum_{n=1}^{\infty} P(|\varepsilon_n| > d_n) \le E|\varepsilon|^{2+\eta}\sum_{n=1}^n \frac{1}{d_n^{2+\eta}} < \infty.$$

Borel-Cantelli Lemma implies that

$$P\{\exists N, |\varepsilon_n| \le d_n \text{ for } n > N\} = 1 \Rightarrow P\{\exists N, |\varepsilon_i| \le d_n, i = 1, 2, \ldots, n, \text{ for } n > N\} = 1$$
$$\Rightarrow P\{\exists N, \varepsilon_{i,1}^{d_n} = 0, i = 1, 2, \ldots, n, \text{ for } n > N\} = 1.$$

Hence,

$$\sup_{x\in[a,b]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,h}(X_i)\varepsilon_{i,1}^{d_n}\Big| = O(n^{-k}), \quad \forall k > 0.$$

For the second term $\varepsilon_{i,2}^{d_n}$, we have $E[\varepsilon_{i,2}^{d_n}|X_i] = 0$, and it is easy to show that

$$\mathrm{Var}(\varepsilon_{i,2}^{d_n}|X_i) = \sigma^2(X_i) + O[d_n^{-\eta} + d_n^{-2(1+\eta)}]$$

and for $k \ge 2$, $E(|\varepsilon_{i,n}^{d_n}|^k|X_i) \le 2^{k-2}d_n^{k-2}E(|\varepsilon_{i,n}^{d_n}|^2|X_i)$. From the proof of uniform convergence of Beta kernel estimator in Chapter 2, $|K_{x/h+1,(1-x)/h+1}(X)| \le \frac{c}{\sqrt{x(1-x)h}}$ holds. Furthermore, the boundedness of $\sigma^2(x)$ over $(0,1)$ implies that

$$\begin{aligned} E|n^{-1}K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}|^k &= n^{-k}E[K^k_{x/h+1,(1-x)/h+1}(X_i)E(|\varepsilon_{i,n}^{d_n}|^k|X_i)] \\ &\le n^{-k}2^{k-2}d_n^{k-2}EK^k_{x/h+1,(1-x)/h+1}(X)\sigma^2(X_i) \\ &\le \Big(\frac{cd_n}{n\sqrt{h}}\Big)^{k-2}E|n^{-1}K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}|^2. \quad (3.31) \end{aligned}$$

Since,

$$\begin{aligned} E|n^{-1}K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}|^2 &= \frac{1}{n^2}E[K^2_{x/h+1,(1-x)/h+1}(X_i)\sigma^2(X_i)][1+o(1)] \\ &= \frac{f(x)\sigma^2(x)}{2n^2\sqrt{\pi x(1-x)h}}[1+o(1)], \end{aligned}$$

the random variable $n^{-1}K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}$ satisfies the Cramér condition. So, invoking

the Bernstein inequality, we can state that for all $c > 0$,

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}\Big| \geq c\sqrt{\log n}\sqrt{\sum_{i=1}^{n} E\Big[\frac{1}{n}K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}\Big]^2}\Big)$$
$$\leq 2\exp(-c^2\log n/8).$$

Take $c = 4$ and $C(x) = \frac{c\sqrt{f(x)\sigma^2(x)}}{2\sqrt{\pi x(1-x)}}$ in the above inequality to obtain

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,h}(X_i)\varepsilon_{i,2}^{d_n}\Big| \geq C(x)\sqrt{\log n/n\sqrt{h}}\Big) \leq \frac{2}{n^2},$$

Using the Borel-Cantelli Lemma and the boundedness of $f(x)\sigma^2(x)/\sqrt{x(1-x)}$ over $x \in [a,b]$, it is easy to show that for each $x \in [a,b]$,

$$\Big|\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}\Big| = O\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big).$$

To bound the sum $\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}$ uniformly for $x \in [a,b]$, we partition the closed interval $[a,b]$ by the equally spaced points x_i, $i = 0,1,\ldots,N_n$ such that $a = x_0 < x_1 < \cdots < x_{N_n} = b$ and $N_n = n^5$. It is easily seen that

$$\begin{aligned}
&\sup_{a\leq x\leq b}\Big|\frac{1}{n}\sum_{i=1}^{n} K_{x_j/h+1,(1-x_j)/h+1}(X_i)\varepsilon_{i,2}^{d_n}\Big| \\
&= \max_{0\leq j\leq N_n}\sup_{x\in[x_j,x_{j+1}]}\Big|\frac{1}{n}\sum_{i=1}^{n} K_{x_j/h+1,(1-x_j)/h+1}(X_i)\varepsilon_{i,2}^{d_n}\Big| \\
&= O\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big).
\end{aligned}$$

In the following, we will prove

$$\max_{0\leq j\leq N_n}\sup_{x\in[x_j,x_{j+1}]}\Big|\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n} - \frac{1}{n}\sum_{i=1}^{n} K_{x_j/h+1,(1-x_j)/h+1}(X_i)\varepsilon_{i,2}^{d_n}\Big| = o\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big)$$

For ease in discussion, from this point onwards, we will denote $K_{x/h+1,(1-x)/h+1}(X)$ as $K(x,X)$ understanding the dependance on h is implicit. For any $x \in [x_j, x_{j+1}]$, a Taylor expansion of $K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}$ at $x = x_j$ up to the first order leads to the following expression for the difference $K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n} - K_{x_j/h+1,(1-x_j)/h+1}(X_i)\varepsilon_{i,2}^{d_n}$:

$$|K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n} - K_{x_j/h+1,(1-x_j)/h+1}(X_i)\varepsilon_{i,2}^{d_n}| \approx (x - x_j)K'(\tilde{x}, X)\varepsilon_{i,2}^{d_n} \tag{3.32}$$

for $\tilde{x} \in [x_j, x_{j+1}]$. We will bound the difference in (3.32) by evaluating $K'(x,X)$ as follows:

$$K'(x,X) = \frac{\mathcal{A}}{B(x/h+1,(1-x)/h+1)} + X^{x/h}(1-X)^{(1-x)/h}\mathcal{B}$$

where, $\mathcal{A}$ and $\mathcal{B}$ correspond to the derivative of $X^{x/h}(1-X)^{(1-x)/h}$ and $1/B(x/h+1,(1-x)/h+1)$, respectively. We can show that,

$$\mathcal{A} = \frac{1}{h}\Big[\log X - \log(1-X)\Big]X^{x/h}(1-X)^{(1-x)/h} \tag{3.33}$$

and

$$\mathcal{B} = \frac{\psi^0((1-x)/h+1) - \psi^0(x/h+1)}{hB(x/h+1,(1-x)/h+1)} \tag{3.34}$$

where, $\psi^0(x)$ represents the digamma function. Exploiting the properties of the digamma function and some straightforward algebra, we can upper bound the numerator of (3.34) as

$$\psi^0((1-x)/h+1) - \psi^0(x/h+1) \leq \frac{1-2x+h}{x}.$$

Substituting the expressions for $\mathcal{A}$ and $\mathcal{B}$ in (3.32) we have,

$$\begin{aligned} K'(x,X) \leq\ & \frac{X^{x/h}(1-X)^{(1-x)/h}}{B(x/h+1,(1-x)/h+1)}\Big[\log X - \log(1-X)\Big] \\ & + \frac{X^{x/h}(1-X)^{(1-x)/h}}{B(x/h+1,(1-x)/h+1)}\frac{1-2x+h}{x}. \end{aligned}$$

Observing that $\frac{X^{x/h}(1-X)^{(1-x)/h}}{B(x/h+1,(1-x)/h+1)}$ corresponds to the beta density, we can bound it based on its value at its mode similar to the derivation in (2.23). Returning to the notation in (3.32), this bound can be written as,

$$K'(\tilde{x}, X) \leq \frac{ph^{-3/2}}{\sqrt{x(1-x)}} \tag{3.35}$$

for some positive constant p. Since $0 \leq x - x_j \leq (b-a)/N_n$, and $\tilde{x} > 1/a$,

$$|K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n} - K_{x_j/h+1,(1-x_j)/h+1}(X_i)\varepsilon_{i,2}^{d_n}| \leq \frac{ph^{-3/2}}{N_n}.$$

This, together with the choice $N_n = n^5$, implies that

$$\left|\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n} - \frac{1}{n}\sum_{i=1}^{n} K_{x_j/h+1,(1-x_j)/h+1}(X_i)\varepsilon_{i,2}^{d_n}\right| \leq \frac{c}{N_n h^2\sqrt{h}} = o\left(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\right).$$

Therefore, we obtain

$$\begin{aligned}
&\sup_{a\leq x\leq b} \left|\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}\right| \leq \max_{0\leq j\leq N_n} \left|\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n}\right| \\
&+ \max_{0\leq j\leq N_n-1} \sup_{x\in[x_j,x_{j+1}]} \left|\frac{1}{n}\sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i)\varepsilon_{i,2}^{d_n} - \frac{1}{n}\sum_{i=1}^{n} K_{x_j/h+1,(1-x_j)h+1}(X_i)\varepsilon_{i,2}^{d_n}\right| \\
&= O\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big).
\end{aligned}$$

This, implies $\sup_{a\leq x\leq b}|V_{10}(x)| = O\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big)$. For $V_{20}(x)$, the similar technique can be used to obtain that $\sup_{a\leq x\leq b}|V_{20}(x)| = o\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big)$. In fact, the only major difference is that,

instead of using (3.31), we should use the inequality

$$\begin{aligned} E|n^{-1}K_{x/h+1,(1-x)/h+1}(X_i)(X_i-x)\varepsilon_{i,2}^{d_n}|^k &= n^{-k}E[K^k_{x/h+1,(1-x)/h+1}(X_i)(X_i-x)E(|\varepsilon_{i,2}^{d_n}|^k|X_i)] \\ &\le n^{-k}2^{k-2}d_n^{k-2}E[K_{x/h+1,(1-x)/h+1}(X)(X_i-x)]^k\sigma^2(X_i) \\ &\le \Big(\frac{cd_n}{n\sqrt{h}}\Big)^{k-2}E|n^{-1}K_{x/h+1,(1-x)/h+1}(X_i)(X_i-x)\varepsilon_{i,2}^{d_n}|^2. \end{aligned}$$

This is based on $|K_{x/h+1,(x-1)/h+1}(X)(X-x)| \le |X-x||K_{x/h+1,(x-1)/h+1}(X)| \le \frac{c}{\sqrt{hx(1-x)}}$. Thus, we have

$$\begin{aligned} E|n^{-1}K_{x/h+1,(1-x)/h+1}(X_i)(X_i-x)\varepsilon_{i,2}^{d_n}|^2 &= \frac{1}{n^2}E[K^2_{x/h+1,(1-x)/h+1}(X_i)(X_i-x)\sigma^2(X_i)][1+o(1)] \\ &= \frac{\sqrt{hx(1-x)}f(x)\sigma^2(x)}{4n^2\sqrt{\pi}}[1+o(1)]. \end{aligned}$$

Together with $x \in [a,b]$, condition (C.1) and (3.10), it is easily show that the first row elements of $(X^TWX)^{-1}$ all bounded, hence

$$\sup_{x\in[a,b]}|V(x)| = O\Big(\frac{\sqrt{\log n}}{\sqrt{n\sqrt{h}}}\Big).$$

Chapter 4

Conclusions

Non parametric statistical learning methods and, in general, unsupervised machine learning methods are becoming increasingly popular in a world where data has become a powerful currency. With the growth in volume, velocity and variety of data, standard parametric regression models may not be always appropriate for inferencing. In many cases, the underlying relationships may be complicated and it may not be feasible or practical to hypothesize parametric forms to capture those relationships. Kernel based non parametric regression estimators present a useful alternative and are an important statistical tool that help identify the relationships between response and predictor variables in a general setup.

The efficacy of kernel based methods depend both on the kernel choice and the smoothing parameter. With insufficient smoothing, the resulting regression estimate is too rough and with excessive smoothing, important features of the underlying relationship is lost. While the choice of the kernel has been shown to have less of an effect on the quality of regression estimate, it is important to choose kernels to best match the support set of the underlying predictor variables. In the past few decades, there have been multiple efforts to quantify the properties of asymmetric kernel density and regression estimators. Unlike classic symmetric kernel based estimators, asymmetric kernels do not suffer from boundary problems.

In this dissertation, a comprehensive analysis of one such asymmetric kernel, namely

the Beta kernel based regression function estimation strategies and their asymptotic performance was characterized. Beta kernel estimates are especially suitable for investigating the distribution structure of predictor variables with compact support. In this dissertation, two types of Beta kernel based non parametric regression estimators were proposed and analyzed. First, in Chapter 2, a Nadaraya-Watson type Beta kernel estimator was introduced within the regression setup. For the first time, the asymptotic conditional bias and variance of the Beta kernel estimator were derived. Additionally, the uniform almost sure consistency and asymptotic normality of the regression estimator was proven. Some implementable bandwidth selection methodologies based on least square cross validation (LSCV) and generalized cross validation (GCV) were provided and tested using both simulation studies and a real data example. The usefulness of the Beta kernel estimation procedure was illustrated by comparing it with the Nadaraya-Watson estimator and the local linear smoother.

In Chapter 3, the focus shifted towards the study of local linear regression estimators. Specifically, the impact of transitioning from using symmetric kernel functions in local linear regression estimators to asymmetric kernel functions was quantified using the Beta kernel based local linear regression estimator as an example. Once again, the asymptotic conditional bias and variance were derived for local linear regression with Beta kernels. This was followed by a rigorous analysis of the asymptotic distribution and uniform almost sure consistency of the estimator. A comparison with the traditional normal kernel based local linear regression estimator using simulated and real life data examples demonstrated the value of employing asymmetric kernels in local linear regression.

Although the proposed Beta kernel and Beta local linear estimators have certain merits, we are not expecting that they can be superior to other competitors in all aspects, such as the classic Nadaraya-Watson type estimators, local linear estimators and Gamma kernel estimators. The significance of the proposed methods in this dissertation is to provide another alternative to estimate regression functions which are supported on an interval. In

real application, collectively using all the available kernel estimators might provide more insight on the structures of the data. For the sake of completeness, the following table summarizes the asymptotic biases and variances of some commonly used nonparametric estimators, as well as the Beta kernel and the Beta local linear estimators.

Estimator	Bias	Variance
NW kernel	$(\frac{m'(x)f'(x)\mu_2(K)+m''(x)\mu_2(K)f(x)}{2f(x)})h^2$	$\frac{\sigma^2(x)R(K)}{nhf(x)}$
Local Linear	$(\frac{m''(x)\mu_2(K)}{2})h^2$	$\frac{\sigma^2(x)R(K)}{nhf(x)}$
Gamma Kernel	$m'(x)h + \frac{xm''(x)h}{2} + \frac{xm'(x)f'(x)h}{f(x)}$	$\frac{\sigma^2(x)}{2nf(x)\sqrt{\pi xh}}$
Beta Kernel	$(1-2x)m'(x)h + \frac{x(1-x)m''(x)h}{2} + \frac{x(1-x)m'(x)f'(x)h}{f(x)}$	$\frac{\sigma^2(x)}{2nf(x)\sqrt{\pi x(1-x)h}}$
Beta Local Linear	$\frac{x(1-x)m''(x)h}{2}$	$\frac{\sigma^2(x)}{2nf(x)\sqrt{\pi x(1-x)h}}$

In summary, the fundamental results and the associated numerical simulations shed light, for the first time, on the use of Beta kernel based methods in non-parametric regression applications. While this dissertation takes the first and necessary steps in this direction, there are many follow on efforts that can feed off this work:

- While we have provided some easily implementable bandwidth selection procedures in Chapter 2 and have used them in Chapter 3 as well, it would be useful to analytically evaluate optimal bandwidth choices for Beta kernel based regression strategies. Theoretical derivation of the optimal bandwidth based on minimizing the asymptotic MSE has been done for traditional kernel estimators and similar approaches can be used to accomplish this task.

- Since we are studying the impact of using asymmetric kernels within the regression setup, it would be helpful to define and evaluate a new metric to measure efficiency. Traditionally, for kernel based density estimation, efficiency of a kernel K relative to K^* represents the ratio of sample sizes necessary to obtain the same minimum AMISE for a given density f. For example, if K has an efficiency of 0.90 - this indicates that the density estimate with an optimal (with respect to AMISE) kernel K^* can achieve

the same minimum AMISE using 90% of the data used by kernel K. Within the family of asymmetric kernels, defining a similar measure of efficiency along with an elegant method to compute it will be helpful in comparing asymmetric kernels.

Bibliography

[Altman(1992)] Altman, N. S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46 (3), 175–185.

[Buja(1989)] Buja, A. e. a., 1989. Linear smoothers and additive models. The Annals of Statistics, 453–510.

[Chaubey(2012)] Chaubey, Y. e. a., 2012. A new smooth density estimator for non-negative density etimator. J. Indian Statistical Association 50, 83–104.

[Chen(1999)] Chen, S. X., 1999. Beta kernel estimators for density functions. Computational Statistics and Data Analysis 31 (2), 131 – 145.
URL http://www.sciencedirect.com/science/article/pii/S0167947399000109

[Chen(2000a)] Chen, S. X., 2000a. Beta kernel smoothers for regression curves. Statistica Sinica, 73–91.

[Chen(2000b)] Chen, S. X., 2000b. Probability density function estimation using gamma kernels. Annals of the Institute of Statistical Mathematics 52 (3), 471–480.

[Chen(2002)] Chen, S. X., 2002. Local linear smoothers using asymmetric kernels. Annals of the Institute of Statistical Mathematics 54 (2), 312–323.

[Cowling and Hall(1996)] Cowling, A., Hall, P., 1996. On pseudodata methods for removing boundary effects in kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological), 551–563.

[Fan(1992)] Fan, J., 1992. Design-adaptive nonparametric regression. Journal of the American statistical Association 87 (420), 998–1004.

[Fan(1993)] Fan, J., 1993. Local linear regression smoothers and their minimax efficiencies. The Annals of Statistics, 196–216.

[Fan and Gijbels(1992)] Fan, J., Gijbels, I., 1992. Variable bandwidth and local linear regression smoothers. The Annals of Statistics, 2008–2036.

[Gasser and Müller(1979)] Gasser, T., Müller, H.-G., 1979. Kernel estimation of regression functions. In: Smoothing techniques for curve estimation. Springer, pp. 23–68.

[Hastie and Loader(1993)] Hastie, T., Loader, C., 1993. Local regression: Automatic kernel carpentry. Statistical Science, 120–129.

[Hille and Phillips(1996)] Hille, E., Phillips, R. S., 1996. Functional analysis and semigroups. Vol. 31. American Mathematical Soc.

[John(1984)] John, R., 1984. Boundary modification for kernel regression. Communications in statistics-Theory and methods 13 (7), 893–900.

[Karlin and Studden(1966)] Karlin, S., Studden, W. J., 1966. Tchebycheff systems: With applications in analysis and statistics. Interscience New York.

[Koul and Song(2013)] Koul, H. L., Song, W., 2013. Large sample results for varying kernel regression estimates. Journal of Nonparametric Statistics 25 (4), 829–853.

[Marron and Ruppert(1994)] Marron, J. S., Ruppert, D., 1994. Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological), 653–671.

[Mnatsakanov and Sarkisian(2012)] Mnatsakanov, R., Sarkisian, K., 2012. Varying kernel density estimation on r+. Statistics & probability letters 82 (7), 1337–1345.

[Müller(1991)] Müller, H.-G., 1991. Smooth optimum kernel estimators near endpoints. Biometrika 78 (3), 521–530.

[Müller and Wang(2007)] Müller, H.-G., Wang, J.-L., 2007. Density and failure rate estimation. Encyclopedia of statistics in quality and reliability.

[Nadaraya(1964)] Nadaraya, E. A., 1964. On estimating regression. Theory of Probability & Its Applications 9 (1), 141–142.

[Priestley and Chao(1972)] Priestley, M., Chao, M., 1972. Non-parametric function fitting. Journal of the Royal Statistical Society. Series B (Methodological), 385–392.

[Scaillet(2004)] Scaillet, O., 2004. Density estimation using inverse and reciprocal inverse gaussian kernels. Nonparametric statistics 16 (1-2), 217–226.

[Schuster(1985)] Schuster, E. F., 1985. Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics-Theory and methods 14 (5), 1123–1136.

[Shi and Song(2015)] Shi, J., Song, W., 2015. Asymptotic results in gamma kernel regression. Communications in Statistics - Theory and Methods 45 (12), 3489–3509.

[Silverman(1986)] Silverman, B. W., 1986. Density estimation for statistics and data analysis. Vol. 26. CRC press.

[Turlach et al.(1993)] Turlach, B. A., et al., 1993. Bandwidth selection in kernel density estimation: A review. Université catholique de Louvain.

[Wand and Jones(1995)] Wand, M. P., Jones, M. C., 1995. Kernel smoothing. Crc Press.

[Weisberg(2005)] Weisberg, S., 2005. Applied linear regression. Vol. 528. John Wiley & Sons.