UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Topics in Transformation-based Statistical Methods

Permalink

https://escholarship.org/uc/item/9wh9m255

Author

Wang, Liang

Publication Date

2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Topics in Transformation-based Statistical Methods

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics

by

Liang Wang

Committee in charge:

Professor Dimitris Politis, Chair
Professor Jelena Bradic
Professor Anthony Gamst
Professor Yixiao Sun
Professor Ronghui Xu

2015

Copyright
Liang Wang, 2015
All rights reserved.

The dissertation of Liang Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2015

DEDICATION

To my family, the coming baby, and friends

EPIGRAPH

The fear of the LORD is the beginning of wisdom;
and the knowledge of the holy is understanding.
—Proverbs 9:10

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge to my academic advisor Professor Politis as the chair of my committee. I'm very grateful for the his patience, guidance and dedication over the past four years. His wisdom and experience consistently lead me to the right path. I would also like to thank all other professors in my committee, including Ronghui Xu, Jelena Bradic, Yixiao Sun and Anthony Gamst. I would also want to acknowledge to Naresh Bhavaraju and Apurv Kamath from Dexcom, Inc., for guiding me in both academic and industry. Special thanks to Dexcom, Inc., for financially supporting me in the past six months. Additional thanks to Wilson Cheung and Srinjoy Das for their help while completing my thesis. I would finally acknowledge everyone who has contributed to the LaTeX style files from which is derived.

Finally, I thank my family, especially my wife Fan Yang and my mother Zhen Dong, for supporting me and teaching me many life lessions, that helps me to continue my work and finish my Ph.D. degree.

Chapter 1, is currently being prepared for the following paper prepared for future submission for publication: Asymptotic Validity of Bootstrap Confidence Intervals in Nonparametric Regression without an Additive Model. Liang Wang; Dimitris Politis. The dissertation author was the primary investigator and author of this material.

Chapter 2, is currently being prepared for the following paper prepared for future submission for publication: Bias Reduction by Transformed Flat-top Kernel Estimator of Density on Compact Support. Liang Wang; Dimitris Politis. The dissertation author was the primary investigator and author of this material.

Chapter 3, is currently being prepared for the following paper prepared for future submission for publication: Bayesian Method for Validation of Continuous Glucose Monitoring System Upgrade. Liang Wang; Naresh Bhavaraju. The dissertation author was the primary investigator and author of this material.

VITA

2009	B. S. in Mathematics, Nankai University, China
2009-2011	M.S. in Statistics, University of California, San Diego
2011-2015	Graduate Teaching Assistant, University of California, San Diego
2015	Ph. D. in Mathematics, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Topics in Transformation-based Statistical Methods

by

Liang Wang

Doctor of Philosophy in Mathematics

University of California, San Diego, 2015

Professor Dimitris Politis, Chair

This thesis is concerned with transformation-based statistical methods in different three areas. For problem of nonparametric regression, the transformed-based technique is called Model-free bootstrap method; for density nonparametric estimation we propose transformed flat-top series estimator; for Bayesian hypothesis testing, we employ a parametric transformation to help in evaluating the system upgrade of glucose monitoring device.

First, we establish the theory of the asymptotic validity of Model-free method on construction of confidence interval under specified assumptions. The spirit of Model-free method is transforming non-i.i.d. data to i.i.d data and then bootstrapping the new i.i.d. data. We also conduct simulations to check the

finite sample properties of Model-free estimators, compared to regular normal approximation and local bootstrap method.

Next, we address the problem of nonparametric estimation of a smooth univariate density on compact support. If the density function has compact support and is non-zero at either boundary, regular kernel estimator will be seriously biased. A lot of bias correction methods were proposed to improve the bias on the boundary. In this chapter we propose the transformed flat-top series estimator, which keeps the same bias order as existing methods at boundary, and improves the bias in the interior region of the support to higher order. Theoretical analysis and simulations are provided, and the results are generally better than corresponding results of many other kernel density estimator with boundary correction.

At last we propose modified Bayesian hypothesis method that can be assessed by type I&II errors and an alternative assessment of standard Bayesian hypothesis testing method, particularly suited to situations where modifications are made to continuous glucose monitoring (CGM) systems already approved. A parametric transformation of observations is the key to test the validation of this system upgrade. Simulations are conducted to assess the risks and benefits of the approach, which show that by careful planning and analysis prospective study sizes can be reduced and better decisions can be made on the effectiveness and safety of the modified systems.

Chapter 1

Asymptotic Validity of Bootstrap Confidence Intervals in Nonparametric Regression without an Additive Model

1.1 Introduction

Consider the regression data $\{(Y_i, X_i)\}, i = 1, \ldots, n$ which are i.i.d. pairs. For simplicity, we suppose that X_i is univariate and random with density function $f(x)$, but the method works in the same way with multivariate X_i. Then the regular nonparametric regression with an additive model is:

$$Y_i = m(X_i) + \sigma \cdot \epsilon_i, \quad i = 1, \ldots, n$$

or more generally

$$Y_i = m(X_i) + \sigma(X_i) \cdot \epsilon_i, \quad i = 1, \ldots, n \tag{1.1.1}$$

Where ϵ_i's are i.i.d. random variables with mean 0 and variance 1. $m(x) = \mathrm{E}(Y|X = x)$ is the target of interest, and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$. In this chapter,

we employ Nadaraya-Watson (N-W) kernel estimator [1]:

$$\hat{m}_{n,h}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \tag{1.1.2}$$

N-W kernel estimator dates back to 1964, and its theories were developed in the following decades, see [2, 3, 4, 5]. The conditional cumulative distribution function (CDF) and quantile estimator based on N-W estimator and their theories were studied by Li&Racine [6] and Qu&Yoon [7]. Since Efron [8] proposed bootstrap resampling method in 1979 (or see review [9]), many resampling methods on nonparametric regression problem were developed. Hardle and Bowman [10] proposed residual bootstrap method, Hall introduce Pivotal bootstrap method [11]. All these methods are very powerful on additive model. On the other side, a non-additive model violates (1.1.1), one special example is

$$Y_i = m(X_i) + \sigma(X_i) \cdot e(X_i, \epsilon_i), i = 1, \ldots, n \tag{1.1.3}$$

Where $\mathrm{E}[e(x,\epsilon)|X = x] = 0, \mathrm{Var}[e(x,\epsilon)|X = x] = 1$, more importantly, $e(x,\epsilon)$ is not additive function (neither is $\log[e(x,\epsilon)]$). See the non-additive model used in section 1.7. In 1991, local bootstrap designed for heteroscedastic model, was first presented by Shi [12], based on Priestley&Chao Estimator [13], which is consistent but not proved. Since then, more general local bootstrap method was also developed on time series problem, see [14, 15, 16]. In 2010, Politis proposed Model-free bootstraping method [17], and developed by Sperlich [18], Politis [19, 20, 21], Pan and Politis [22]. Model-free bootstrap method has been applied in nonparametric regression problem without an additive model for confidence interval and in time series for predictive interval. Since local bootstrap and Model-free bootstrap are both suitable for non-additive model problem but lack of theoretical support, this chapter provides theoretical analysis, and compares the two methods in simulation. We also reveal the connection between these two methods under specified assumptions. In the followings, we introduce the spirit of local bootstrap and model free method briefly; more details can be found in section 1.3 and section 1.4. Let $m(x)$ be $\mathrm{E}(Y|x)$ and $\hat{m}(x)$ be N-W estimator.

(i) Local bootstrap: Instead of resampling data i.i.d. as other bootstrap method, local bootstrap resamples data independently but non-identically. More specifically, we estimate the conditional probability mass function (p.m.f.) of $Y|x$ based on N-W estimator first, and then in each cycle of bootstrap, resample one of Y_i's at different X_i by the p.m.f. estimator. Denote the bootstrap data at X_i by Y_i^*

(ii) Model free: The idea of model free is that we transform the non-i.i.d. data to i.i.d. data first, and then do the resampling i.i.d.. More particularly, by probability integral transform theory, we can transform each observation Y_i to uniform $[0,1]$ random variable U_i, and then U_i's are i.i.d. random variables. After we obtain bootstrap datasets $U_i^\star$ from resampling U_i's, use quantile function of $Y|x$ to transform $U_i^\star$'s back, and denote this final bootstrap data at different X_i by $Y_i^\star$.

(iii) Construct bootstrap estimation $\hat{m}^*(x)$ by bootstrap data $\{Y_i^*\}_1^n$ (or $\{Y_i^\star\}_{i=1}^n$) and $\{X_i\}_{i=1}^n$. Iterate the whole procedure B (e.g. 1000) times, then the distribution of $\hat{m}^*(x)$ is the estimator of distribution of $\hat{m}(x)$, which leads to confidence interval.

The rest chapter is organized as followed: In section 1.2, we address all the assumptions for the theoretical analysis in this chapter and some previous work on N-W estimator. section 1.3 and 1.4 introduce local bootstrap and model free bootstrap more specifically and necessary lemmas for final proof. Main results and theorems are presented in section 1.5. In section 1.6 we discuss the details of parameter setup, the choice of transformed function and selection of bandwidth. Some simulation results are presented in section 1.7. All the proofs are in Appendix.

1.2 Assumptions

This chapter is motivated to prove validity of Model-free method for both additive model and non-additive model. Recall N-W estimator (1.1.2), and $f(x)$

is density function of $X_1, \ldots, X_n$, $m(x) = \mathrm{E}(Y|X = x)$, let $f_{Y|X}(y|x)$ be the conditional density function of Y given $X = x$, $D(y|x)$ is the conditional CDF of Y given $X = x$. Now we address the following assumptions:

Assumption(i). $\{X_i, Y_i\}, i = 1, \ldots, n$ are i.i.d pairs.

Assumption(ii). $f(x)$ and $m(x)$ are both twice differentiable, and their the derivative functions all satisfy the Lipschitz condition $|g(x) - g(y)| \leq C|x - y|$ for some $C > 0$. (where $g(\cdot) = m''\text{or}f''$). Moreover, $\sigma^2(x) = \mathrm{Var}(Y_i|X_i = x)$ is a continuous function.

Assumption(iii). $n \to \infty, h \to 0, nh \to \infty$.

Assumption(iv). The Kernel function $K(\cdot)$ is symmetric, bounded, Lipschitz continuous, twice differentiable function, and satisfies $K(x) \geq 0$, $\int K(x)dx = 1$, $\int x^2 K(x)dx < \infty$. Let the Lipschitz constant be C_1.

Assumption(v). $\mathrm{E}\left(|Y|^3|X = x\right) < \infty$. $f_{Y|X}(y|x)$ is twice differentiable with respect to x. Moreover, there exists $B_1(y) > 0, B_2(y) > 0$, that $\frac{\partial f(y|x)}{\partial x} \leq B_1(y)$, $\frac{\partial^2 f(y|x)}{\partial x^2} \leq B_2(y)$, and $\int B_i(y)dy < \infty, i = 1, 2$

Assumption(vi). $\mathrm{Var}\left(|Y_i|^3|X_i = x\right) < \infty$, where $g(\cdot)$ is the same function in assumption (v).

Assumption(vii). $\inf_{x \in \mathcal{S}} f(x) \geq \delta > 0$, where $\mathcal{S}$ is a compact subset of $\mathbb{R}$ that excludes the boundary of the support of X. And $D(y|x)$ is twice differentiable with respect to both y and x in $\mathcal{S}$.

Assumption(viii). The support of X is the compact set $\mathcal{S}$ in assumption (vii).

Assumption(ix).$nh = o(n^{\frac{4}{5}})$ (or $h = o(n^{-\frac{1}{5}})$).

Li and Racine [5] proved the following theorems under assumptions (i)-(iv):

$$\mathrm{E}[\hat{m}_{n,h}(x)] = m(x) + h^2 B_s(x) + O(h^3) \tag{1.2.1}$$

$$\mathrm{Var}[\hat{m}_{n,h}(x)] = \frac{1}{nh}\frac{\kappa\sigma^2(x)}{f(x)} + O\Big(\frac{h}{n}\Big) \tag{1.2.2}$$

$$\sqrt{nh}(\hat{m}_{n,h}(x) - m(x) - h^2 B_s(x)) \xrightarrow{d} N\left(0, \frac{\kappa\sigma^2(x)}{f(x)}\right) \tag{1.2.3}$$

Where

$$B_s(x) = \frac{\kappa_2}{2}\frac{2f'(x)m'(x) + f(x)m''(x)}{f(x)} \tag{1.2.4}$$

Here $\kappa_2 = \int v^2 K(v)dv, \kappa = \int K^2(v)dv$. Notice again that this theorem is valid for i.i.d pairs, which implies for both additive model (1.1.1) and non-additive model (1.1.3).

Remark 1.2.1. For the proof of bootstrap of i.i.d. pairs of data in the following sections, we need slightly stronger assumption than assumption (ii) for $Y|X = x$, i.e. not only $m(x) = \mathrm{E}(Y|X = x)$ and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$, but also the smoothness of $f_{Y|X}(y|x)$ with respect to x. Therefore the assumption(v) is very necessary.

Remark 1.2.2. McMurry and Politis [23] proposed infinite order kernel method for nonparametric regression. Unfortunately, flat-top kernel function violates the assumption (iv) and is not theoretical analyzed in this chapter.

1.3 Local Bootstrap

Shi [12] proposed local bootstrap in 1991 and assumed x_i's are design points in $[0, 1]$. In this chapter, we discuss more general situation that X_i are random with N-W estimator. The procedure of local bootstrap is as following:

(i) For each observation X_i, we have estimator of p.m.f.:

$$\widehat{G}_{X_i} : \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_n \\ w_{1i} & w_{2i} & \cdots & w_{ni} \end{pmatrix} \tag{1.3.1}$$

Where $w_{ji} = \dfrac{K\left(\frac{X_j - X_i}{h}\right)}{\sum_{l=1}^{n} K\left(\frac{X_l - X_i}{h}\right)}$ for $i, j = 1, 2, \cdots, n$, so $\sum_j w_{ji} = 1$.

(ii) At each X_i, we resample Y_i^* from distribution $\widehat{G}_{X_i}$.

(iii) Compute bootstrap kernel estimator:

$$\hat{m}^*_{n,h}(x) = \frac{\sum_{i=1}^n Y_i^* K\Big(\frac{X_i - x}{h}\Big)}{\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big)} \tag{1.3.2}$$

Shi did not discuss the properties of $\hat{m}^*_{n,h}(x)$, which is addressed in section 1.5. To obtain further properties, the following lemma is needed:

Lemma 1.3.1. *Under assumptions (i)-(v),*

$$\mathrm{E}_{X_i,Y_i}\Big[g(Y_i)K^l\Big(\frac{X_i - x}{h}\Big)\Big] = h\Big\{A(x)\mathrm{E}[g(Y)|X = x] + O(h^2)\Big\} \tag{1.3.3}$$

where $g(\cdot)$ is a continuous function, in this chapter, $g(x) = x^k$ or $g(x) = |x|^k, k = 1, 2, 3$, $A(x) = f_X(x)\int K^l(u)du$, and l is a fixed number, $f_X(x)$ is density function of $X_1, \ldots, X_n$. Also, E_{X_i,Y_i} denote the joint expectation of $(X_i.Y_i)$.

Notice i is removed because all (X_i, Y_i) are integrated, it implies $O(h^2)$ here is not related to i anymore. This lemma also works for $Y^*|X = x$, that is under the same assumptions, with $K(\cdot)$ needs to be twice differentiable in assumption(iv):

$$\mathrm{E}_{X_i,Y_i^*}\Big[g(Y_i^*)K^l\Big(\frac{X_i - x}{h}\Big)\Big] = h\Big\{A(x)\mathrm{E}[g(Y^*)|X = x] + O(h^2)\Big\} \tag{1.3.4}$$

Lemma 1.3.1 also leads to the following preliminary result:

Theorem 1.3.2. *Under the assumptions (i)-(vi) we have:*

$$|\mathrm{E}^*[g(Y^*)|X = x] - \mathrm{E}[g(Y)|X = x]| = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big) \tag{1.3.5}$$

where $g(\cdot)$ is the same continuous function in lemma 1.3.1.

This theorem is essentially a more general version of theorem 3.2 in Shi [12] for Priestley&Chao Estimator [13], and it holds for Nadaraya-Waston estimator with non-additive model here.

1.4 Model free

Remember that the spirit of Model free is transforming non-i.i.d. data to i.i.d. data. For both additive and non-additive regression models, the key is probability integral transform theorem. More specifically, if the conditional distribution function $D_x(\cdot)$ of Y is

$$D_x(y) = P(Y \leq y | X = x) \tag{1.4.1}$$

In this chapter, $D_x(y)$ is assumed to be absolutely continuous w.r.t y, then $D_x(Y)$ follows Uniform $[0, 1]$. However, since the true conditional distribution $D_x(y)$ is unknown, so Politis [17] introduced the kernel estimators of conditional CDF:

$$\widehat{D}_x(y) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq y\}} K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \tag{1.4.2}$$

$\widehat{D}_x(y)$ has some good properties which have already been showed by Li and Racine [5]. Nevertheless, it has a lot of discontinuities, makes itself unqualified as an estimator in this chapter. To solve this issue, Politis [17] used an alternative option, the smooth kernel estimator of conditional CDF:

$$\widetilde{D}_x(y) = \frac{\sum_{i=1}^{n} \Lambda\left(\frac{Y_i - y}{h_0}\right) K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \tag{1.4.3}$$

Where $\Lambda(\cdot)$ is an absolutely continuous and strictly increasing CDF over its support. We also define:

$$\widehat{D}_x^{-1}(u) = \min_{i=1,\ldots,n} \{Y_i : u \leq \widehat{D}_x(Y_i)\} \tag{1.4.4}$$

and

$$\widetilde{D}_x^{-1}(u) = \inf_y \{y : u \leq \widetilde{D}_x(y)\} \tag{1.4.5}$$

What are the transform-back functions. Politis recommended to use $\widehat{D}_X^{-1}(\cdot)$ in [19] and $\widetilde{D}_x^{-1}(\cdot)$ in [17], which will be discussed below in section 1.6, but notice that all the proofs of this chapter are built on $\widehat{D}_x^{-1}(\cdot)$. In this section, we denote model free bootstrap data by $Y^\star$, and local bootstrap data by Y^*. Politis [17] proposed 3 ways to generate Model free bootstrap data $Y_1^\star, \cdots, Y_n^\star$:

Method 1. Get transformed data $U_1, \cdots, U_n$ by computing $U_i = \widetilde{D}_{X_i}(Y_i)$. Then re-sample randomly from the transformed data $U_1, \cdots, U_n$ to get bootstrap pseudo-data $u_1^\star, \cdots, u_n^\star$. Finally $Y_i^\star = \widehat{D}_{x_i}^{-1}(u_i^\star)$.

Method 2. Generate bootstrap pseudo-data $u_1^\star, \cdots, u_n^\star$ i.i.d. from Uniform(0,1) distribution, then $Y_i^\star = \hat{D}_{x_i}^{-1}(u_i^\star)$.

Method 3. Let $\widetilde{D}_x^{(t)}$ denote the estimator $\widetilde{D}_x(\cdot)$ as computed from the delete-X_t, Y_t dataset, i.e., $\{(Y_i, X_i, i = 1, \ldots, t-1, t+1, \ldots, n\}$. Now let

$$u_t^{(t)} = \widetilde{D}_{x_t}^{(t)}(Y_t) \quad \text{for } t = 1, \ldots, n$$

And we will sample randomly from predictive u-data $U_1^{(1)}, \ldots, U_n^{(n)}$ to create bootstrap pseudo-data $u_1^\star, \ldots, u_n^\star$, which leads to $Y_i^\star = \widehat{D}_{x_i}^{-1}(u_i^\star)$.

Remark 1.4.1. In the rest of paper, we follow the same notation given by Politis [24], and denote method 1 as Model-free (MF), method 2 as Limit Model-free (LMF), method 3 as Predictive Model-free (PMF). In addition, at X_i, if $\widehat{D}_x^{-1}(\cdot)$ is applied to transform $u_i^\star$ to bootstrap data $Y_i^\star$, which is a discrete step function, then it implies we resample $Y_i^\star$ from Y_i from a p.m.f. $\widehat{H}_{X_i}$. This function $\widehat{H}_{X_i}$ assigns different weights to $Y_1, \ldots, Y_n$.

Let x_f is the point of interest, recall N-W estimator $\hat{m}_{n,h}(x_f)$ (1.1.2), and denote $\check{m}_{n,h}(x_f)$ the estimator proposed by Politis [17].

$$\check{m}_{n,h}(x_f) = \frac{1}{n}\sum_{i=1}^{n} \widehat{D}_{x_f}^{-1}(u_i)$$

Notice that

$$\begin{aligned}\hat{m}_{n,h}(x_f) &= \int y \widehat{D}_{x_f}(y)dy = \int_0^1 \widehat{D}_{x_f}^{-1}(u)du \\ &= \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}\end{aligned}$$

It implies $\check{m}_{n,h}(x_f)$ and $\hat{m}_{n,h}(x_f)$ are consistent with large sample size. Then the resampling algorithm for Model-free confidence intervals for $m(x_f)$ is:

(a) Pick one Model-free method from MF, LMF and PMF, and obtain bootstrap data $u_1^\star, \ldots, u_n^\star$ and $Y_1^\star, \cdots, Y_n^\star$

(b) Based on the pseudo data $\{(Y_i^\star, x_i), i = 1, \ldots, n\}$, re-estimate the conditional CDF $D_x(\cdot)$; Denote the bootstrap estimates by $\widehat{D}_x^\star(\cdot)$ and $\widetilde{D}_x^\star(\cdot)$.

(c) Calculate a replicate of the bootstrap confidence interval root: $\hat{m}_{n,h}(x_f) - \hat{m}_{n,h}^\star(x_f)$, or $\check{m}_{n,h}(x_f) - \check{m}_{n,h}^\star(x_f)$, where

$$\hat{m}_{n,h}^\star(x_f) = \frac{\sum_{i=1}^n Y_i^\star K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} \tag{1.4.6}$$

and

$$\check{m}_{n,h}^\star(x_f) = \frac{1}{n}\sum_{i=1}^n \widehat{D}_{x_f}^{\star -1}(u_i^\star) \tag{1.4.7}$$

(d) Steps (a)-(c) in the above are repeated B times, and the B bootstrap root replicates are collected in the form of an empirical distribution with α-quantile denoted by $q(\alpha)$.

(e) Then, the Model-free $(1-\alpha)100\%$ equal-tailed confidence interval for $m(x_f)$ is

$$[\hat{m}_{n,h}(x_f) + q(\alpha/2), \hat{m}_{n,h}(x_f) + q(1-\alpha/2)]$$

The following lemmas and theorems reveal MF, LMF and PMF method has the rate of convergence to true $m(x)$ in probability same as local bootstrap.

Theorem 1.4.2. *LMF resampling method and local bootstrap resampling method are equivalent. More specifically, at X_i, recall $\widehat{G}_{X_i}$ (1.3.1) for locall bootstrap and $\widehat{H}_{X_i}$ introduce in Remark 1.4.1. Then for LMF, $\widehat{H}_{X_i} = \widehat{G}_{X_i}$.*

Notice that this theorem only shows LMF and local bootstrap has equivalent re-sampling method. But if the statistic are different, see (1.4.6) and (1.4.7), the result will have finite sample difference. Such difference will be relieved with a large sample size.

Lemma 1.4.3. *In method 1, under assumptions (i)-(vi), if If we claim $U_1, \cdots, U_n$ are random observations from* Unif(0,1)*, then a similar result as Lemma 1.3.2 holds, that is:*

$$|\mathrm{E}^\star[g(Y^\star)|X = x] - \mathrm{E}[g(Y)|X = x]| = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big) \tag{1.4.8}$$

where $g(\cdot)$ is the same continuous function in lemma 1.3.1.

The lemma above can be easily derived from lemma 1.4.2.

Lemma 1.4.4. *Under assumptions (i)-(viii), we have:*

$$\sup_{x\in\mathcal{S}} |\widetilde{D}_x(y) - D_x(y)| = O\Big(h_0^2 + h^2 + \Big[\frac{\ln(n)}{nh}\Big]^{\frac{1}{2}}\Big) \quad \text{a.s.}$$

Where $D_x(\cdot)$ is the real conditional CDF of Y, and $\mathcal{S}$ is a compact set.

Then, Polya Theorem could guarantee us that:

Corollary 1.4.5. *Under assumptions (i)-(viii), we have:*

$$\sup_{x\in\mathcal{S}} \sup_{y} |\widehat{D}_x(y) - D_x(y)| = o(1) \text{ a.s.}$$

Remark 1.4.6. Notice the above corollary holds for both MF and PMF. It helps our main theorems to support both of them.

Lemma 1.4.7. *Let $f_n : \mathbb{R}^{2n} \to \mathbb{R}$, Z_n is $2n$ dimensional random vector and its range is a subspace of $\mathbb{R}^{2n}$. If we know that as $n \to \infty$*

$$|f_n(Z_n) - U| \to 0 \text{ a.s.}$$

where U is a random variable, then

$$|g_n[Z_n, f_n(Z_n)] - g_n(Z_n, U)| \to 0 \text{ a.s.}$$

as $n \to \infty$, where g_n is a continuous function of both two arguments.

Theorem 1.4.8. *If we use method 2, under assumptions (i)-(viii):*

$$|\mathrm{E}^\star[g(Y^\star)|X = x] - \mathrm{E}[g(Y)|X = x]| = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)$$

where $g(\cdot)$ is the same continuous function in lemma 1.3.1.

Remark 1.4.9. Some proofs are inspired by [4, 25, 26, 27]. The theorems above are the keys to prove the convergence of all three Model-free methods in the next section, by setting local bootstrap as a transition in the middle, which is also helpful to understand the connection between local bootstrap and model free methods. If we use $\widetilde{D}_x(y)$ instead, it can be seen as a generalized smooth local bootstrap.

1.5 Main Result

From now on, denote:

$$\widetilde{K}\Big(\frac{X_i - x}{h}\Big) = \frac{K\Big(\frac{X_i - x}{h}\Big)}{\frac{1}{nh}\sum_{j=1}^{n} K\Big(\frac{X_j - x}{h}\Big)} = \frac{K\Big(\frac{X_i - x}{h}\Big)}{\hat{f}_X(x)} = \frac{K\Big(\frac{X_i - x}{h}\Big)}{f_X(x) + o_p(1)} \tag{1.5.1}$$

Where $f_X(x)$ is the density function of $X_1, \ldots, X_n$ and $\hat{f}_X(x)$ is its kernel estimator.

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{j=1}^{n} K\Big(\frac{X_j - x}{h}\Big)$$

Therefore $\widetilde{K}\Big(\frac{X_i - x}{h}\Big)$ is bounded and satisfy all the condition of K in probability. So (1.1.2) can be written as:

$$\hat{m}_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} Y_i \widetilde{K}\Big(\frac{X_i - x}{h}\Big) \tag{1.5.2}$$

And Lemma 1.3.1 can be easily extended.

Lemma 1.5.1.

$$\mathrm{E}_{X_i,Y_i}\Big[g(Y_i)\widetilde{K}^l\Big(\frac{X_i - x}{h}\Big)\Big] = h\Big\{\widetilde{A}(x)\mathrm{E}[g(Y)|X = x] + O(h^2)\Big\} \tag{1.5.3}$$

where $g(\cdot)$ is the same function in lemma 1.3.1, $\widetilde{A}(x) = f_X(x)\int \widetilde{K}^l(u)du$, and l is a fixed number.

As with lemma 1.3.1, this one also works for Y_i^* and $Y_\star, i = 1, \ldots, n$. For simply describing the following theorems, we denote both of local bootstrap and model-free bootstrap by Y_i^*. Then under assumptions (i)-(viii), we can have the following theorems for both local bootstrap and model free bootstrap:

Theorem 1.5.2.

$$|\mathrm{E}^*[\hat{m}^*_{n,h}(x)] - \hat{m}_{n,h}(x)| = O_p(h^2) \tag{1.5.4}$$

Theorem 1.5.3.

$$nh|\mathrm{Var}^*[\hat{m}^*_{n,h}(x)] - \mathrm{Var}[\hat{m}_{n,h}(x)]| = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big) \tag{1.5.5}$$

Finally, by Lyapunov CLT, we can get:

Theorem 1.5.4. *For any fixed x,*

$$\sqrt{nh}(\hat{m}^*_{n,h}(x) - \mathrm{E}[\hat{m}^*_{n,h}(x)]) \xrightarrow{d} N\left(0, \frac{\kappa\sigma^2(x)}{f(x)}\right) \tag{1.5.6}$$

Recall 1.2.1, and by Theorem 1.5.4, we have:

$$\sup_u \left| P(\sqrt{nh}(\hat{m}^*_{n,h}(x) - \mathrm{E}^*\hat{m}^*_{n,h}(x)) \le u) - P(\sqrt{nh}(\hat{m}_{n,h}(x) - \mathrm{E}\hat{m}_{n,h}(x)) \le u) \right| \xrightarrow{p.} 0$$

It is difficult to estimate $h^2 B_s(x)$, see (1.2.4), thus we can employ under smoothing method, that means under assumptions (ix), we have

$$\sup_u \left| P(\sqrt{nh}(\hat{m}^*_{n,h}(x) - \mathrm{E}^*\hat{m}^*_{n,h}(x)) \le u) - P(\sqrt{nh}(\hat{m}_{n,h}(x) - m_{n,h}(x)) \le u) \right| \xrightarrow{p.} 0$$

Then obtain equal-tailed $(1-\alpha)100\%$ bootstrap confidence interval:

$$[\hat{m}_{n,h}(x) + q^*(\alpha/2), \hat{m}_{n,h}(x) + q^*(1-\alpha/2)]$$

where $q^*(\alpha)$ denote the α-quantile of empirical distribution function of $\hat{m}_{n,h}(x) - \hat{m}^*_{n,h}(x)$.

1.6 Setup of Parameters

1.6.1 The Transform and Transform-back Functions

In this section, we discuss the selection of between step and smooth kernel estimation of conditional CDF of Y. Politis used the smooth kernel estimator in [17, 19], i.e. $\widetilde{D}_x(\cdot)$, see (1.4.3). On the other hand, however, the step

estimation, $\widehat{D}_x(\cdot)$, see (1.4.3), makes the transformed data unnatural and too equidistant in application.

Another important issue is the selection of inverse transform function, i.e. either resampling from transformed data or directly from uniform$[0,1]$, after obtaining u_i^*'s, we need to transform them back to bootstrap data $Y^*|x$. This chapter only shows the validity when we use the discrete inverse function (1.4.4) $\widehat{D}_x^{-1}(\cdot)$. Then the question arises if we use a smooth inverse transformed function, i.e. $\widetilde{D}_x^{-1}(\cdot)$. In fact, as $\widetilde{D}_x(\cdot)$ shares the same good properties as $\widehat{D}_x(\cdot)$, it is supposed to be valid. In addition, the differences between $\widetilde{D}_x^{-1}(\cdot)$ and $\widehat{D}_x(\dot{c})$ are similar to the situation when Efron [8] proposed bootstrap in the beginning, he discussed resampling from regular empirical distribution function and from a kernel distribution estimation. The reason is, if we suppose u_i^* are uniformly from $[0,1]$, then $\widehat{D}_x^{-1}(u_i^*)$ is equivalent to a random variable from $\widehat{D}_x(\cdot)$, so does $\widetilde{D}_x(u_i^*)$. Efron [8] discussed them by comparing the MSE of $\int g(x)d\widehat{F}(x)$ and $\int g(x)d\widetilde{F}(x)$ see [28]. Similarly, define:

$$g(\widehat{E}|X=x) = \int g(y)d\widehat{D}_x(y)$$

$$g(\widetilde{E}|X=x) = \int g(y)d\widetilde{D}_x(y)$$

If we assume $x_1,\ldots,x_n$ are deterministic, and do some simple computation, can obtain:

$$\text{MSE}\left[g(\widetilde{E}|X=x)\right] = \text{MSE}\left[g(\widehat{E}|X=x)\right] + h_0^2 \cdot \frac{\sum_{i=1}^n \left\{\int g(y_i)g''(y_i)dD_{x_i}(y_i)\right\} K\left(\frac{x_i-x}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i-x}{h}\right)} + O(h_0^4)$$

This means, the second term on the left side, let us call it $\triangle$, determine which MSE is greater. For example, if $\triangle > 0$, then finally we think $\widehat{D}_x^{-1}(\cdot)$ is better. However, $\triangle$ is difficult to estimate, but at least we can see, if $g(\cdot)$ is identical function, then

$$\text{MSE}\left[g(\widetilde{E})\right] = \text{MSE}\left[g(\widetilde{E})\right] = O(h_0^4)$$

Consequently their difference is trivial. The same problem was also discussed by Li and Racine [5, 6].

1.6.2 Selection of Bandwidth

Since the smooth kernel estimation of conditional CDF is involved, see lemma 1.4.4, we have to select two different bandwidths, h and h_0. For $\hat{h}$, we are using the same cross validation method in [17], and it has good performances. For h_0, [17] used function **bw.nrd0()** which is a fast bandwidth selection for density estimation and available in R. Nevertheless, according to Li and Racine [5], $h_0 = O(n^{-\frac{2}{5}})$ and $h = O(n^{-\frac{1}{5}})$ by minimizing weighted IMSE. Another cross validation for h_0 does not guarantee to reach the correct order but increase the complexity, thus we suggest the same method in Pan and Politis [22], that simply use $\hat{h}_0 = \hat{h}^2$. In the application, we find that both two selections of h_0 work well, and actually give very close outputs. The work on bandwidth h and especially h_0 requires further study.

1.7 Simulation

1.7.1 Additive Model

The additive model in the simulation is (1.1.1), with $m(x) = \sin(x)$, $\sigma(x) = 1/2$, and errors ϵ_i i.i.d N(0,1) or two sided exponential (Laplace) rescaled to unit variance. Neverthess, $\sigma(x)$ is not used and estimated from the data. For each model, sample size $n = 100$, and totally 500 datasets are created with design points $x_1, \ldots, x_n$ with equal grid on $[0, 2\pi]$. All the kernel functions for N-W estimator, $\widehat{D}_x(\cdot)$ and $\widetilde{D}_x(\cdot)$ are a normal kernel in $\mathbb{R}$.

$$Y_i = \sin(x_i) + \frac{1}{2}\epsilon_i, i = 1, \ldots, n$$

where ϵ_i i.i.d. follows $N(0, 1)$ or laplace distribution with variance 1. In each cell of the following tables, first line gives estimated coverage probability, second line gives mean of length of confidence interval, third line gives standard error of length. "Norm" denote normal approximation by (1.2.3). "MB" and "PRMB" are two different model-based bootstrap methods proposed by Politis [19]. "LB" stands for local bootstrap. For LMF, MF, PMF, we use statistics (1.4.7). For LMF using N-W, we use (1.4.6). All the intervals have confidence level 90%.

• The standard error of the reported coverage probability levels over the 500 replications is $\sqrt{0.1 \times 0.9/500} \approx 0.013$

• Since the true model is $m(x) = \sin(x)$, this simulation has some symmetry that helps us to adjust the CVRs. To elaborate, note that for any $x \in [0, \pi]$, we have $|m(x)| = |m(2\pi - x)|$, and the same symmetry holds for the derivatives of $m(x)$ as well due to the sinusoidal structure. Recall (1.2.1), since $x_1, \ldots, x_n$ are design points, it is easy to see the bias term totally depends on second derivative of $m(x)$. Due to the symmetry, $\hat{m}_{n,h}(x)$ at all the symmetric pairs of point are supposed to have exactly same limit distribution, and hence the same expected CVRs in all methods. All the pairs of point have very nice symmetry, except in the normal error case $m(x)$ visually is always greater at $x = 0.15\pi$ than $x = 1.85\pi$. However, more simulations on 4 more 500 replication show that such trend is random, this issue is more likely caused by datasets themselves. Back to adjusting CVRs by symmetry, it is recommended by Politis [19] to take the average as final estimated CVR. For normal error case table 1 at $x = 0.15\pi$ and $x = 1.85\pi$, the CVR would be better estimated by the average of 0.842 and 0.796, i.e. 0.819.

• Theorem 1.4.2, the equivalence between the local bootstrap and LMF is verified in the simulation. Notice that only LMF using (1.4.6) will produce the very similar result as local bootstrap. Finite sample difference exists when using (1.4.7), and can be reduced with larger sample size.

• Since CDF of laplace distribution violate assumption(vii), and is not continuously differentiable at 0, lemma 1.4.4 might fail or the convergence has a slower rate, thus the $u_1, \ldots, u_n$ calculated from $(X_i, Y_i), i = 1, \ldots, n$ are not uniformly distributed very well, see Fig 1.1. This is one possible reason MF and PMF have serious over-coverage problem for laplace error.

• All the methods appears to have issue that CVR has peak at $x = \pi$ and valley at $x = \pi/2$ and $x = 3\pi/2$. As mentioned above, $\hat{m}_{n,h}(x)$ has different bias at different points, see (1.2.1). But in the simulation we use global bandwidth, and finally lead to different CVR at different points. A further study on local bandwidth selection might be pursued in the future. In addition, at some points, the

over coverage could be explain by ”bias-leakage” in Politis [17] and [19].

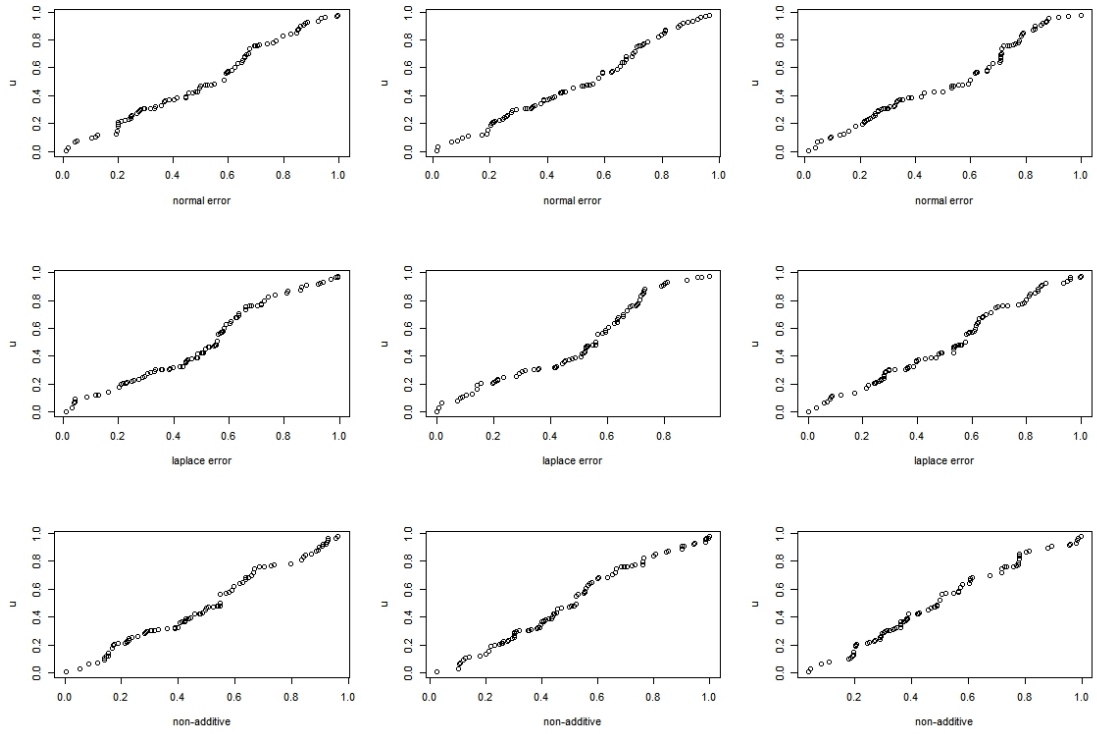


Figure 1.1: QQ plots for all three models

1.7.2 Non-additive Model

In this subsection we recall non-additive model (1.1.3). For easy comparison to section 1.7.1, we use the following model:

$$Y_i = \sin(x_i) + \frac{1}{2}\epsilon_{x_i}, i = 1, \ldots, n$$

where $x_i \in [0, 2\pi]$, $\epsilon_x = \frac{c_x Z + (1-c_x)W}{\sqrt{c_x^2 + (1-c_x)^2}}$, and $c_x = x/2\pi$, $Z \sim N(0,1)$ independent of W that will be distributed as $\frac{1}{2}\chi_2^2 - 1$. Thus ϵ_x has mean 0 and variance 1. $E(Y|X = x) = \sin(x)$ and $\mathrm{Var}(Y|X = x) = 1/4$, $x_1 \ldots, x_n$ are design points equally spaced on $[0, 2\pi$, with sample size $n = 100$. However, ϵ_x has skewness depending on x that violating i.i.d. assumption. The result is presented in table

Table 1.1: Simulation Result for i.i.d. Normal Error

$x_f/pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
Norm	0.842	0.826	0.768	0.824	0.886	0.822	0.796	0.850	0.796
	0.3852	0.3723	0.3711	0.3710	0.3710	0.3711	0.3711	0.3723	0.3852
	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
MB	0.790	0.752	0.740	0.764	0.840	0.778	0.760	0.782	0.752
	0.3865	0.3523	0.3392	0.3598	0.3835	0.3621	0.3377	0.3498	0.3793
	0.004	0.003	0.004	0.003	0.003	0.003	0.004	0.003	0.004
PRMB	0.856	0.832	0.808	0.844	0.886	0.850	0.838	0.854	0.826
	0.4657	0.4255	0.4134	0.4341	0.4621	0.4346	0.4104	0.4224	0.4580
	0.006	0.007	0.008	0.006	0.006	0.006	0.007	0.006	0.006
LB	0.846	0.816	0.800	0.830	0.886	0.840	0.832	0.824	0.812
	0.4403	0.4036	0.3853	0.4129	0.4405	0.4133	0.3866	0.4015	0.4336
	0.004	0.004	0.004	0.004	0.003	0.003	0.004	0.004	0.004
LMF	0.846	0.808	0.804	0.818	0.884	0.824	0.818	0.836	0.816
using	0.4398	0.4023	0.3871	0.4118	0.4382	0.4126	0.3851	0.3985	0.4330
N-W	0.004	0.004	0.004	0.004	0.003	0.003	0.004	0.004	0.004
LMF	0.880	0.844	0.864	0.868	0.880	0.870	0.866	0.852	0.842
	0.5244	0.4907	0.4706	0.5118	0.5498	0.5115	0.4679	0.4851	0.5146
	0.004	0.004	0.004	0.004	0.003	0.003	0.004	0.004	0.004
MF	0.862	0.836	0.828	0.848	0.888	0.852	0.840	0.842	0.826
	0.4579	0.4287	0.4147	0.4408	0.4698	0.4420	0.4123	0.4248	0.4494
	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
PMF	0.928	0.898	0.904	0.926	0.936	0.936	0.916	0.910	0.894
	0.5482	0.5152	0.5005	0.5310	0.5613	0.5319	0.5006	0.5122	0.5412
	0.005	0.005	0.005	0.005	0.005	0.004	0.005	0.005	0.005

Table 1.2: Simulation Result for i.i.d. Laplace Error

$x_f/pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
Norm	0.846	0.792	0.816	0.858	0.884	0.856	0.814	0.846	0.854
	0.3820	0.3701	0.3691	0.3690	0.3691	0.3691	0.3690	0.3700	0.3820
	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
MB	0.788	0.740	0.756	0.808	0.826	0.824	0.760	0.750	0.818
	0.3648	0.3454	0.3307	0.3545	0.3761	0.3535	0.3210	0.3355	0.3627
	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.004	0.004
PRMB	0.870	0.834	0.864	0.894	0.924	0.894	0.846	0.850	0.904
	0.4555	0.4322	0.4119	0.4414	0.4679	0.4394	0.4006	0.4185	0.4513
	0.006	0.006	0.005	0.005	0.005	0.006	0.006	0.006	0.006
LB	0.866	0.790	0.828	0.850	0.908	0.870	0.832	0.818	0.880
	0.4242	0.4008	0.3830	0.4108	0.4373	0.4066	0.3733	0.3904	0.4188
	0.005	0.004	0.004	0.003	0.004	0.004	0.004	0.004	0.004
LMF	0.866	0.798	0.838	0.868	0.914	0.878	0.828	0.826	0.884
using	0.4210	0.3975	0.3819	0.4079	0.4333	0.4072	0.3691	0.3865	0.4180
N-W	0.004	0.004	0.004	0.003	0.004	0.004	0.004	0.004	0.004
LMF	0.900	0.890	0.924	0.932	0.934	0.922	0.908	0.898	0.896
	0.5053	0.4879	0.4648	0.5100	0.5476	0.5072	0.4546	0.4746	0.5022
	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.005
MF	0.868	0.806	0.848	0.890	0.910	0.876	0.836	0.862	0.872
	0.4119	0.3932	0.3765	0.4078	0.4404	0.4064	0.3703	0.3869	0.4124
	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
PMF	0.946	0.914	0.914	0.940	0.964	0.952	0.922	0.928	0.946
	0.5172	0.4962	0.4808	0.5161	0.5499	0.5147	0.4737	0.4909	0.5157
	0.004	0.005	0.004	0.004	0.004	0.004	0.004	0.005	0.004

Table 1.3: Simulation Result for i.i.d. Non-additive Error

$x_f/pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
Norm	0.836	0.834	0.780	0.848	0.876	0.804	0.762	0.838	0.814
	0.3824	0.3700	0.3689	0.3688	0.3689	0.3688	0.3688	0.3700	0.3824
	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
MB	0.746	0.720	0.736	0.754	0.796	0.772	0.718	0.756	0.768
	0.3593	0.3342	0.3146	0.3530	0.3810	0.3511	0.3287	0.3418	0.3697
	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.003	0.003
PRMB	0.808	0.798	0.806	0.850	0.884	0.872	0.822	0.846	0.850
	0.4822	0.4463	0.4205	0.4722	0.5061	0.4393	0.4393	0.4567	0.4908
	0.009	0.009	0.009	0.009	0.010	0.008	0.009	0.009	0.009
LB	0.800	0.782	0.806	0.836	0.872	0.834	0.778	0.810	0.804
	0.4175	0.3885	0.3697	0.4098	0.4422	0.4102	0.3814	0.4006	0.4332
	0.006	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004
LMF	0.798	0.780	0.796	0.824	0.870	0.844	0.760	0.810	0.820
using	0.4185	0.3891	0.3671	0.4114	0.4425	0.4074	0.3832	0.3997	0.4335
N-W	0.006	0.005	0.005	0.005	0.005	0.003	0.004	0.004	0.004
LMF	0.822	0.790	0.782	0.858	0.890	0.876	0.832	0.848	0.828
	0.4977	0.4713	0.4448	0.5090	0.5548	0.5056	0.4639	0.4854	0.5130
	0.007	0.005	0.005	0.006	0.005	0.004	0.005	0.004	0.004
MF	0.794	0.796	0.784	0.824	0.868	0.844	0.772	0.822	0.804
	0.4074	0.3851	0.3690	0.4028	0.4394	0.4108	0.3830	0.3983	0.4210
	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
PMF	0.870	0.888	0.876	0.902	0.938	0.930	0.880	0.894	0.886
	0.5167	0.4897	0.4688	0.5265	0.5509	0.5181	0.4821	0.4971	0.5214
	0.007	0.006	0.005	0.006	0.005	0.005	0.006	0.005	0.005

1.3. We could observe that PMF seems the best method although over coverage take place $x = \pi$ and $x = 1.25\pi$.

1.8 Appendix: Proof

Proof of lemma 1.3.1.

$$
\begin{aligned}
&\mathrm{E}_{X_i,Y_i}\Big[g(Y_i)K^l\Big(\frac{X_i-x}{h}\Big)\Big] \\
&= \int\int g(y_i)K^l(\frac{x_i-x}{h})f_{X_i,Y_i}(x_i,y_i)dx_i dy_i \\
&= h\int\int g(y_i)K^l(u_i)f_{X_i,Y_i}(x+u_ih,y_i)du_i dy_i \\
&= h\int\int g(y_i)K^l(u_i)\Big[f_{X_i,Y_i}(x,y_i)+\frac{\partial f_{X_i,Y_i}(x,y_i)u_ih}{\partial x}+\frac{1}{2}\frac{\partial^2 f_{X_i,Y_i}(x,y_i)u_i^2h^2}{\partial x^2}\Big]du_i dy_i \\
&= h\Bigg\{\int K^l(u_i)du_i\int g(y_i)f_{X_i,Y_i}(x,y_i)dy_i + h\int u_iK^l(u_i)du_i\int g(y_i)\frac{\partial f_{X_i,Y_i}(x,y_i)}{\partial x}dy_i \\
&\quad + \frac{h^2}{2}\int u_i^2K^l(u_i)du_i\int g(y_i)\frac{\partial^2 f_{X_i,Y_i}(x,y_i)}{\partial x^2}dy_i + O(h^3)\Bigg\} \\
&= h\Bigg\{\int K^l(u_i)du_i\int g(y_i)f_{Y_i|X_i}(y_i|x)f_{X_i}(x)dy_i + \frac{h^2}{2}\int u_i^2K^l(u_i)du_i\int g(y_i) \\
&\quad \frac{\partial^2 f_{X_i,Y_i}(x,y_i)}{\partial x^2}dy_i + O(h^3)\Bigg\} \\
&= h\Bigg\{f_X(x)\int K^l(u)du\cdot\mathrm{E}[g(Y)|X=x] + \frac{h^2}{2}\int u^2K^l(u)du\int g(y)\frac{\partial^2}{\partial x^2}f_{X,Y}(x,y)dy + O(h^3)\Bigg\} \\
&= h\Big\{A(x)\cdot\mathrm{E}[g(Y)|X=x] + O(h^2)\Big\}
\end{aligned}
$$

where $A(x) = f_X(x)\int K^l(u)du$. And $O(h^2)$ is not related to i anymore, so they have the common bound. □

Proof of theorem 1.3.2.

By definition of Big O notation in probability, we just need to show:

$$\mathrm{E}^*[g(Y^*)|X=x] - \mathrm{E}[g(Y)|X=x] = O_p\Big(h^2+\frac{1}{\sqrt{nh}}\Big)$$

First, by the definition of local bootstrap algorithm,

$$\mathrm{E}^*[g(Y^*)|X=x] = \frac{\sum_{i=1}^n g(Y_i)K\Big(\frac{X_i-x}{h}\Big)}{\sum_{i=1}^n K\Big(\frac{X_i-x}{h}\Big)}$$

$$= \frac{1}{\widehat{f}_X(x)} \cdot \frac{1}{nh} \sum_{i=1}^{n} g(Y_i) K\Big(\frac{X_i - x}{h}\Big)$$
$$= \frac{1}{f_X(x) + o_p(1)} \cdot \frac{1}{nh} \sum_{i=1}^{n} g(Y_i) K\Big(\frac{X_i - x}{h}\Big)$$

By lemma 1.3.1, we have:

$$\begin{aligned}
\mathrm{E}_{X,Y}\Big[\frac{1}{nh}\sum_{i=1}^{n} g(Y_i)K\Big(\frac{X_i - x}{h}\Big)\Big] &= \frac{1}{nh}\sum_{i=1}^{n} \mathrm{E}_{X_i,Y_i}\Big[g(Y_i)K^l\Big(\frac{X_i - x}{h}\Big)\Big] \\
&= \frac{1}{nh}\sum_{i=1}^{n} h\Big\{f_X(x)\int K(u)du \cdot \mathrm{E}[g(Y)|X = x] + O(h^2)\Big\} \\
&= f_X(x) \cdot \mathrm{E}[g(Y)|X = x] + O(h^2)
\end{aligned}$$

As $\{(X_i, Y_i\}_{i=1}^n$ are independent and under assumption (v), we also have

$$\begin{aligned}
&\mathrm{Var}_{X,Y}\Big[\frac{1}{nh}\sum_{i=1}^{n} g(Y_i)K\Big(\frac{X_i - x}{h}\Big)\Big] \\
&= \frac{1}{(nh)^2}\sum_{i=1}^{n} \mathrm{Var}\Big[g(Y_i)K\Big(\frac{X_i - x}{h}\Big)\Big] \\
&= \frac{1}{(nh)^2}\sum_{i=1}^{n} \left\{\mathrm{E}\Big[g(Y_i)K\Big(\frac{X_i - x}{h}\Big)\Big]^2 - \Big(\mathrm{E}\Big[g(Y_i)K\Big(\frac{X_i - x}{h}\Big)\Big]\Big)^2\right\} \\
&= \frac{1}{(nh)^2}\sum_{i=1}^{n} \left(h\Big\{f_X(x)\int K^2(u)du \cdot \mathrm{E}[g^2(Y)|X = x] + O(h^2)\Big\}\right. \\
&\quad \left. - h^2\Big\{f_X(x)\int K(u)du \cdot \mathrm{E}[g(Y)|X = x] + O(h^2)\Big\}^2\right) \\
&= \frac{1}{nh}\Big[\int K^2(u)du f_X(x)\mathrm{E}[g^2(Y)|X = x] + O(h^2)\Big] - \frac{1}{n}\Big[f_X(x)\mathrm{E}[g(Y)|X = x] + O(h^2)\Big]^2 \\
&= O\Big(\frac{1}{nh}\Big)
\end{aligned}$$

By Markov's Inequality, it is easy to get

$$\frac{1}{nh}\sum_{i=1}^{n} g(Y_i)K\Big(\frac{X_i - x}{h}\Big) - f_X(x)\mathrm{E}[g(Y_i)] = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)$$

Finally, we get:

$$\begin{aligned}
&\mathrm{E}^*[g(Y)|x] - \mathrm{E}[g(Y)|x] \\
&= \frac{1}{f_X(x) + o_p(1)} \cdot \frac{1}{nh}\sum_{i=1}^{n} g(Y_i)K\Big(\frac{X_i - x}{h}\Big) - \mathrm{E}[g(Y)|x] \\
&= \frac{1}{f_X(x) + o_p(1)} \cdot \left[\frac{1}{nh}\sum_{i=1}^{n} g(Y_i)K\Big(\frac{X_i - x}{h}\Big) - [f_X(x) + o_p(1)]\mathrm{E}[g(Y)|x]\right]
\end{aligned}$$

$$= \frac{O_p(h^2 + 1/\sqrt{nh}) - o_p(1)\mathrm{E}[g(Y)|x]}{f_X(x) + o_p(1)}$$
$$= O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)$$

□

Proof of theorem 1.4.2.

Recall 1.3.1. We already know that, in local bootstrap, we use $\widehat{G}_{x_i}$ to generate Y_i^*. To show the equivalence, we will prove there exists another kind of distribution function $\widehat{H}_{x_i}$ to generate $Y_i^\star$, and $\widehat{G}_{x_i} = \widehat{H}_{x_i}$.

$$\widehat{H}_{x_i} : \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_n \\ \widetilde{w}_{1i} & \widetilde{w}_{2i} & \cdots & \widetilde{w}_{ni} \end{pmatrix}$$

The reason we can define such a function for model free, is that no matter what u_i^* we get from Uniform(0,1), by the transformed function $\widehat{D}_x^{-1}(\cdot), Y^\star$ will be always one point from $\{Y_1, \cdots, Y_n\}$. Thus for each Y_j ,there is a probability we choose it as $Y_i^\star$. For each i, j, we define $Y_j^- = \max_{k=1,\ldots,n}\{Y_k, Y_k < Y_j\}$. Then, we have:

$$\begin{aligned}
\widetilde{w}_{ij} &= P(Y_i^\star = Y_j) \\
&= P(\widehat{D}_{x_i}^{-1}(u_i^*) = Y_j) \\
&= P(\mathbb{1}_{\{\exists k, Y_k < Y_j\}} \cdot \widehat{D}_{x_i}(Y_j^-) < u_i^* \leq \widehat{D}_{x_i}(Y_j)) \\
&= \frac{\sum_{t=1}^n \mathbb{1}_{\{Y_t \leq Y_j\}} K\Big(\frac{X_i - X_t}{h}\Big)}{\sum_{t=1}^n K\Big(\frac{X_i - X_t}{h}\Big)} - \mathbb{1}_{\{\exists k, Y_k < Y_j\}} \cdot \frac{\sum_{t=1}^n \mathbb{1}_{\{Y_t \leq Y_j^-\}} K\Big(\frac{X_i - X_t}{h}\Big)}{\sum_{t=1}^n K\Big(\frac{X_i - X_t}{h}\Big)} \\
&= \frac{K\Big(\frac{X_i - X_j}{h}\Big)}{\sum_{t=1}^n K\Big(\frac{X_i - X_t}{h}\Big)}
\end{aligned}$$

□

Proof of lemma 1.4.3.

Recall that we suppose each $u_i, i = 1 \cdots, n$ is uniformly distributed. then, by the algorithm of Model free method 2,

$$\mathrm{E}^\star[g(Y^\star)|X = x] = \frac{1}{n}\sum_{i=1}^{n} g[\widehat{D}_x^{-1}(u_i)]$$

Then, consider

$$|\mathrm{E}^\star[g(Y^\star)|x] - \mathrm{E}[g(Y)|x]| \leq |\mathrm{E}^\star[g(Y^\star)|x] - \mathrm{E}^*[g(Y^*)|x]| + |\mathrm{E}^*[g(Y^*)|x] - \mathrm{E}[g(Y)|x]| \tag{1.8.1}$$

Let E^* denote expectation in local bootstrap world, and $\mathrm{E}^\star$ denote expectation in model free bootstrap world. For the right side of inequality above, the second term is proved in theorem 1.3.2. For the first term, by theorem 1.4.2, we know $\mathrm{E}^*[g(Y^*)|X = x] = \mathrm{E}_u g[\widehat{D}_x^{-1}(u)]$, where u comes from uniform[0,1] distribution. Let random vector $U = (u_1, \ldots, u_n)$, independent from u. Then we have:

$$\begin{aligned}
\mathrm{E}\{\mathrm{E}^\star[g(Y^\star)|x] - \mathrm{E}^*[g(Y^*)|x]\} &= \mathrm{E}_{X,Y,U}\Big[\frac{1}{n}\sum_{i=1}^{n} g[\widehat{D}_x^{-1}(u_i)] - \mathrm{E}_u g[\widehat{D}_x^{-1}(u)]\Big] \\
&= \mathrm{E}_{X,Y}\Big\{\mathrm{E}_U\Big[\frac{1}{n}\sum_{i=1}^{n} g[\widehat{D}_x^{-1}(u_i)\Big|X,Y]\Big] - \mathrm{E}_u g\Big[\widehat{D}_x^{-1}(u)\Big|X,Y\Big]\Big\} \\
&= 0
\end{aligned} \tag{1.8.2}$$

Then consider the variance:

$$\begin{aligned}
&\mathrm{Var}_{X,Y,U}\big\{\mathrm{E}^\star[g(Y^\star)|X = x] - E^*[g(Y^*)|X = x]\big\} \\
&= \mathrm{E}_{X,Y,U}\Big\{\mathrm{E}^\star[g(Y^\star)|X = x] - E^*[g(Y^*)|X = x]\Big\}^2 \\
&= \mathrm{E}_{X,Y}\Big(\mathrm{E}_U\big\{\mathrm{E}^\star[g(Y^\star)|X = x] - E^*[g(Y^*)|X = x]\big|X,Y\big\}^2\Big) \\
&= \mathrm{E}_{X,Y}\Big(\mathrm{E}_U\Big\{\frac{1}{n}\sum_{i=1}^{n} g[\widehat{D}_x^{-1}(u_i)] - \mathrm{E}_u g[\widehat{D}_x^{-1}(u)]\Big|X,Y\Big\}^2\Big) \\
&= \mathrm{E}_{X,Y}\Big(\mathrm{Var}_U\Big\{\frac{1}{n}\sum_{i=1}^{n} g[\widehat{D}_x^{-1}(u_i)]\Big|X,Y\Big\}\Big) \\
&= \mathrm{E}_{X,Y}\Big(\frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}_U\big\{g[\widehat{D}_x^{-1}(u_i)]\big|X,Y\big\}\Big) \\
&= \frac{1}{n}\mathrm{E}_{X,Y}\Big(\mathrm{Var}^*[g(Y^*)|X = x]\Big) \\
&= \frac{1}{n}\mathrm{E}_{X,Y}\Big(\mathrm{E}^*[g(Y^*)|X = x]^2 - \big\{\mathrm{E}^*[g(Y^*)|X = x]\big\}^2\Big)
\end{aligned}$$

For the second term, in theorem 1.3.2, we already show that

$$\mathrm{E}^*[g(Y^*)|X=x] - \mathrm{E}[g(Y)|X=x] = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)$$

and

$$\mathrm{Var}_{X,Y}\Big(\mathrm{E}^*[g(Y^*)|X=x]\Big) = O\Big(\frac{1}{nh}\Big)$$

which implies

$$\mathrm{E}_{X,Y}\Big(\mathrm{E}^*[g(Y^*)|X=x]\Big)^2 = \mathrm{Var}_{X,Y}\Big(\mathrm{E}^*[g(Y^*)|X=x]\Big) + \Big[\mathrm{E}_{X,Y}\Big(\mathrm{E}^*[g(Y^*)|X=x]\Big)\Big]^2$$

is bounded. Then let us see the first term:

$$\begin{aligned}
&\mathrm{E}_{X,Y}\Big\{\mathrm{E}^*[g(Y^*)|X=x]^2\Big\} \\
&= \mathrm{E}_{X,Y}\Big\{\frac{1}{nh}\sum_{i=1}^{n} g^2(Y_i)\widetilde{K}\Big(\frac{X_i - x}{h}\Big)\Big\} \\
&= \frac{1}{nh}\sum_{i=1}^{n} \mathrm{E}_{X,Y}\Big\{g^2(Y_i)\widetilde{K}\Big(\frac{X_i - x}{h}\Big)\Big\} \\
&= \frac{1}{nh}\sum_{i=1}^{n} h\Big\{f_X(x)\int \widetilde{K}(u)du \cdot \mathrm{E}[g^2(Y)|X=x] + O(h^2)\Big\} \quad \text{by lemma 1.5.1} \\
&= \widetilde{A}(x)\mathrm{E}[g^2(Y)|X=x] + O(h^2) \\
&< \infty
\end{aligned}$$

Thus $\mathrm{Var}_{X,Y,U}\{\mathrm{E}^\star[g(Y^\star)|X=x] - E^*[g(Y^*)|X=x]\} = O(1/n)$. This and 1.8.2, Chebyshev inequality, and the definition of big O in probability notation leads to:

$$\mathrm{E}^\star[g(Y^\star)|X=x] - E^*[g(Y^*)|X=x] = O_p\Big(\frac{1}{\sqrt{n}}\Big)$$

Recall 1.8.1, we have:

$$\begin{aligned}
|\mathrm{E}^\star[g(Y^\star)|X=x] - \mathrm{E}[g(Y)|X=x]| &= O_p\Big(\frac{1}{\sqrt{n}}\Big) + O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big) \\
&= O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)
\end{aligned}$$

□

Proof of lemma 1.4.4.

Let $\widehat{d}(x) = \frac{1}{nh}\sum_{i=1}^{n}\Lambda\Big(\frac{Y_i-y}{h_0}\Big)K\Big(\frac{X_i-x}{h}\Big)$, then we write:

$$\begin{aligned}\sup_{x\in\mathcal{S}}|\widetilde{D}_x(y)-D_x(y)| &= \sup_{x\in\mathcal{S}}\left|\frac{\widehat{d}(x)}{\widehat{f}(x)}-D_x(y)\right| \\ &\leq \frac{\sup_{x\in\mathcal{S}}\left|\widehat{d}(x)-\mathrm{E}\big(\widehat{d}(x)\big)+\mathrm{E}\big(\widehat{d}(x)\big)-D_x(y)\widehat{f}(x)\right|}{\inf_{x\in\mathcal{S}}|\hat{f}(x)|} \\ &\leq \frac{\sup_{x\in\mathcal{S}}\left|\widehat{d}(x)-\mathrm{E}\big(\widehat{d}(x)\big)\right|}{\inf_{x\in\mathcal{S}}|\hat{f}(x)|}+\frac{\sup_{x\in\mathcal{S}}\left|\mathrm{E}\big(\widehat{d}(x)\big)-D_x(y)\widehat{f}(x)\right|}{\inf_{x\in\mathcal{S}}|\hat{f}(x)|}\end{aligned}$$

Where $\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}K\Big(\frac{X_i-x}{h}\Big)$. By assumption (ii) and thm 1.4 from [5], i.e $\sup_{x\in\mathcal{S}}|\hat{f}(x)-f(x)| \xrightarrow{a.s.} 0$, for large sufficiently enough n we can find another δ' s.t. $\inf_{x\in\mathcal{S}}\hat{f}(x) \geq \delta' > 0$. And for the second term, by theorem 6.2 in Li and Racine [5], we have:

$$\begin{aligned}&\Big\{\mathrm{E}\big[\widehat{d}(x)\big]-D_x(y)\widehat{f}(x)\Big\} \\ &= \frac{h_0^2}{2}\int u^2K(u)du\cdot\frac{\partial^2 D_x(y)}{\partial y^2}f(x)+\frac{h^2}{2}\int u^2K(u)du\Big[\frac{\partial^2 D_x(y)}{\partial x^2}f(x) \\ &+2\frac{\partial f(x)}{\partial x}\cdot\frac{\partial D_x(y)}{\partial x}\Big]+o(h_0^2+h^2) \\ &= O(h_0^2+h^2)\end{aligned}$$

By assumption (vii), $\frac{\partial^2 D_x(y)}{\partial y^2}$ and $\sup_{x\in\mathcal{S}}\Big[\frac{\partial D_x(y)}{\partial x}f(x)+2\frac{\partial f(x)}{\partial x}\cdot\frac{\partial D_x(y)}{\partial x}\Big]$ is bounded, which implies

$$\sup_{x\in\mathcal{S}}\Big|\mathrm{E}\big(\widehat{d}(x)\big)-D_x(y)\widehat{f}(x)\Big| = O(h_0^2+h^2)$$

Then just need to show

$$\sup_{x\in\mathcal{S}}\Big|\widehat{d}(x)-\mathrm{E}\big(\widehat{d}(x)\big)\Big| = O\Big(\Big[\frac{\ln(n)}{nh}\Big]^{\frac{1}{2}}\Big) \quad \text{a.s.} \tag{1.8.3}$$

As $\mathcal{S}$ is compact, it can be covered by a finite number L_n of interval $\{I_k\}_1^n$ with length l_n, and $L_n = \text{constant}/l_n$. We write:

$$\sup_{x\in\mathcal{S}}\Big|\widehat{d}(x)-\mathrm{E}\big(\widehat{d}(x)\big)\Big| = \max_{1\leq k\leq L_n}\sup_{x\in\mathcal{S}\bigcap I_k}|\widehat{d}(x)-\mathrm{E}\big(\widehat{d}(x)\big)|$$

$$
\begin{aligned}
&\leq \max_{1\leq k\leq L_n} \sup_{x\in\mathcal{S}\bigcap I_k} |\widehat{d}(x) - \widehat{d}(x_{k,n})| \\
&\quad + \max_{1\leq k\leq L_n} |\widehat{d}(x_{k,n}) - \mathrm{E}[\widehat{d}(x_{k,n})]| \\
&\quad + \max_{1\leq k\leq L_n} \sup_{x\in\mathcal{S}\bigcap I_k} |\mathrm{E}[\widehat{d}(x_{k,n})] - \mathrm{E}[\widehat{d}(x)]| \\
&= Q_1 + Q_2 + Q_3
\end{aligned}
$$

For Q_2, the sup can be ignored because it only concern $x_{k,n}$ but not x, and $x_{k,n}$ is the central point of the interval I_k. We will show Q_1 and Q_3 in the last part, just consider Q_2 first. To show $Q_2 = O(\eta_n)$ a.s., by Borel-Cantelli Lemma, we just need to show

$$\sum_{n=1}^{\infty} P[Q_2 > \eta_n] < \infty$$

Let $W_n(x) = \widehat{d}(x) - \mathrm{E}[\widehat{d}(x)] = \sum_i Z_{n,i}$. where

$$Z_{n,i} = \frac{1}{nh}\Big\{\Lambda\Big(\frac{Y_i - y}{h_0}\Big)K\Big(\frac{X_i - x}{h}\Big) - \mathrm{E}\Big[\Lambda\Big(\frac{Y_i - y}{h_0}\Big)K\Big(\frac{X_i - x}{h}\Big)\Big]\Big\}$$

For any $\eta_n > 0$, we have

$$
\begin{aligned}
P[Q_2 > \eta_n] &= P[\max_{1\leq k\leq L_n} |W_n(x_{k,n})| > \eta_n] \\
&\leq \sum_{k=1}^{L_n} P[|W_n(x_{k,n})| > \eta_n] \\
&\leq L_n \sup_{x\in\mathcal{S}} P[|W_n(x)| > \eta_n]
\end{aligned}
$$

Since $K(\cdot)$ is bounded, let its supremum is A_1, moreover, $\Lambda(\cdot)$ is a CDF and so $|\Lambda(\cdot) \leq 1|$, thus we have $Z_{n,i} \leq \frac{2A_1}{nh}$. Define $\lambda_n = (nh\ln(n))^{\frac{1}{2}}$, then $\lambda_n|Z_{n,i}| \leq 2A_1\left[\frac{\ln(n)}{nh}\right]^{\frac{1}{2}}$. Then we can choose sufficiently large n such that $\lambda_n|Z_{n,i}| \leq \frac{1}{2}$ for all $i = 1, 2, \cdots, n$. Now use inequality

$$\exp(x) \leq 1 + x + x^2 \quad \text{as} \quad |x| \leq \frac{1}{2}$$

and

$$1 + x \leq \exp(x) \quad \text{as} \quad x \geq 0$$

We have:

$$E[\exp(\pm\lambda_n Z_{n,i})] \leq 1 + E[\pm\lambda_n Z_{n,i}] + E[\lambda_n^2|Z_{n,i}^2|] \leq \exp(E[\lambda_n^2|Z_{n,i}^2|]) \quad \text{a.s.}$$

By the Markov inequality, $(a > 0)$

$$P[X > c] \leq \frac{E[\exp(Xa)]}{\exp(ac)} \tag{1.8.4}$$

Thus,

$$\begin{aligned}
P[|W_n(x)| > \eta] &= P\Big[\Big|\sum_{i=1}^{n} Z_{n,i}\Big| > \eta\Big] \\
&= P\Big[\sum_{i=1}^{n} Z_{n,i} > \eta\Big] + P\Big[-\sum_{i=1}^{n} Z_{n,i} > \eta\Big] \\
&\leq \frac{E[\exp(\lambda_n \sum_{i=1}^{n} Z_{n,i})] + E[\exp(-\lambda_n \sum_{i=1}^{n} Z_{n,i})]}{\exp(\lambda_n \eta)} \\
&\leq 2\exp(-\lambda_n \eta) \prod_{i=1}^{n} \Big[\exp\Big(\lambda_n^2 E Z_{n,i}^2\Big)\Big] \\
&\leq 2\exp(-\lambda_n \eta)\Big[\exp\Big(\frac{A_2 \lambda_n^2}{nh}\Big)\Big]
\end{aligned}$$

Where we use

$$\begin{aligned}
EZ_{n,i}^2 &= \frac{1}{(nh)^2}\mathrm{Var}\Big[\Lambda\Big(\frac{Y_i - y}{h_0}\Big)K\Big(\frac{X_i - x}{h}\Big)\Big] \\
&\leq \frac{1}{n^2h^2}E\Big[\Lambda\Big(\frac{Y_i - y}{h_0}\Big)K\Big(\frac{X_i - x}{h}\Big)\Big]^2 \\
&\leq \frac{1}{n^2h^2}\int K^2\Big(\frac{x_i - x}{h}\Big) f_X(x_i)dx_i \\
&= \frac{1}{n^2h}\int K^2(u)) f_X(x + uh)du \\
&= \frac{1}{n^2h}\int K^2(u))\Big[f_X(x) + \frac{\partial f_X(x) uh}{\partial x} + O(h^2)\Big]du \\
&= \frac{1}{n^2h}\Big\{f_X(x)\int K^2(u)du + h\frac{\partial f_X(x)}{\partial x}\int uK^2(u)du + O(h^2)\Big\} \\
&= \frac{f(x)\int K^2(u)du}{n^2h}[1 + O(h^2)] \\
&\leq \frac{A_2}{n^2h}
\end{aligned}$$

As $f(x)$ is bounded, and $\int uK^2(u)du = 0$ since K is symmetric. Because A_2 is independent of x, we get

$$\sup_{x\in\mathcal{S}} P[|W_n(x)| > \eta_n] \leq 2\exp\Big(-\lambda_n\eta_n + \frac{A_2\lambda_n^2}{nh}\Big)$$

And let $\lambda_n = [(nh)\ln(n)]^{\frac{1}{2}}$ and $\eta_n = C_4 \ln(n)/\lambda_n = C_4[\ln(n)/(nh)]^{\frac{1}{2}}$, where C_4 is positive. Choosing large enough C_4, we get

$$\Big(-\lambda_n \eta + \frac{A_2 \lambda_n^2}{nh} \Big) = (-C_4 + A_2) \ln n = \alpha \ln n$$

where $\alpha = (-C_4 + A_2)$ is negative, and it can be small enough (by the choice of C_4) so that:

$$\sum_{n=1}^{\infty} P[Q_2 > \eta_n] \leq \sum_{n=1}^{\infty} \sup_{x \in \mathcal{S}} P[|W_n(x)| > \eta_n] \leq \sum_{n=1}^{\infty} \frac{L_n}{n^{\alpha}} < \infty$$

by choosing $L_n = \sqrt{\frac{n}{h^3 \ln n}}$. Thus, based on all the above and Borel-Cantelli Lemma, we know that,

$$Q_2 = O\Big(\Big[\frac{\ln(n)}{nh} \Big]^{\frac{1}{2}} \Big) \quad \text{a.s.}$$

Now consider Q_1 and Q_3. By the Lipschitz condition on $K(\cdot)$, we know that

$$\begin{aligned} \sup_{\mathcal{S} \bigcap I_k} \Big| \Lambda\Big(\frac{Y_i - y}{h_0} \Big) K\Big(\frac{X_i - x}{h} \Big) - \Lambda\Big(\frac{Y_i - y}{h_0} \Big) K\Big(\frac{X_i - x_{k,n}}{h} \Big) \Big| &\leq \frac{C_1}{h} \sup_{\mathcal{S} \bigcap I_k} ||x - x_{k,n}|| \\ &\leq \frac{C_1 l_n}{h} \end{aligned}$$

By using the same choice of L_n above, we have $l_n = \text{constant} \cdot \sqrt{\frac{h^3 \ln n}{n}}$, and

$$|Q_1| \leq \frac{C_1 l_n}{h^2} = O\Big(\Big[\frac{\ln(n)}{nh} \Big]^{\frac{1}{2}} \Big) \quad \text{a.s.}$$

And by exactly same arugement, we can also show

$$|Q_3| = O\Big(\Big[\frac{\ln(n)}{nh} \Big]^{\frac{1}{2}} \Big) \quad \text{a.s.}$$

Thus, finally, we get

$$\sup_{x \in \mathcal{S}} |\widetilde{D}_x(y) - D_x(y)| = O\Big(h_0^2 + h^2 + \Big[\frac{\ln(n)}{nh} \Big]^{\frac{1}{2}} \Big) \quad \text{a.s.}$$

□

Proof of Lemma 1.4.7.

By the property of continuity of $g_n(\cdot)$, we know that if $|f_n(Z_n) - U| \to 0$ as $n \to \infty$, then $|g_n[Z_n, f_n(Z_n)] - g_n(Z_n, U)| \to 0$ as $n \to \infty$. It implies

$$1 \geq \mathrm{P}\left(\lim_{n\to\infty}[g_n(Z_n, f_n(Z_n)) - g_n(Z_n, U)] = 0\right) \geq \mathrm{P}\left(\lim_{n\to\infty}[f_n(Z_n) - U] = 0\right) = 1$$

Consequently as $n \to \infty$

$$|g_n(Z_n, f_n(Z_n)) - g_n(Z_n, U)| \to 0 \text{ a.s.}$$

□

Proof of Theorem 1.4.8.

First, by Probability Integral Transform theorem, we have

$$\mathrm{E}[g(Y)|X = x] = \mathrm{E}\{g[D_x^{-1}(U)]\}$$

The first expectation is with respect to $Y|x$, the expectation on the right side is with respect to r.v. U, that follows uniform [0,1]. By Law of Large Number, we have

$$\mathrm{E}\{g[D_x^{-1}(U)]\} = \frac{1}{n}\sum_{i=1}^{n} g[D_x^{-1}(U_i)] + o_p\left(\frac{1}{\sqrt{n}}\right)$$

where U_i follows Unif[0,1] i.i.d. On the other hand, by the algorithm of model free and Lemma 1.4.7, we know that

$$\begin{aligned}\mathrm{E}^\star[g(Y^\star)|X = x] &= \frac{1}{n}\sum_{i=1}^{n} g\left(\widetilde{D}_x^{-1}[\widetilde{D}_{x_i}(Y_i)]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} g\left(\widetilde{D}_x^{-1}[D_{x_i}(Y_i)]\right) \text{ a.s.}\end{aligned}$$

Then, to show the theorem, we just need to prove

$$\left|\frac{1}{n}\sum_{i=1}^{n} g\left(\widetilde{D}_x^{-1}[D_{x_i}(Y_i)]\right) - \frac{1}{n}\sum_{i=1}^{n} g[D_x^{-1}(U_i)]\right| = O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right)$$

Notice that

$$\left|\frac{1}{n}\sum_{i=1}^{n} g\left(\widetilde{D}_x^{-1}[D_{x_i}(Y_i)]\right) - \frac{1}{n}\sum_{i=1}^{n} g[D_x^{-1}(U_i)]\right|$$

$$\leq \left| \frac{1}{n} \sum_{i=1}^{n} g\left(\widetilde{D}_x^{-1}[D_{x_i}(Y_i)] \right) - \frac{1}{n} \sum_{i=1}^{n} g\left(D_x^{-1}[D_{x_i}(Y_i)] \right) \right|$$
$$+ \left| \frac{1}{n} \sum_{i=1}^{n} g\left(D_x^{-1}[D_{x_i}(Y_i)] \right) - \frac{1}{n} \sum_{i=1}^{n} g[D_x^{-1}(U_i)] \right| \tag{1.8.5}$$

$D_{x_i}(Y_i)$ essentially follows uniform[0,1] by Probability Integral Transform Theorem. Thus the second term is $o_p\left(\frac{1}{\sqrt{n}}\right)$ by Law of Large Number. The first term is however, complicated since $\widetilde{D}_x^{-1}(\cdot)$ includes X_i, Y_i. But it is straightforward if we can show

$$\sup_{x_i, y_i} \left| g\left(\widetilde{D}_x^{-1,i}[D_{x_i}(y_i)] \right) - g\left(D_x^{-1}[D_{x_i}(y_i)] \right) \right| = O_p\left(h^2 + \frac{1}{\sqrt{nh}} \right)$$

where $\widetilde{D}_x^{-1,i}(\cdot)$ denote the same function $\widetilde{D}_x^{-1}(\cdot)$ but with variables x_i, y_i instead of random variables X_i, Y_i. However, the equality above might fail because $D_{x_i}(y_i)$ can reach 0 or 1, which leads $\widetilde{D}_x^{-1}(\cdot)$ and $D_x^{-1}(\cdot)$ to infinity. Fortunately, we just study these functions in probability, which implies $Y|X = x$, and $D_x(Y)$ are both bounded in probability. $\forall \epsilon > 0, \exists M_\epsilon$, s.t.

$$\mathrm{P}(|Y|X = x| > M_\epsilon) < \epsilon$$

For the same ϵ, there also exists $\widetilde{M}_\epsilon > 0$, such that

$$\mathrm{P}\left(D_x(Y) < 1 - \widetilde{M}_\epsilon \text{ and } D_x(Y) > \widetilde{M}_\epsilon \right) < \epsilon$$

Here notice that $M_\epsilon, \widetilde{M}_\epsilon$ does not depend on i. Now we consider to prove

$$\sup_{x_i; y_i \in [1-\widetilde{M}_\epsilon, \widetilde{M}_\epsilon]} \left| g\left(\widetilde{D}_x^{-1,i}[D_{x_i}(y_i)] \right) - g\left(D_x^{-1}[D_{x_i}(y_i)] \right) \right| = O_p\left(h^2 + \frac{1}{\sqrt{nh}} \right)$$

Since the function inside is continuous, we can attain the maximum. Denote the maximizer by $x_0^{(n)}, y_0^{(n)}$, and let $u_0^{(n)} = D_{x_0^{(n)}}(y_0^{(n)})$, then

$$\sup_{x_i; y_i \in [1-\widetilde{M}_\epsilon, \widetilde{M}_\epsilon]} \left| g\left(\widetilde{D}_x^{-1,i}[D_{x_i}(y_i)] \right) - g\left(D_x^{-1}[D_{x_i}(y_i)] \right) \right|$$
$$= \left| g\left(\widetilde{D}_x^{-1,i,0}(u_0^{(n)}) \right) - g\left(D_x^{-1}(u_0^{(n)}) \right) \right| \tag{1.8.6}$$

Let $s_n = D_x^{-1}(u_0^{(n)}), \tilde{s}_n = \widetilde{D}_x^{-1,i,0}(u_0^{(n)}), \tilde{t}_n = \widetilde{D}_x^{i,0}(s_n)$. where $\widetilde{D}_x^{-1,i,0}(\cdot)$ and $\widetilde{D}_x^{i,0}(\cdot)$ denote the same function $\widetilde{D}_x^{-1}(\cdot)$ and $\widetilde{D}_x(\cdot)$ but with maximizer $x_0^{(n)}, y_0^{(n)}$ instead of random variables X_i, Y_i. By the definition of $s_n, \tilde{s}_n, \tilde{t}_n$, we have

$$u_0^{(n)} = D_x(s_n) = \widetilde{D}_x^{i,0}(\tilde{s}_n)$$

$$s_n = D_x^{-1}(u_0^{(n)}) = \widetilde{D}_x^{-1,i,0}(\tilde{t}_n)$$

Then we go back the right side of equation (1.8.6). In the following procedures, we are using the tricks $D_x^{-1}(u_0^{(n)}) = \widetilde{D}_x^{-1,i,0}(\tilde{t}_n), u_0^{(n)} = D_x(s_n), \tilde{t}_n = \widetilde{D}_x^{i,0}(s_n)$,

$$\left| g\left(\widetilde{D}_x^{-1,i,0}(u_0^{(n)})\right) - g\left(D_x^{-1}(u_0^{(n)})\right)\right| = \left| g\left(\widetilde{D}_x^{-1,i,0}(u_0^{(n)})\right) - g\left(\widetilde{D}_x^{-1,i,0}(\tilde{t}_n)\right)\right|$$

Since $g(x) = x^k$ or $g(x) = |x|^k, k = 1, 2, 3$, let $h(x) = x^k$ only, then

$$\begin{aligned}
\left| g\left(\widetilde{D}_x^{-1,i,0}(u_0^{(n)})\right) - g\left(D_x^{-1}(u_0^{(n)})\right)\right| &= \left| g\left(\widetilde{D}_x^{-1,i,0}(u_0^{(n)})\right) - g\left(\widetilde{D}_x^{-1,i,0}(\tilde{t}_n)\right)\right| \\
&\leq \left| h\left(\widetilde{D}_x^{-1,i,0}(u_0^{(n)})\right) - h\left(\widetilde{D}_x^{-1,i,0}(\tilde{t}_n)\right)\right| \\
&= \frac{h'\left(\widetilde{D}_x^{-1,i,0}(c_n)\right)}{\widetilde{D}_x'^{i,0}[\widetilde{D}_x^{-1,i,0}(c_n)]}\left| u_0^{(n)} - \tilde{t}_n\right| \\
&= \frac{h'\left(\widetilde{D}_x^{-1,i,0}(c_n)\right)}{\widetilde{D}_x'^{i,0}[\widetilde{D}_x^{-1,i,0}(c_n)]}\left| \widetilde{D}_x^{i,0}(s_n) - D_x(s_n)\right|
\end{aligned}$$

The third line is derived by mean value theorem. First we consider the equation $\left|\widetilde{D}_x^{i,0}(s_n) - D_x(s_n)\right|$

$$\left|\widetilde{D}_x^{i,0}(s_n) - D_x(s_n)\right| \leq \left|\widetilde{D}_x^{i,0}(s_n) - \widetilde{D}_x(s_n)\right| + \left|\widetilde{D}_x(s_n) - D_x(s_n)\right|$$

On the right side, the second term is shown by Q. Li and J.S. Racine [5] in theorem 6.2, that

$$\left|\widetilde{D}_x(s_n) - D_x(s_n)\right| = O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right)$$

Thus, we only need to concern the first term $\left|\widetilde{D}_x^{i,0}(s_n) - \widetilde{D}_x(s_n)\right|$. Actually their difference is very minor

$$\widetilde{D}_x^{i,0}(x) = \frac{\sum_{j\neq i}^n \Lambda\left(\frac{Y_j - y}{h_0}\right) K\left(\frac{X_j - x}{h}\right)}{\sum_{j\neq i}^n K\left(\frac{X_j - x}{h}\right)}$$

$$= \frac{\sum_{j=1}^{n} \Lambda\Big(\frac{Y_j-y}{h_0}\Big)K\Big(\frac{X_j-x}{h}\Big) + \Lambda\Big(\frac{y_i-y}{h_0}\Big)K\Big(\frac{y_i-x}{h}\Big) - \Lambda\Big(\frac{Y_i-y}{h_0}\Big)K\Big(\frac{X_i-x}{h}\Big)}{\sum_{j=1}^{n} K\Big(\frac{X_j-x}{h}\Big) + K\Big(\frac{x_i-x}{h}\Big) - K\Big(\frac{X_i-x}{h}\Big)}$$

Let

$$a_n = \sum_{j=1}^{n} \Lambda\Big(\frac{Y_j-y}{h_0}\Big)K\Big(\frac{X_j-x}{h}\Big), \delta_{a,i} = \Lambda\Big(\frac{y_i-y}{h_0}\Big)K\Big(\frac{y_i-x}{h}\Big), \triangle_{a,i} = \Lambda\Big(\frac{Y_i-y}{h_0}\Big)K\Big(\frac{X_i-x}{h}\Big)$$

$$b_n = \sum_{j=1}^{n} K\Big(\frac{X_j-x}{h}\Big), \delta_{b,i} = K\Big(\frac{x_i-x}{h}\Big), \triangle_{b,i} = K\Big(\frac{X_i-x}{h}\Big)$$

Then

$$\begin{aligned}
\widetilde{D}_x^{i,0}(y) &= \frac{a_n + \delta_{a,i} - \triangle_{a,i}}{b_n + \delta_{b,i} - \triangle_{b,i}} \\
&= \frac{a_n}{b_n}\left(\frac{1 - \frac{\triangle_{a,i}-\delta_{a,i}}{a_n}}{1 - \frac{\triangle_{b,i}-\delta_{b,i}}{b_n}}\right) \\
&= \frac{a_n}{b_n}\left(1 - \frac{\triangle_{a,i} - \delta_{a,i}}{a_n}\right)\left(1 + \frac{\triangle_{b,i} - \delta_{b,i}}{b_n} + \left[\frac{\triangle_{b,i} - \delta_{b,i}}{b_n}\right]^2 + o\left[\frac{\triangle_{b,i} - \delta_{b,i}}{b_n}\right]^2\right) \\
&= \frac{a_n}{b_n} + O_p\left(\frac{1}{n}\right) \\
&= \widetilde{D}_x(y) + O_p\left(\frac{1}{n}\right)
\end{aligned}$$

Above all, we have that

$$\left|\widetilde{D}_x^{i,0}(s_n) - D_x(s_n)\right| = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)$$

It also imples

$$\left|u_0^{(n)} - \tilde{t}_n\right| = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)$$

We know that $u_0^{(n)}$ is bounded in probability. With a large enough n, $\tilde{t}_n$ is also bounded in probability. And c_n is a value between $u_0^{(n)}$ and $\tilde{t}_n$, and also bounded in probability. This leads to $\frac{h'\left(\widetilde{D}_x^{-1,i,0}(c_n)\right)}{\widetilde{D}_x'^{i,0}[\widetilde{D}_x^{-1,i,0}(c_n)]}$ is bounded in probability. Thus

$$\sup_{x_i;y_i\in[1-\widetilde{M}_\epsilon,\widetilde{M}_\epsilon]} \left|g\left(\widetilde{D}_x^{-1,i}[D_{x_i}(y_i)]\right) - g\left(D_x^{-1}[D_{x_i}(y_i)]\right)\right| = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)$$

Now we go back to (1.8.5), it is very straightforward to see our target equation holds

$$|\mathrm{E}^\star[g(Y^\star)|X = x] - \mathrm{E}[g(Y)|X = x]| = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)$$

□

Proof of Theorem 1.5.2.

Recall lemma 1.5.1, we have:

$$\begin{aligned}
\mathrm{E}^*\widehat{m}^*_{n,h}(x) &= \frac{1}{nh}\sum_{i=1}^{n}\mathrm{E}^*\Big[Y_i^*K\Big(\frac{X_i-x}{h}\Big)\Big] \\
&= \frac{1}{nh}\sum_{i=1}^{n}\Big[h\frac{f_X(x)}{f_X(x)+o_p(1)}\int K(u)du\mathrm{E}^*(Y^*|X=x)+O(h^2)\Big] \\
&= \frac{f_X(x)}{f_X(x)+o_p(1)}\mathrm{E}^*(Y^*|X=x)+O(h^2) \\
&= \widehat{m}_{n,h}(x)+O(h^2) \quad \text{in probability}
\end{aligned}$$

which implies $|\mathrm{E}^*\widehat{m}^*_{n,h}(x)-\widehat{m}_{n,h}(x)| = O_p(h^2)$ □

Proof of Theorem 1.5.3.

$$\begin{aligned}
&nh\Big|\mathrm{Var}^*[\widehat{m}^*_{n,h}(x)]-\mathrm{Var}[\widehat{m}_{n,h}(x)]\Big| \\
&= nh\Big|\frac{1}{(nh)^2}\sum_{i=1}^{n}\mathrm{Var}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]-\frac{1}{(nh)^2}\sum_{i=1}^{N}\mathrm{Var}\Big[Y_i\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]\Big| \\
&\le \frac{1}{nh}\sum_{i=1}^{n}\Big|\mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]^2-\Big(\mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]\Big)^2-\mathrm{E}\Big[Y_i\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]^2 \\
&\quad +\Big(\mathrm{E}\Big[Y_i\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]\Big)^2\Big| \\
&\le \frac{1}{nh}\sum_{i=1}^{n}\Big|\mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]^2-\mathrm{E}\Big[Y_i\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]^2\Big| \\
&\quad +\frac{1}{nh}\sum_{i=1}^{n}\Big|\Big(\mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]\Big)^2-\Big(\mathrm{E}\Big[Y_i\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]\Big)^2\Big| \\
&= \frac{1}{nh}\sum_{i=1}^{n}h\Big|f_X(x)\int\widetilde{K}^2(u)du\cdot(\mathrm{E}^*(Y^{*2}|X=x)-\mathrm{E}(Y^2|X=x))+O(h^2)\Big|+ \\
&\quad \frac{1}{nh}\sum_{i=1}^{n}\Big|\Big(\mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]+\mathrm{E}\Big[Y_i\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]\Big)\Big(\mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]-\mathrm{E}\Big[Y_i\widetilde{K}\Big(\frac{X_i-x}{h}\Big)\Big]\Big)\Big| \\
&\le f_X(x)\int\widetilde{K}^2(u)du\cdot|\mathrm{E}^*(Y^{*2}|X=x)-\mathrm{E}(Y^2|X=x)|+O(h^2)+\frac{1}{nh}\sum_{i=1}^{n}\Big|h\Big[f_X(x)\int\widetilde{K}(u)du \\
&\quad \cdot(\mathrm{E}^*(Y^*||x)+\mathrm{E}(Y|x))+O(h^2)\Big]\cdot h\Big[f_X(x)\int\widetilde{K}(u)du\cdot(\mathrm{E}^*(Y^*|x)-\mathrm{E}(Y|x))+O(h^2)\Big]\Big| \\
&\le f_X(x)\int\widetilde{K}^2(u)du\cdot|\mathrm{E}^*(Y^{*2}||x)-\mathrm{E}(Y^2|x)|+h\Big(f_X(x)\int\widetilde{K}(u)du\Big)^2|\mathrm{E}^*(Y^*|x)+ \\
&\quad \mathrm{E}(Y|x)|\cdot|\mathrm{E}^*(Y^*|x)-\mathrm{E}(Y|x)|
\end{aligned}$$

Here notice that

$$|\mathrm{E}^*(Y^*|x)|+\mathrm{E}(Y|x)| \le |\mathrm{E}^*(Y^*|x)|+|\mathrm{E}(Y|x)|$$

$$\begin{aligned}&\leq |\mathrm{E}(Y|x)| + |\mathrm{E}^*(Y^*|x) - \mathrm{E}(Y|x)| + |\mathrm{E}(Y|x)|\\ &< \infty\end{aligned}$$

Thus, recall theorem 1.3.2&1.4.8, we can easily get:

$$nh\Big|\mathrm{Var}^*[\widehat{m}^*_{n,h}(x)] - \mathrm{Var}[\widehat{m}_{n,h}(x)]\Big| = O_p\Big(h^2 + \frac{1}{\sqrt{nh}}\Big)$$

□

Proof of Theorem 1.5.4.

Define $u^*_i = \frac{1}{nh}Y^*_i\widetilde{K}\Big(\frac{X_i - x}{h}\Big)$ and $u_i = \frac{1}{nh}Y_i\widetilde{K}\Big(\frac{X_i - x}{h}\Big)$. Then we have $\widehat{m}^*_{n,h}(x) = \sum_{i=1}^n u^*_i$, $\widehat{m}_{n,h}(x) = \sum_{i=1}^n u_i$. Also, let

$$S^{*2}_n = \sum_{i=1}^n \mathrm{Var}^*(u^*_i) = \mathrm{Var}^*[\widehat{m}^*_{n,h}(x)], \quad S^2_n = \sum_{i=1}^n \mathrm{Var}(u_i) = \mathrm{Var}[\widehat{m}_{n,h}(x)]$$

Our goal is to show

$$\sqrt{nh}(\hat{m}^*_{n,h}(x) - \mathrm{E}[\hat{m}^*_{n,h}(x)]) \xrightarrow{d} N\left(0, \frac{\kappa\sigma^2(x)}{f(x)}\right)$$

which is equivalent to

$$\sqrt{nh}\sum_{i=1}^n (u^*_i - \mathrm{E}^* u^*_i) \xrightarrow{d} N\left(0, \frac{\kappa\sigma^2(x)}{f(x)}\right)$$

It is easy to see that $\{u^*_i\}$ are independent but not identical, thus we can use Lyapunov Central Limit Theorem to show the above equation. Actually, we only need to show Lyapunov Condition

$$\frac{\sum_{i=1}^n \mathrm{E}^*|u^*_i - \mathrm{E}^* u^*_i|^3}{S^{*3}_n} \xrightarrow{p.} 0$$

First, recall 1.2.2 and theorem 1.5.3, we have:

$$S^{*3}_n \leq |S^{*3}_n - S^3_n| + S^3_n = o_p(1) + S^3_n \xrightarrow{p.} \frac{1}{(nh)^{\frac{3}{2}}}\frac{\kappa\sigma^2(x)}{f_X(x)}$$

Simply, we just need

$$S^{*3}_n = O_p\Big(\frac{1}{(nh)^{\frac{3}{2}}}\Big) \tag{1.8.7}$$

$$
\begin{aligned}
&\mathrm{E}^*|u_i^* - \mathrm{E}^* u_i^*|^3 \\
&= \frac{1}{(nh)^3}\mathrm{E}^*\Big|Y_i^*\widetilde{K}\Big(\frac{X_i - x}{h}\Big) - \mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i - x}{h}\Big)\Big]\Big|^3 \\
&\le \frac{1}{(nh)^3}\Bigg\{\mathrm{E}^*\Big|Y_i^*\widetilde{K}\Big(\frac{X_i - x}{h}\Big)\Big|^3 + 3\Big|\mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i - x}{h}\Big)\Big]\Big| \cdot \mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i - x}{h}\Big)\Big]^2 \\
&\quad + 3\Big(\mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i - x}{h}\Big)\Big]\Big)^2 \cdot \mathrm{E}^*\Big|Y_i^*\widetilde{K}\Big(\frac{X_i - x}{h}\Big)\Big| + \Big|\mathrm{E}^*\Big[Y_i^*\widetilde{K}\Big(\frac{X_i - x}{h}\Big)\Big]\Big|^3\Bigg\} \\
&\le \frac{1}{(nh)^3}\Big\{h\widetilde{A}_3(x)\mathrm{E}^*|Y^*|x|^3 + 3h^2\widetilde{A}_1(x)\widetilde{A}_2(x)|\mathrm{E}^*(Y|x)^*| \cdot \mathrm{E}^*(Y^{*2}|x) + \\
&\quad 3h^3\widetilde{A}_1^3(x)(E^*(Y^*|x)^2 \cdot \mathrm{E}^*|Y^*|x| + h^3\widetilde{A}_1^3|\mathrm{E}^*(Y^*|x)|^3 + O(h^2)\Big\} \\
&= O\Big(\frac{1}{n^3h^2}\Big) \quad \text{in probability} \\
&= O_p\Big(\frac{1}{n^3h^2}\Big)
\end{aligned}
$$

Here we use $\mathrm{E}^*|Y^*|x|^k \le \Big|\mathrm{E}^*|Y^*|x|^k - E|Y|x|^k\Big| + E|Y|x|^k < \infty$.

$$\sum_{i=1}^{n}\mathrm{E}^*|u_i^* - \mathrm{E}^* u_i^*|^3 = O_p\Big(\frac{1}{(nh)^2}\Big)$$

Recall 1.8.7, we have

$$\frac{\sum_{i=1}^{n}\mathrm{E}^*|u_i^* - \mathrm{E}^* u_i^*|^3}{S_n^{*3}} = O_p\Big(\frac{1}{\sqrt{nh}}\Big)$$

By Lyapunov Central Limit Theorem, we have

$$\frac{\widehat{m}_{n,h}^*(x) - \mathrm{E}^*[\widehat{m}_{n,h}^*(x)]}{\sqrt{\mathrm{Var}^*[\widehat{m}_{n,h}^*(x)]}} = \frac{\sum_{i=1}^{n}(u_i^* - \mathrm{E}^* u_i^*)}{S_n^*} \xrightarrow{d} N(0,1)$$

With the result of theorem 1.5.3, $\mathrm{Var}^*[\widehat{m}_{n,h}^*(x)] \xrightarrow{p} \mathrm{Var}[\widehat{m}_{n,h}(x)] \to \frac{1}{nh}\frac{\kappa\sigma^2(x)}{f_X(x)}$, we can finally get

$$\sqrt{nh}(\hat{m}_{n,h}^*(x) - \mathrm{E}[\hat{m}_{n,h}^*(x)]) \xrightarrow{d} N\left(0, \frac{\kappa\sigma^2(x)}{f(x)}\right)$$

□

1.9 Acknowledgements

Chapter 1, is currently being prepared for the following paper prepared for future submission for publication: Asymptotic Validity of Bootstrap Confidence Intervals in Nonparametric Regression without an Additive Model. Liang Wang; Dimitris Politis. The dissertation author was the primary investigator and author of this material.

Chapter 2

Bias Reduction by Transformed Flat-top Series Estimator of Density on Compact Support

2.1 Introduction

Suppose observed sample $X_1, \ldots, X_n$ follows univariate density distribution $f_X(x)$ with compact support, without loss of generality, let $f_X(x)$ has support $[0, 1]$. The standard kernel density estimator is

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.1.1}$$

where $K(\cdot)$ is a kernel function, that usually is symmetric and satisfies $\int K(x)dx = 1$. h is the bandwidth. It is well known that such density kernel estimators suffer boundary effects since the kernel function will place positive weight outside of the bounded support. A numerous methods have been developed in the literature to reduce the boundary bias problem. Schuster [29], Silverman [30], Cline and Hart [31] considered the reflection method, which will reduce the boundary bias to $O(h^2)$ when the derivative of density at boundary is zero, otherwise the boundary effect is still slight improved with bias $O(h)$. The boundary kernel method, usually taken different kernel gives up symmetry property at points in

boundary region, was proposed and discussed by the following authors, Gasser and Muller [32], Gasser, Muller and Mammitzsch [33], Jones [34], Muller [35], Zhang and Karunamuni [36]. The local linear method is more general and fast implemented without many assumptions on density, was developed by Cheng et al. [37], Cheng [38], Zhang and Karunamuni [39]. The pseudo-data method, that generates pseudo-data by linear interpolation of order statistics to estimate density function with original data together, was proposed by Cowling and Hall [40].

The transformation method, also employed in this chapter, was first proposed and discussed by Wand, Marron and Ruppert [41], and then further developed by Wen and Wu [42]. Hall and Park [43] presented "empirical translation correction" method. The beta kernel estimator, that uses the beta density as kernel function was first proposed by Chen [44]. Jones and Henderson [45] discussed a Gaussian copula based estimator. Zhang et al. [46] combined the pseudo-data, transformation and reflection methods. Marron and Ruppert [47] combined transformation and reflection methods.

In this chapter, we proposed to combine the transformation method and flat-top series method. Politis and Romano [48] first proposed flat-top kernel spectral density estimator. Later Politis and Romano [49] extended flat-top kernel to multivariate density estimation and discovered this kind of kernel has infinite order property. After that, Politis proposed flat-top series estimator of the density with compact support [50] in 2000 and proposed adaptive bandwidth choice method for spectral density and probability density function with flat-top kernel [51] in 2003. Moreover, flat-top kernel method was further developed in [23, 52, 53] and discussed in [54, 55]. More details of flat-top kernel estimator and flat-top series estimator are introduced in section 2.2.

The rest of this section briefly introduces the spirit of transformed method in nonparametric density problem. Denote the transform function by $g(\cdot)$ that is a fixed one to one and monotonically increasing function, suppose random variable X follows the density $f_X(x)$, let $Y = g(X)$, then,

$$f_X(x) = f_Y[g(x)]g'(x)$$

If a density estimator $\hat{f}_Y$ based on $\{Y_i = g(X_i)\}_{i=1}^n$, then the transformed estimator is:

$$\hat{f}_X(x) = \hat{f}_Y[g(x)]g'(x) \tag{2.1.2}$$

If $g(x)$ is smooth, we have $\sup_{x\in[0,1]} g'(x) \leq M$ for some positive value M,

$$\begin{aligned}\left|\mathrm{E}[\hat{f}_X(x)] - f_X(x)\right| &= \left|\mathrm{E}\left[\hat{f}_Y[g(x)]g'(x)\right] - f_Y[g(x)]g'(x)\right| \\ &\leq M \cdot \left|\mathrm{E}\left(\hat{f}_Y[g(x)]\right) - f_Y[g(x)]\right| \end{aligned} \tag{2.1.3}$$

Similarly, we also have

$$\mathrm{Var}\left[\hat{f}_X(x)\right] \leq M^2 \cdot \mathrm{Var}\left[\hat{f}_Y[g(x)]\right] \tag{2.1.4}$$

Marron and Ruppert [47] intended to construct a function $g(\cdot)$ that satisfies $f_Y'(0) = 0$, and then used reflection method to obtain boundary bias order $O(h^2)$. This chapter constructs $g(\cdot)$ that satisfies some other conditions and then uses flat-top series estimator. The rest of the text is organized as follows: Section 2.2 introduces the infinite order flat-top kernel estimator and flat-top series estimator, along with some theorems. Section 2.3 presents the transformed flat-top series estimator and shows the higher order bias in the interior region. The selection of parameters and applications are discussed in section 2.4 and 2.5. Section 2.6 conducts simulation study and final comments are given in section 2.7.

2.2 Flat-top Estimator

2.2.1 Review of Flat-top Kernel Density Estimator

For regular kernel estimator (2.1.1) of density on $\mathbb{R}$,. we call a kernel function $K(\cdot)$ has q order, if $K(\cdot)$ has finite moments up to qth order, and moments of order up to $q-1$ equal to zero. Suppose a density function $f_X(x)$ has up to r continuous and bounded derivative, then it is known that

$$\mathrm{Bias}\left[\hat{f}_X(x)\right] = \mathrm{E}\hat{f}_X(x) - f_X(x) = c_{f,K}(x)h^k + o(h^k)$$

where $k = \min(q, r)$ and $c_{f,K}(x)$ is a bounded function depending on $K(\cdot), f$ and f'. Since regular positive kernel functions are usually density functions themselves and satisfy symmetry, they always have up to second order, and consequently the bias is always $O(h^2)$. In order to get a better bias order $O(h^k)$, where $k \geq 2$, the idea of choosing a kernel of higher order q is developed and even dates back to Parzen [56] and Bartlett [57]. More references on high-order kernels includes: [30, 33, 58, 59, 60, 61, 62, 63, 64, 65].

However, high order kernel method needs very complicated technique to estimate the smoothness of density and hard to select bandwidth. On the other side, infinity order flat-top kernel results in a bias of order $O(h^r)$ no matter how large is r. In addition Politis [51] proposed several simpler bandwidth selection methods for flat-top kernel than the ones for high order kernels. In the following definitions, we denote $\Omega(\cdot)$ as flat-top kernel function instead of $K(\cdot)$.

Definition 2.2.1. Let c be a positive value, the kernel Ω_c is said to be a member of the general family of univariate flat-top kernels of infinite order if

$$\Omega_c(x) = \frac{1}{2\pi} \int \omega_c(s) e^{-isx} ds$$

where the Fourier transform $\omega_c(s)$ satisfies the following properties:

(i) $\omega_c(s) = 1$ for all $|s| \leq c$;

(ii) $\int |\omega_c(s)|^2 ds < \infty$;

(iii) $\omega_c(s) = \omega_c(-s)$ for any $s \in \mathbb{R}$.

The above 3 properties will guarantee the infinite order, a finite variance and $\Omega_c(x)$ is real valued. In this chapter, c is a positive value less than 1, and

$$\omega_c(s) = \begin{cases} 1, & |s| \leq c \\ \eta_\omega(s), & c < |s| \leq 1 \\ 0, & |s| > 1 \end{cases} \tag{2.2.1}$$

$\eta_\omega(s)$ here determines the shape. Like regular kernel estimator, with a bandwidth h, we define

$$\Lambda_c(x) = \frac{1}{h} \Omega_c \left(\frac{x}{h} \right) \text{ and } \lambda_c(s) = \int \Lambda_c(x) e^{isx} dx = \omega_c(hs)$$

Denote $\phi_X(s) = \int e^{isx} f_X(x)dx$ the characteristic function, and let the sample characteristic function be $\hat{\phi}_X(s) = \frac{1}{n}\sum_i e^{isX_i}$, then flat-top kernel estimator of $f_X(x)$ is

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \Lambda_c(x - X_i) = \frac{1}{2\pi}\int \lambda_c(s)\hat{\phi}_X(s)e^{-isx}ds \tag{2.2.2}$$

Theorem 2.2.2 (Politis and Romano [49]). *If there is an $r > 0$, such that*

$$\int |s|^r |\phi_X(s)|ds < \infty \tag{2.2.3}$$

Assume that $n \to \infty$, and let $h \sim An^{-1/(2r+1)}$, for some constant $A > 0$, it follows that

$$\sup_{x\in\mathbb{R}} Bias\left[\hat{f}_X(x)\right] = o(h^r)$$

and

$$\sup_{x\in\mathbb{R}} \text{MSE}\left[\hat{f}_X(x)\right] = O\left(n^{-\frac{2r}{2r+1}}\right)$$

Here condition (2.2.3) implies that $f_X(x)$ has up to r bounded and continuous derivatives.

2.2.2 Flat-top Series Density Estimator

Now suppose $f_X(x)$ is very smooth density function on $[0,1]$, and $\tilde{f}_X(x)$ is periodic function with period 1 that is the extension of $f_X(x)$ on the whole real line. Let the characteristic function here be $\phi_X(s) = \int_0^1 e^{isx} f_X(x)dx$, and the sample characteristic function still be $\hat{\phi}_X(s) = \frac{1}{n}\sum_i e^{isX_i}$, recall Fourier series formula:

$$\tilde{f}_X(x) = \sum_{s\in\mathbb{Z}} e^{-i2\pi sx}\phi_X(s)$$

and (2.2.2): The definition of flat-top series estimator in Politis [50] is:

$$\hat{f}_X(x) = \sum_{s\in\mathbb{Z}} \lambda_c(s)e^{-i2\pi sx}\hat{\phi}_X(s) \tag{2.2.4}$$

This estimator is also a periodic function with period 1 on the whole real line, but we only need the part on interval $[0,1]$. Then we have the following theorems:

Theorem 2.2.3. *If $f_X(x)$ is continuous and differentiable density function defined $[0,1]$, $0 < |f_X(1) - f_X(0)| < \infty$, and $f_X^{(1)} \in L^1[0,1]$, then*

$$\sup_{x\in[0,1]} \text{Var}\left[\hat{f}_X(x)\right] = O\left[\frac{1}{nh}\log\left(\frac{1}{h}\right)\right] \tag{2.2.5}$$

According to the proof in appendix, we can also obtain the following corollary:

Corollary 2.2.4. *If $f_X(x)$ is continuously differentiable density function defined $[0,1]$, $|f_X(1) - f_X(0)| = 0$, $0 < |f_X^{(1)}(1) - f_X^{(1)}(0)| < \infty$ but is finite, both of $f_X^{(1)}, f_X^{(2)} \in L^1[0,1]$, then*

$$\sup_{x\in[0,1]} \text{Var}\left[\hat{f}_X(x)\right] = O\left(\frac{1}{nh}\right) \tag{2.2.6}$$

Theorem 2.2.5 (Politis [50])**.** *If there exists $r > 0$, such that*

$$\sum_{s\in\mathbb{Z}} |s|^r \phi_X(s) < \infty \tag{2.2.7}$$

let $h \sim An^{-1/(2r+1)}$ from some constant $A > 0$, it follows that

$$\sup_{x\in[0,1]} \textit{Bias}\left[\hat{f}_X(x)\right] = o(h^r)$$

and

$$\sup_{x\in[0,1]} \text{MSE}\left[\hat{f}_X(x)\right] = O\left(n^{-\frac{2r}{2r+1}}\right) \tag{2.2.8}$$

Notice the condition (2.2.7) here implies $\tilde{f}_X$ has r bounded and continuous derivatives, more particularly, it means f_X has r bounded and continuous derivatives

$$f_X(0) = f_X(1), f_X^{(1)}(0) = f_X^{(1)}(1), f_X^{(2)}(0) = f_X^{(2)}(1), \ldots, f_X^{(r)}(0) = f_X^{(r)}(1),$$

Actually, this condition is too strong. If we consider rectangular flat-top function, more specifically, $\omega_c(s)$ with $c = 1$ and $\eta_\omega = 0$, we could relax the condition as follows:

Theorem 2.2.6. *Let $f_X(x)$ denote the density function on $[0,1]$ and $\tilde{f}_X(x)$ be the periodic extension of $f_X(x)$. If $\tilde{f}_X(x)$ has $(r-1)$th bounded and continuous derivative ($r \geq 1$), and $\tilde{f}_X^{(r)}(x)$ is continuous except the jump discontinuities at $x \in \mathbb{Z}$. More specifically, $f^{(r)}(x)$ is continuous on $[0,1]$, but $0 < |f_X^{(r)}(0) - f_X^{(r)}(1)| < \infty$. Moreover, $f_X^{(r+1)}(x)$ is continuous on $[0,1]$. Then for $h \sim An^{-1/(2r+1)}$, we still have result (2.2.8) and bias*

$$\sup_{x\in[0,1]} \textit{Bias}\left[\hat{f}_X(x)\right] = O(h^r)$$

Proof is provided in appendix.

2.3 Transformation-based Flat-top Series Estimator

Although theorem 2.2.5 and theorem 2.2.6 guarantee very good properties of flat-top series estimator, the conditions are too strong. In many situations, density f_X would not satisfy the condition that its periodic extension $\tilde{f}_X$ has up to rth bounded and continuous derivative. In this section, we will assume,

Condition 2.3.1. *f_X is rth continuously differentiable and bounded on $[0,1]$ ($r \geq 2$), but $f_X(0) \neq f_X(1)$.*

This implies $\tilde{f}_X$ is discontinuous.

2.3.1 Parametric Transform Function

In this chapter, the transform function is estimated from a parametric family containing a "target" transformation such that density function satisfies $f_Y(0) = f_Y(1)$, where Y is the transformed random variable, and density function f_X satisfies Condition 2.3.1. Nevertheless, the proof of theorem 2.2.6 in appendix shows that if $f_Y^{(1)}(0) \neq f_Y^{(1)}(1)$, then the bias of flat-top series estimator has order at most $O(h)$ universally. Thus, we need to modify the "target" transformation to be:

$$f_Y(0) = f_Y(1); \quad f_Y^{(1)}(0) = f_Y^{(1)}(1) \tag{2.3.1}$$

Let $\Theta \subset \mathbb{R}^k$ and $G = \{g_\theta, \theta \in \Theta\}$ be a parametric family of transformations with domain $[0,1]$. Then we need the following condition throughout:

Condition 2.3.2. *Each $g_\theta \in G$ is three times continuously differentiable with respect to x and has a strictly positive derivative from $[0,1]$ to $[0,1]$. Derivatives at 0 and 1 are one sided. And there exists a $K > 0$, s.t.*

$$\max_{j=0,1,2,3} \{ \sup_{x\in[0,1]} |g_\theta^{(j)}(x) - g_{\theta_0}^{(j)}(x)| \} \leq K||\theta - \theta_0|| \tag{2.3.2}$$

where derivatives of g_θ are with respect to x but not θ, and $||\cdot||$ denotes the Euclidean norm on $\mathbb{R}^k$.

One example of this condition is satisfied, is any polynomial with coefficients as linear functions of θ. As (2.1.2) in section 2.1, but since g is no longer fixed, write $f_Y(\cdot; g)$ in stead of f_Y, flat-top series estimator $\hat{f}_Y(\cdot; g)$ instead of $\hat{f}_Y$, characteristic function $\phi_Y(\cdot; g)$ instead of ϕ_Y, and $\hat{\phi}_Y(\cdot; g)$ instead of $\hat{\phi}_Y$, then we formally define the transformed flat-top series estimator is

$$\hat{f}_X(x; g) = \hat{f}_Y[g(x); g]g'(x) \tag{2.3.3}$$

Where $g \in G$. Now assume $Y_{\theta_0} = g_{\theta_0}(X)$ satisfies (2.3.1), for simplicity, denote Y_{θ_0} by Y_0, and g_{θ_0} by g_0. We shall be concerned that θ_n is the estimator of θ_0 based on $\{X_i\}_{i=1}^n$. Let $Y_{\theta_n} = g_{\theta_n}(X)$, similarly as above, let Y_n denote Y_{θ_n}, and g_n denote g_{θ_n}. The actual transformed flat-top series estimator is

$$\begin{aligned} \hat{f}_X(x; g_n) &= \hat{f}_{Y_n}[g_n(x); g_n]g_n'(x) \\ &= \hat{f}_Y[g(x); g]\big|_{g=g_n} g_n'(x) \end{aligned} \tag{2.3.4}$$

Where

$$\hat{f}_{Y_n}[g_n(x); g_n] = \sum_{s\in\mathbb{Z}} \lambda_c(s) e^{-i2\pi s g_n(x)} \hat{\phi}_{Y_n}(s; g_n) \tag{2.3.5}$$

As in Marron and Ruppert [47], we let the expectation of flat-top series estimator of f_{Y_n} given g_n be:

$$\mathrm{E}\hat{f}_{Y_n}(y; g_n) = \left[\mathrm{E}\hat{f}_Y(y; g)\right]\Big|_{g=g_n}$$

$$= \mathrm{E}\left[\sum_{s\in\mathbb{Z}} \lambda_c(s)\hat{\phi}_Y(s;g)e^{i2\pi sy}\right]\Bigg|_{g=g_n}$$
$$= \sum_{s\in\mathbb{Z}} \lambda_c(s)\phi_{Y_n}(s;g_n)e^{i2\pi sy} \tag{2.3.6}$$

which is a sort of modified partial sum of Fourier series, where $\phi_{Y_n}(\cdot;g_n)$ is the characteristic function of Y_n.

Theorem 2.3.3. *Suppose Condition 2.3.1 holds with $r = 2$, $\theta_0 \in \Theta$, and Condition 2.3.2 holds. Define $G_{\triangle_n} = \{g_n : ||\theta_n - \theta_0|| \le \triangle_n$ and $\theta_n \in \Theta\}$. Let $\triangle_n = \frac{1}{\log 1/h}$, where $h = h(n) \to 0$ as $n \to \infty$ for some function $h(\cdot)$, then for any $g_n \in G_{\triangle_n}$*

$$\sup_{y\in[0,1]} \left|\hat{f}_{Y_n}(y;g_n) - \mathrm{E}\hat{f}_{Y_n}(y;g_n)\right| = o\left(\frac{1}{\sqrt{nh}}\right) \tag{2.3.7}$$

Notice that if $||\theta_n - \theta_0|| = o_p(\triangle_n)$, then $\mathrm{P}\left(g_n \in G_{\triangle_n}\right) \to 1$, so by theorem 2.3.3,

$$\sup_{y\in[0,1]} \left|\hat{f}_{Y_n}(y;g_n) - \mathrm{E}\hat{f}_{Y_n}(y;g_n)\right| = o_p\left(\frac{1}{\sqrt{nh}}\right) \tag{2.3.8}$$

Suppose $\theta_n \to \theta_0$ at rate $q(n)$ where $q(n) \to 0$ as $n \to \infty$ for some function $q(\cdot)$, then Y_n does not really satisfies (2.3.1). By Condition 2.3.2, it is very easy to show that

$$|f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)| = O[q(n)], \quad |f'_{Y_n}(1;g_n) - f'_{Y_n}(0;g_n)| = O[q(n)] \tag{2.3.9}$$

Similarly, if $\theta_n \to \theta_0$ in probability at rate $q(n)$, then

$$|f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)| = O_p[q(n)], \quad |f'_{Y_n}(1;g_n) - f'_{Y_n}(0;g_n)| = O_p[q(n)] \tag{2.3.10}$$

By definition 2.2.1, (2.3.6) can be seen as a modified partial sum of Fourier series. Although Gibbs Phenomenon will be caused by (2.3.9), it is well known that the difference between the true function and partial sum of Fourier series is bounded by the size of the jump of discontinuity at discontinuous points, in another word, the difference between $\mathrm{E}\hat{f}_{Y_n}(y;g_n)$ and $f_{Y_n}(y;g_n)$ at 0 and 1, is also bounded by $|f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)| = q(n)$, see Lemma 2.3.7. This implies

flat-top series estimator $\hat{f}_{Y_n}(y; g_n)$ always has the order of bias $q(n)$, actually the same rate as $\theta_n \to \theta_0$, at boundary points. Fortunately, we are able to improve the bias to higher order in the interior of the support. Before introducing the order of bias, recalling (2.2.1), we first introduce the following theorems to understand the performance of transformation-based flat-top series estimator when $c = 1$.

2.3.2 Rectangular Flat-top Function

$$\omega(s) = \begin{cases} 1, & |s| \leq 1 \\ 0, & |s| > 1 \end{cases}$$

and $\lambda_c(s) = \omega_c(hs)$, it implies

$$\mathrm{E}\hat{f}_{Y_n}(y; g_n) = \sum_{s \leq 1/h} \phi_{Y_n}(s) e^{i2\pi sy} = S_{N_h} \tilde{f}_{Y_n}(y; g_n) \tag{2.3.11}$$

the modified partial sum of Fourier series now becomes regular partial sum of Fourier series, and $N_h = \lfloor 1/h \rfloor$, here $\tilde{f}_{Y_n}(y; g_n)$ is the periodic extension of $f_{Y_n}(y; g_n)$. The following lemmas are showing the bias at boundary points:

Lemma 2.3.4. *Define*

$$\phi_0(x) = \frac{1}{2}\left(\frac{1}{2} - x\right), 0 \leq x \leq 1$$

Where $\phi(x)$ is periodic extension function of $\phi_0(x)$. Let $\phi_n(x)$ is the n^{th} partial sum of Fourier sereis:

$$\phi_n(x) = \sum_{k=1}^{n} \frac{\sin(2k\pi x)}{2k\pi}$$

For any $x \notin \mathbb{Z}$, i.e. x is away from the discontinuity of $\phi(x)$, we have

$$|\phi(x) - \phi_n(x)| = O\left(\frac{1}{n}\right)$$

Remark 2.3.5. Notice that if $x \in [a, b]$ where $[a, b]$ is an interval away from discontinuity, then we can also have:

$$\sup_{x \in [a,b]} |\phi(x) - \phi_n(x)| = O\left(\frac{1}{n}\right) \tag{2.3.12}$$

Remark 2.3.6. If $f(x)$ is discontinuous at a point a, then we can define a continuous function

$$f^*(x) = f(x) - 2[f(a^+) - f(a^-)]\phi(x-a)$$

Clearly $f^*(a) = \frac{f(a^+)+f(a^-)}{2}$ and $f^*(x)$ is now continuous, very helpful for showing the next lemma. The reason we construct $f^*(x)$ by $\phi(x)$ instead of by a step function is that $\phi_n(x)$ has a better form and more convenient for the proof.

Lemma 2.3.7 (Theorem F, E. Hewitt [66]). *Let f be a real-valued periodic function on the real line $\mathbb{R}$ with period 1, and suppose that f and its derivative f' are both continuous except for a finite number of finite jump discontinuities in the interval [0,1]. let $S_n f(x)$ be the n^{th} partial sum of the Fourier series of the function f, computed at the point x. Let a be a point of discontinuity of f. The distance between overshoot and undershoot of Gibbs Phenomenon at point a is at most $\frac{2}{\pi}Si(\pi)|f(a^+) - f(a^-)|$, where $Si(x) = \int_0^x \frac{sin(t)}{t}dt$.*

Remark 2.3.8. Recall (2.3.11), that the expectation of rectangular flat-top series estimator is Fourier series. Lemma 2.3.7 essentially tells us, for any $0 < a \leq b < 1$:

$$\begin{aligned}\sup_{y\in[0,a)\bigcup(b,1]} \left|\mathrm{E}\hat{f}_{Y_n}(y;g_n) - f_{Y_n}(y;g_n)\right| &\leq \frac{2}{\pi}Si(\pi)|\tilde{f}_{Y_n}(0^+;g_n) - \tilde{f}_{Y_n}(0^-;g_n)| \\ &= \frac{2}{\pi}Si(\pi)|f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)|\end{aligned}$$

With (2.3.10), then we have the following theorem:

Theorem 2.3.9. *If $f_X(x)$ is continuously differentiable function on $[0,1]$, and $\theta_n - \theta_0 = O_p(h^i), i = 1,2$, then for any $0 < a \leq b < 1$:*

$$\sup_{y\in[0,a)\bigcup(b,1]} \left|\mathrm{E}\hat{f}_{Y_n}(y;g_n) - f_{Y_n}(y;g_n)\right| = O_p(h^i), \quad i = 1,2$$

Lemma 2.3.10. *Let f be a real valued function on the $[0,1]$. $\tilde{f}$ is periodic extension of f. If $\tilde{f}(x)$ is continuous (i.e. $f(0) = f(1)$), $\tilde{f}^{(1)}(x)$ is continuous except jump discontinuities at $x \in \mathbb{Z}$, in another word, $f^{(1)}(x)$ is continuous on $[0,1]$, but $0 < |f^{(1)}(0) - f^{(1)}(1)| < \infty$. Moreover, $f^{(2)}$ and $f^{(3)}$ is continuous on $[0,1]$. Then*

$$\sup_{x\in[0,1]} |S_n f(x) - f(x)| \leq \frac{K_1}{n}|f^{(1)}(1) - f^{(1)}(0)| + O\left(\frac{1}{n^2}\right)$$

Notice all the above lemmas and theorems are discussing the rectangular flat-top series estimator for transformed random variable Y_n. Recall (2.3.4), the following theorem is back to random variable X:

Theorem 2.3.11. *Assume Condition 2.3.1 and Condition 2.3.2 hold, $h = An^{-1/5}$, if $\theta_n = \theta_0 + O_p(h)$, then for any $0 < a \leq b < 1$, the transformed rectangular flat-top series density estimator converges to true density function:*

$$\sup_{x\in[a,b]} \left|\hat{f}_X(x; g_n) - f_X(x)\right| = O_p\left(\frac{1}{\sqrt{nh}} + h^2\right) = O_p\left(n^{-2/5}\right)$$

Remark 2.3.12. Although we only need $\theta_n = \theta_0 + O_p(h)$ in the theorem above, to get bias order $O(h^2)$ in any closed interval inside $[0, 1]$, by theorem 2.3.9, the bias order at boundary is still $O(h)$. Thus we ought to employ θ_n such that $\theta_n = \theta_0 + O_p(h^2)$ to obtain:

$$\sup_{x\in[0,1]} \left|\hat{f}_X(x; g_n) - f_X(x)\right| = O_p\left(\frac{1}{\sqrt{nh}} + h^2\right) = O_p\left(n^{-2/5}\right)$$

If $h = An^{-1/5}$.

2.3.3 Infinitely Differentiable Flat-top Function

Recall (2.2.1), $\lambda_c(s) = \omega_c(hs)$ and (2.3.6) with $c < 1$ and some function η_ω, the mean of flat-top series estimator is a modified partial sum of Fourier series. In fact, it is well known that when the function is discontinuous, regular n^{th} partial sum of Fourier series has bad convergent properties caused by Gibbs Phenomenon, that not only failure of convergence at discontinuity, but also slow global convergent rate $O(1/n)$. This is essentially caused by discontinuity of $\omega_c(s)$. This section introduces the situation when $\omega_c(s)$ is infinitely differentiable, first see the definition of η_ω (see McMurry and Politis, 2004),

$$\eta_\omega(s) = \exp\left(\frac{-b}{(|s|-1)^2} \exp\left[\frac{-b}{(|s|-c)^2}\right]\right) \tag{2.3.13}$$

Where $b > 0$ and $c < |s| < 1$. c determines the width of region that flat-top function is 1, and b allows us to alter the shape of η_ω. The selection of these 2

parameters are discussed later. Apparently function η_ω is able to connect the region where $\omega(s)$ is 0 and the region $\omega(s)$ is 1 in a manner such that $\omega(s)$ is infinitely differentiable for all s, even including where $|s| = c$ and $|s| = 1$. Remember again the mean of flat-top series estimator is a modified partial sum of Fourier series, and see the following lemma:

Lemma 2.3.13 (Theorem 3.4, Gottlieb and Shu [67])**.** *Let f be piecewise C^{r+1} function with one point of discontinuity ξ, and y is one point in $[0,1]$ away from ξ.*

$$f_n^\omega(y) = \sum_{s\in\mathbb{Z}} \omega_c\left(\frac{s}{n}\right) c_n(f) e^{i2\pi sy}$$

Where $c_n(f)$ is the coefficient of Fourier series.

$$|f(y) - f_n^\omega(y)| \leq \frac{C\cdot K(f)}{n^r d(y)^r} + \frac{C||f||_{L^2[0,1]}}{n^{r+1/2}} \tag{2.3.14}$$

where C is some constant, $d(y) = \min_{k=-1,0,1}|y-\xi+k|$, and

$$K(f) = \sum_{l=0}^{r} d(y)^l \left[f^{(l)}(\xi^+) - f^{(l)}(\xi^-)\right] \int_{-\infty}^{\infty} \left|G_l^{(r+1-l)}(\eta)\right| d\eta$$

Where $G_l(\eta) = \frac{\omega_c(\eta)-1}{\eta^l}$.

Notice $\int_{-\infty}^{\infty} |G_l^{(r+1-l)}(\eta)|d\eta < \infty$ because $\omega_c(\eta) - 1 = 0$ here when $|\eta| \leq c$. This lemma shows that convergence rate of this modified partial sum of Fourier series is $O(n^{-r})$ when y is away from discontinuity ξ.

Remark 2.3.14. One amazing property of such modified partial sum of Fourier series, by definition of $K(f)$, is that its convergent rate $O(n^{-r})$ is not influenced by the continuity of $f^{(l)}(\cdot)$ at ξ. It implies when practitioners transform the random variable, it does not have to satisfy (2.3.1). More specifically, suppose $\xi = 0$ and discussing $\tilde{f}_{Y_n}(\cdot; g_n)$, the convergent rate $O(N_h^{-r})$ of bias remains, even when $\tilde{f}_{Y_n}(\cdot; g_n)$ and $\tilde{f}^{(1)}_{Y_n}(\cdot; g_n)$ are discontinuous at 0, or in another word, even when $|f^{(i)}_{Y_n}(1; g_n) - f^{(i)}_{Y_n}(0; g_n)| \neq 0$ where $i = 0, 1$. Thus the bias of the point away from discontinuity always has order $O(N_h^{-r}) = O(h^r)$. But for the points around discontinuity, it is easy to show that Lemma 2.3.7 and Theorem 2.3.9 still apply for infinitely differentiable flat-top function, and therefore we still suggest

to employ an estimator $\theta_n = \theta_0 + O_p(h^2)$ so that the bias at boundary has order $O_p(h^2)$.

Remark 2.3.15. The definition of $K(f)$ implies the lemma does not work for the density function which is infinite at boundary, or any order derivative is infinite at boundary.

Theorem 2.3.16. *Assume Condition 2.3.2 holds, $f_X(x)$ is the density function on $[0,1]$ and $f_X(0) \neq f_X(1)$. Moreover, $f_X(x)$ is rth continuously differentiable and bounded function on $[0,1]$. If $\theta_n = \theta_0 + O_p(h^2)$, then for any $x \in (0,1)$, the transformed infinitely differentiable flat-top series estimator converges:*

$$\hat{f}_X(x; g_n) = f_X(x) + O_p\left(\frac{1}{\sqrt{nh}} + h^r\right)$$

for $x = 0, 1$

$$\hat{f}_X(x; g_n) = f_X(x) + O_p\left(\frac{1}{\sqrt{nh}} + h^2\right)$$

In the following sections, we are more interested in the higher order bias and therefore mainly discuss the transformed infinitely differentiable flat-top series estimator.

2.4 Selection of Parameters

2.4.1 Selection of b and c

Recall infinite differentiable flat top function η_ω 2.3.13, which is first introduced by McMurry and Politis [23], See Fig.2.1 Where they advised a small value of c is preferred, because large c will make η_ω more rectangular and thus undesirable. We find large b has the same problem. η_ω is supposed to correspond to a kernel with very large side lobes, similar to those of the Dirichlet kernel. As a result, in application we advise to pick $b = 1/4$ and $c = 0.05$.

2.4.2 Selection of Bandwidth h

The problem of the bandwidth selection in the whole compact support needs further study. The first question arises by theorem 2.3.16, that is the bias

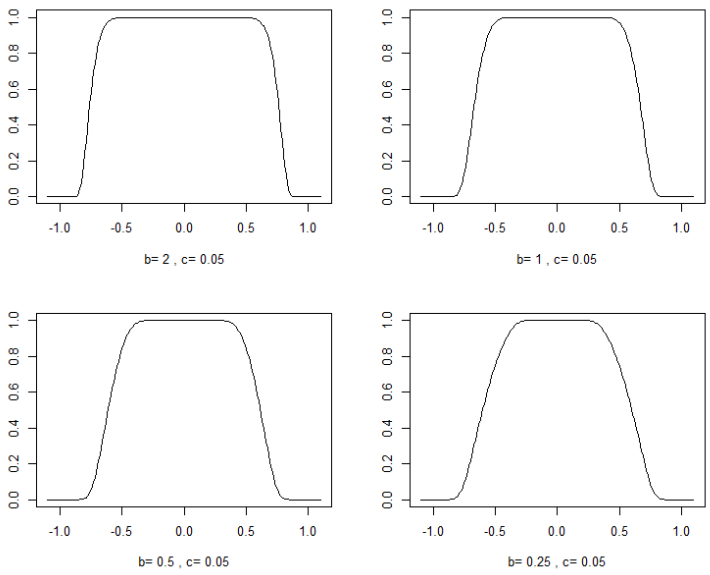


Figure 2.1: $\eta_\omega(s)$ with $c = 0.05$ and different b

orders at boundary area and interior area are different if $r > 2$. If a global bandwidth $h_1 = An^{-1/(2r+1)}$ is selected, where A_1 is a constant, then in the interior area we have

$$\hat{f}_X(x; g_n) - f_X(x) = O\left(n^{-\frac{r}{2r+1}}\right)$$

But in the boundary area,

$$\hat{f}_X(x; g_n) - f_X(x) = O\left(n^{-\frac{2}{2r+1}}\right)$$

Which is worse than the other boundary correction kernel estimator with bandwidth $h_2 = A_2 n^{-1/5}$, where A_2 is a constant:

$$\hat{f}_X(x) - f_X(x) = O\left(n^{-\frac{2}{5}}\right)$$

On the other hand, if we select two different bandwidths for interior region and boundary region, then discontinuity issue takes place at the border of these two regions. In application, we recommend to employ h_2 that optimize $||\theta_n - \theta_0||$, see condition 2.3.2. Thus,

$$\hat{f}_X(x; g_n) - f_X(x) = O\left(n^{-\frac{2}{5}}\right), \quad x = 0, 1$$

In the interior area, that $[h_2, 1-h_2]$ is advisable, if we select a bandwidth h_2 different from h_1, then define a local bandwidth $h = h(t)$, such that

$$h(t) = \begin{cases} h_1, & t = 0, 1 \\ h_2, & h_1 \leq t \leq 1 - h_1 \\ \gamma_h(t), & 0 < t < h_1, \text{ or } 1 - h_1 < t < 1 \end{cases}$$

Where $\gamma_h(t)$ makes $h(t)$ be continuous function. $\gamma_h(t)$ as a linear function is suggested in this chapter, and can be smoother if necessary. Selection of the bandwidth h_1 in the interior region is an open question. As in Politis [51], we can make a practical recommendation. In fact, as the trick applied in the proof, the inverse of bandwidth of flat-top series estimator performs as a threshold in the frequency domain. If the extension of density function, $\tilde{f}_X(x)$ is very smooth, the characteristic function, as the basis of frequency domain, $\phi_X(s)$ will decay very fast and be negligible when $|s| > \frac{1}{h}$. The bandwidth thus should be chosen so that the characteristic function at low frequency to pass without disturbed, while damping out the characteristic function at high frequency. With this in mind, according to Politis [51], we propose the following rule of thumb.

1. Pick h_2 to minimize $||\theta_n - \theta_0||$, obtain transform function $g_n(\cdot)$ and transformed observation $Y_n = g_n(X)$;

2. Let $\rho(s) = \phi_{Y_n}(s)/\phi_{Y_n}(0)$, and $\hat{\rho}(s) = \hat{\phi}_{Y_n}(s)/\hat{\phi}_{Y_n}(s)$. Let $\hat{m}$ be the smallest positive real number such that $|\hat{\rho}(\hat{m}+s)| < 2\sqrt{\log n/n}$, for all $s \in (0, K_n)$, where K_n is a positive nondecreasing real-valued function of n such that $K_n = o(\log n)$, Then $h_1 = 1/(2\hat{m})$.

3. Estimate $f_{Y_n}(\cdot; g_n)$ by (2.3.5) with infinite differentiable flat-top function and local bandwidth $h(t)$. Finally get estimator of $f_X(\cdot)$ by (2.3.4).

Unfortunately, the adaptive bandwidth choice method might not always be the most appropriate for flat-top series transformed flat-top series estimator when the density function is not smooth enough. Moreover, the periodic extension of estimated transformed function $f_{Y_n}(y; g_n)$ is not really continuous, and thus its

characteristic function might not always decay fast. Thus we present to estimate bandwidth by least square cross validation, to minimize

$$\int_0^1 [\hat{f}_{Y_n}(y;g_n) - f_{Y_n}(y;g_n)]^2 dy = \int_0^1 \hat{f}^2_{Y_n}(y;g_n)dy - 2\int_0^1 \hat{f}_{Y_n}(y;g_n)f_{Y_n}(y)dy + \int_0^1 f^2_{Y_n}(y;g_n)dy$$

The last term on the right is independent from h. The first term is easy to compute because this integral is on closed interval $[0,1]$. We will use leave-one-out method to estimate the second term:

$$\int_0^1 \hat{f}_{Y_n}(y;g_n)f_{Y_n}(y;g_n)dy \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}^{-i}_{Y_n}(y_i;g_n)$$

where $\hat{f}^{-i}_{Y_n}$ is the flat-top series estimator without observation y_i.

2.5 Application

2.5.1 Polynomial Transformation

In this section, we propose polynomial function as transformation with the asymptotical properties discussed in the previous section. Since we mainly discuss estimator with infinite differentiable flat-top function, by lemma 2.3.13 the transformation target is just

$$f_Y(0) = f_Y(1)$$

Let $Y_0 = g_0(X)$ be the desired transformation random variable. Since g_0 is one to one function from $[0,1]$ to $[0,1]$, we also have $X = g_0^{-1}(Y)$. Then by theorem for derivative of inverse function:

$$f_{Y_0}(y) = \frac{f_X[g_0^{-1}(y)]}{g_0'[g_0^{-1}(y)]} \tag{2.5.1}$$

which implies

$$f_{Y_0}(0) = \frac{f_X(0)}{g_0'(0)} \quad f_{Y_0}(1) = \frac{f_X(1)}{g_0'(1)}$$

Remark 2.5.1. If $f_X(0) > f_X(1) = 0$, then we need $g_0'(0)$ to be 0 or $g_0'(1)$ to be infinite, so that the equation above holds. In addition, the rate of $g_0'(t) \to 0$ as $t \to 0$ is also very important. Thus the situation $f_X(0) > f_X(1) = 0$ is discussed in next subsection. The following discussion are all based on assumption $f_X(0)$ and $f_X(1)$ both positive.

A polynomial function and its coefficients will be selected by estimating $f_X(0), f_X(1)$. In fact, we need to solve all the following equations:

$$\begin{cases} g_0(0) = 0 \\ g_0(1) = 1 \\ g_0'(x) \geq 0, 0 \leq x \leq 1 \\ f_{Y_0}(0) = f_{Y_0}(1) \end{cases}$$

If considering quadratic polynomial $g_0(x) = ax^2 + bx$, $g_0(0) = 0$ is satisfied. We need $a + b = 1$ to make $g_0(1) = 1$ holds. In addition, since $f_X(x)$ is always non-negative, derived by the last equation, we have

$$\frac{f_X(0)}{b} = \frac{f_X(0)}{g_0'(0)} = \frac{f_X(1)}{g_0'(1)} = \frac{f_X(1)}{2a + b}$$

And

$$a = \frac{f_X(1) - f_X(0)}{f_X(1) + f_X(0)} \quad b = \frac{2f_X(0)}{f_X(1) + f_X(0)}$$

By some simple analysis, $g_0'(x) \geq 0$ always holds for $x \in [0, 1]$, no matter $f_X(0) > f_X(1)$ or $f_X(0) < f_X(1)$. It implies quadratic polynomial g_0 always guarantees the above equation has valid solution. It is easy to see that if

$$\hat{f}_X(0) = f_X(0) + O_p(h^2)$$

$$\hat{f}_X(0) = f_X(1) + O_p(h^2)$$

and let $g_n(x) = a_n x^2 + b_n$, where

$$a_n = \frac{\hat{f}_X(1) - \hat{f}_X(0)}{\hat{f}_X(1) + \hat{f}_X(0)} \quad b_n = \frac{2\hat{f}_X(0)}{\hat{f}_X(1) + \hat{f}_X(0)}$$

Then $g_n(x) = g_0(x) + O_p(h^2)$ for any fixed x. In this chapter, the simple boundary correction method by Jones [34], available in R package **evmix**, is applied

to estimate $f_X(0)$ and $f_X(1)$. This method is essentially equivalent to the kernel weighted local linear fitting at the boundary, thus implemented fast and is able to provide very accurate estimate at boundaries among many existing methods.

2.5.2 Algorithm

Notice that the polynomial transform method only works for the situation $f_X(0)$ and $f_X(1)$ are different and positive. If one of them is zero, then any polynomial transform method is invalid. Politis [68] addressed a nonparametric hypothesis testing method, that could test whether a density distribution has zero value at point of interest. Thus, we propose the following procedures and modified methods if the assumption $f_X(0) \neq f_X(1) > 0$ is violated:

(1) Testing hypothesis:

$$H_0^1 : f_X(0) = 0; \qquad H_1^1 : f_X(0) > 0$$
$$H_0^2 : f_X(1) = 0; \qquad H_1^2 : f_X(1) > 0$$
$$H_0^3 : f_X(0) = f_X(1); \; H_1^3 : f_X(0) \neq f_X(1)$$

The last hypothesis is feasible based on the theorems provided in Jones [34]. Apparently, if we reject all the three null hypothesis, then the transformed flat-top series estimator discussed in previous sections can be applied directly.

(2) Fail to reject both H_0^1 and H_0^2:
Then we assume $f_X(0) = f_X(1) = 0$, and therefore transformation is not necessary. Infinite differentiable flat-top series estimator could be applied directly. Notice that flat-top series estimator behave similar to other higher order kernel density estimator and might produce negative value, thus a final step to re-normalize the density estimator is required.

(3) Reject both H_0^1 and H_0^2 , but fail to reject H_0^3:
Then we assume $f_X(0) = f_X(1) > 0$, and infinite differentiable flat-top series estimator could be directly applied. Notice that it might produce nega-

tive value as well, and therefore re-normalization is required after obtaining the density estimate.

(4) Reject only one of H_0^1 and H_0^2:
Without loss of generality, here we assume $f_X(1) > f_X(0) = 0$. The polynomial transformation introduced previously does not work and needs to be modified. Let $Z = X^k$, where $k > 1$ is an integer, and first start with $k = 2$. Now we need to test again hypothesis $H_0^4 : f_Z(0) = 0$. Increase k one by one if we keep failing to reject this null hypothesis. Suppose null hypothesis H_0^4 is finally rejected when $k = k_0$, and also reject $H_0^5 : f_Z(0) = f_Z(1)$. quadratic polynomial transformed flat-top series estimator could be now applied on datasets $\{Z_i\}_{i=1}^n$, i.e. $Y_n = g_n(Z)$, and

$$\hat{f}_Z(z; g_n) = \hat{f}_{Y_n}[g_n(z); g_n] \cdot g_n'(z)$$

Since $Z = X^{k_0}$ is strictly increasing function on $X \in [0, 1]$, the final estimator of $f_X(x)$ is

$$\hat{f}_X(x; g_n, k_0) = \hat{f}_Z(x^{k_0}; g_n, k_0) \cdot k_0 x^{k_0 - 1}$$

If $f_X(0) > f_X(1) = 0$ holds, then let $Z = (1 - X)^{k_0}$, which is a decreasing function on $[0, 1]$, thus the true density function is

$$f_X(x) = -f_Z[(1 - x)^{k_0}; g_n, k_0] \cdot [(1 - x)^{k_0}]'$$

And

$$\hat{f}_X(x; g_n, k_0) = \hat{f}_Z[(1 - x)^{k_0}; g_n, k_0] \cdot k_0 (1 - x)^{k_0 - 1}$$

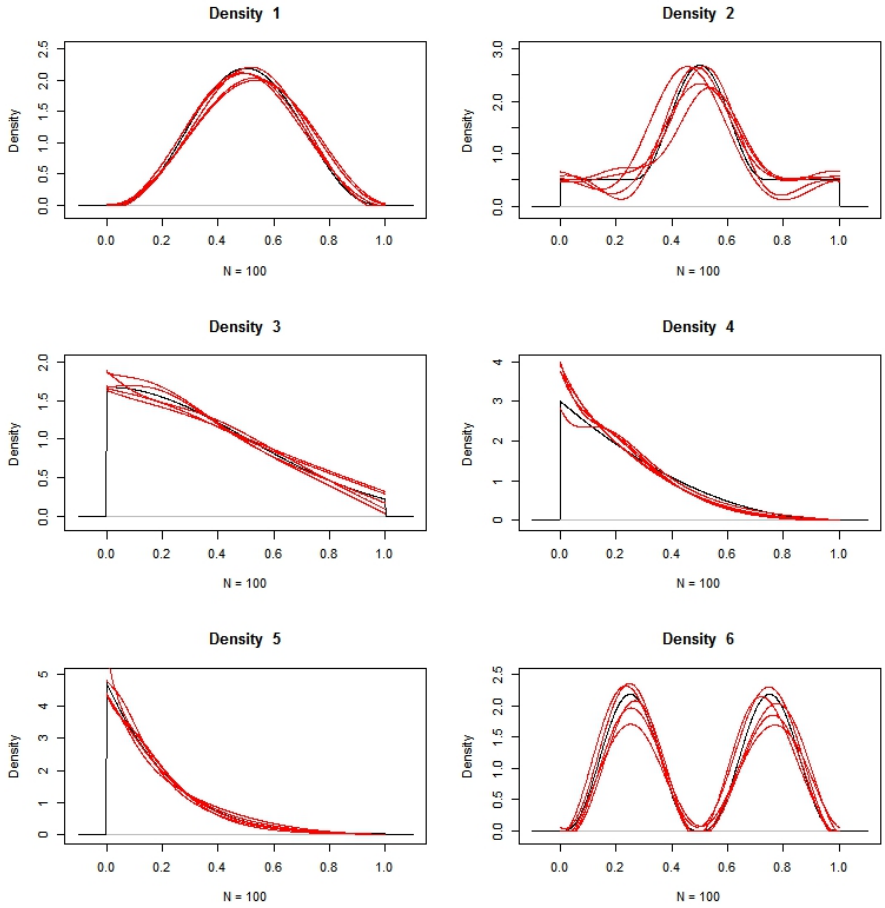


Figure 2.2: Different Algorithms

Remark 2.5.2. The last modified polynomial algorithm is essentially using a higher order polynomial instead of quadratic polynomial as transform function. Assume $f_X(1) > f_X(0) = 0$ and recall (2.5.1), if $k = k_0$, then

$$f_Z(z) = \frac{f_X(z^{1/k_0})}{k_0 z^{(1-k_0)/k_0}}$$

One problem might arise when $z^{(1-k_0)/k_0}$ decay faster than $f_X(z^{1/k_0})$ as $z \to 0$, i.e.

$$\frac{z^{(1-k_0)/k_0}}{f_X(z^{1/k_0})} = o(1)$$

Which implies $f_Z(z) \to \infty$ as $z \to 0$. In our simulation, the estimator still performs good if in this situation, but is not theoretically supported and might not have higher order of bias anymore.

The following plots are some simulation examples of six selected density function, each one of them include five repetitions. More details are introduced in the section of simulation.

2.6 Simulation

In this section, and Figure 2.2, we estimate six different selected densities below. In this table, Beta$(\cdot,\cdot)$ is the Beta density distribution, Beta$_{[a,b]}(\cdot,\cdot)$ denotes the corresponding Beta density function rescaled to interval $[a,b]$. "Truncated" means the original density function is truncated to interval $[0,1]$. We investigate the finite sample performance of original flat-top series estimator and transformed flat-top series estimator using Monte Carlo method.

Table 2.1: 6 Densities in the Simulations

#	Density	Description
1	Beta(4,4)	$f_X(0) = f_X(1) = 0$
2	$\frac{1}{2}$Beta(1,1)+$\frac{1}{2}$Beta$_{[1/4,3/4]}(4,4)$	$f_X(0) = f_X(1) > 0$
3	Truncated $N\left(0, \frac{1}{4}\right)$	$f_X(0) > f_X(1) > 0$
4	Beta(1,3)	$f_X(0) > f_X(1) = 0$
5	Truncated $N\left(-1, \frac{1}{4}\right)$	$f_X(0) > f_X(1) \approx 0$
6	$\frac{1}{2}$Beta$_{[0,1/2]}(4,4) + \frac{1}{2}$Beta$_{[1/2,1]}(4,4)$	Bimodal Distribution

We also compare them to first beta kernel estimator of Chen [44] (denoted by $\hat{f}_{\text{Beta}}$), the Gaussian copula based estimator of Jones and Henderson [45] (denoted by $\hat{f}_{\text{copula}}$), transformed kernel estimator of Wand et. al. [41] (denoted by $\hat{f}_{\text{TKE}}$) and modified transformed kernel estimator of Wen and Wu [42] (denoted by $\hat{f}_{\text{MTKE}}$). All these estimators have similar spirit to the transformed estimator since they all use locally varying kernel functions. For flat-top series estimator (2.2.4) or (2.3.5), the bandwidth is selected by cross-validation ($\hat{f}_{\text{CV}}$) and adaptive bandwidth choice method ($\hat{f}_{\text{abc}}$). For TKE and MTKE we use plug-in bandwidth by Wen and Wu [42]. For beta kernel and Gaussian copula we use rule of

thumb bandwidth by Jones and Henderson [45].

For each distribution, we conduct simulations with sample size $n = 100$ and replicate them 1000 times. Since TKE and MTKE always explode seriously at 0 and 1, we evaluate average mean integrated square errors of estimators on an equally spaced grid on $[0.001, 0.999]$ with an increment 0.001. The simulation results are listed in Table 2.2, and all the values have been multiplied by 1000 to be better readable. For each density, the minimum Average MISE is highlighted in bold font.

The simulation results are very consistent to the theoretical analysis. Among all the estimators, flat-top estimators provide best AMISE for four out of six densities when $n = 100$. All these 4 densities are either satisfied $f_X(0) = f_X(1)$ or $f_X(0) \neq f_X(1) > 0$. In these 4 densities, adaptive bandwidth is powerful same as in the other literature of flat-top based estimator. However, the derivative of density 2 is continuous up to order 2, and limited adaptive bandwidth method as we mentioned previously, also see Politis [51]. Although cross-validation performs stable, it might not be able provide $h = A_1 n^{-1/5}$ and is implemented much slower.

It is not surprising beta kernel estimator provides the best performance in density 4 which is a beta density function. Flat-top and beta kernel give very close AMISE in density 5. The reason that flat-top estimator fails to be the best in density 4 and 5, is that the modified algorithm in section 2.5.2 is not really theoretically supported, and the success of the choice of k_0 introduced in section 2.5.2 might rise some problems.

2.7 Conclusion

A transformed flat-top series estimator is proposed for density function in compact support. This method is developed from boundary correction method, transformed method and flat-top series estimator. We also establish the theoretical properties and show its higher order of bias in the interior region of the support. We use adaptive bandwidth choice and cross-validation as for bandwidth

Table 2.2: Simulation Result with Sample Zize $n = 100$

	Average MISE (Std. Dev.) $\times 10^{-3}$					
Methods	Density 1	Density 2	Density 3	Density4	Density 5	Density 6
$\hat{f}_{\text{Beta}}$	61.92	82.10	23.59	**20.28**	**42.72**	193.07
	(32.30)	(41.89)	(18.77)	**(16.41)**	**(35.80)**	(33.59)
$\hat{f}_{\text{copula}}$	38.05	92.05	34.33	57.19	108.38	140.04
	(25.04)	(45.20)	(22.81)	(34.12)	(74.55)	(31.22)
$\hat{f}_{\text{TKE}}$	42.65	98.38	60.67	85.31	121.21	158.00
	(27.23)	(48.47)	(41.01)	(67.13)	(97.87)	(39.61)
$\hat{f}_{\text{MTKE}}$	30.67	93.42	35.96	46.01	101.29	93.65
	(24.85)	(40.36)	(43.38)	(26.30)	(51.72)	(76.74)
$\hat{f}_{\text{CV}}$	41.25	**74.95**	27.49	51.41	59.40	60.46
	(65.74)	**(70.59)**	(43.58)	(73.75)	(104.77)	(60.53)
$\hat{f}_{\text{abc}}$	**18.68**	124.69	**17.66**	39.27	45.99	**53.83**
	(17.98)	(85.08)	**(21.30)**	(41.33)	(57.76)	**(86.02)**

selection and present their performance in the simulation, and compare them to four other popular estimators for density with compact support. In summary, infinite differentiable flat-top series estimator and transformed flat-top series estimator have the best AMISE among many other methods, for densities with same values or positive values at two boundaries, and not perform good for densities with zero at only one boundary. Development of the proposed estimators on density with a pole or infinite derivative, and extension to multivariate density problem will be pursued in our future work.

2.8 Appendix

Proof of Theorem 2.2.3.

First, denote $\tilde{f}_X$ the periodic extension function of f_X. The Fourier series

coefficient of $\tilde{f}_X$ is

$$\begin{aligned}
|c_n(\tilde{f}_X)| &= \left| \int_0^1 f_X(x) e^{-i2\pi nx} dx \right| \\
&= \left| -\frac{1}{i2\pi n} \left[f_X(1) - f_X(0) - \int_0^1 f_X^{(1)}(x) e^{-i2\pi nx} dx \right] \right| \\
&\leq \frac{1}{2\pi|n|} \left[|f_X(1) - f_X(0)| + ||f_X^{(1)}||_{L^1[0,1]} \right] \\
&= \frac{M_1}{|n|}
\end{aligned} \tag{2.8.1}$$

where $M_1 = \frac{1}{2\pi} \left[|f_X(1) - f_X(0)| + ||f_X^{(1)}||_{L^1[0,1]} \right]$. Because $X_1, \ldots, X_n$ i.i.d. follow density function f_X, and recall the definition of $\lambda_c(s)$, we have

$$\begin{aligned}
\text{Var}\left[\hat{f}_X(x)\right] &= \text{Var}\left[\sum_{s\in\mathbb{Z}} \lambda_c(s) e^{-i2\pi sx} \hat{\phi}_X(s) \right] \\
&\leq \text{Var}\left[\sum_{|s|\leq N_h} e^{-i2\pi sx} \left(\frac{1}{n} \sum_{k=1}^{n} e^{i2\pi sX_k} \right) \right] \\
&= \text{Var}\left[\frac{1}{n} \sum_{k=1}^{n} \sum_{|s|\leq N_h} e^{-i2\pi s(x-X_k)} \right] \\
&= \frac{1}{n} \text{Var}\left[\sum_{|s|\leq N_h} e^{-i2\pi s(x-X_1)} \right] \\
&= \frac{1}{n} \left(\sum_{|s|\leq N_h} \text{Var}\left[e^{-2\pi s(x-X_1)} \right] + \sum_{s,t:s\neq t,|s|,|t|\leq N_h} \text{Cov}\left[e^{-i2\pi s(x-X_1)}, e^{-2\pi t(x-X_1)} \right] \right)
\end{aligned} \tag{2.8.2}$$

Where $N_h = \lfloor 1/h \rfloor$. For the first term in the last line,

$$\begin{aligned}
\text{Var}\left[e^{-i2\pi s(x-X_1)} \right] &= \left| e^{-i2\pi sx} \right|^2 \text{Var}\left[e^{-i2\pi sX_1} \right] \\
&= \text{Var}\left[e^{-i2\pi sX_1} \right] \\
&= \text{E}\left(\left[\left(e^{-i2\pi sX_1} \right) - \text{E}\left(e^{-i2\pi sX_1} \right) \right] \overline{\left[\left(e^{-i2\pi sX_1} \right) - \text{E}\left(e^{-i2\pi sX_1} \right) \right]} \right) \\
&= \text{E}\left(\left[\left(e^{-i2\pi sX_1} \right) - \text{E}\left(e^{-i2\pi sX_1} \right) \right] \left[\left(e^{i2\pi sX_1} \right) - \text{E}\left(e^{i2\pi sX_1} \right) \right] \right) \\
&= \text{E}\left(e^{-i2\pi sX_1} e^{i2\pi sX_1} \right) - \text{E}\left(e^{-i2\pi sX_1} \right) \text{E}\left(e^{i2\pi sX_1} \right) \\
&\leq 1 + |c_s \tilde{f}_X|^2 \\
&\leq 1 + \frac{M_1^2}{s^2}
\end{aligned} \tag{2.8.3}$$

For the second term,

$$\begin{aligned}
\operatorname{Cov}\left[e^{-i2\pi s(x-X_1)}, e^{-2\pi t(x-X_1)}\right] &= \mathrm{E}\left(\left[\left(e^{-i2\pi sX_1}\right) - \mathrm{E}\left(e^{-i2\pi sX_1}\right)\right]\overline{\left[(e^{-i2\pi tX_1}) - \mathrm{E}\left(e^{-i2\pi tX_1}\right)\right]}\right) \\
&= \mathrm{E}\left(\left[\left(e^{-i2\pi sX_1}\right) - \mathrm{E}\left(e^{-i2\pi sX_1}\right)\right]\left[\left(e^{i2\pi tX_1}\right) - \mathrm{E}\left(e^{i2\pi tX_1}\right)\right]\right) \\
&= \mathrm{E}\left(e^{-i2\pi sX_1}e^{i2\pi tX_1}\right) - \mathrm{E}\left(e^{-i2\pi sX_1}\right)\mathrm{E}\left(e^{i2\pi tX_1}\right) \\
&\leq \mathrm{E}\left[e^{-i2\pi(s-t)X_1}\right] + |c_s\tilde{f}_X||c_t\tilde{f}_X| \\
&\leq |c_{s-t}\tilde{f}_X| + \frac{M_1^2}{|st|} \qquad (2.8.4) \\
&\leq \frac{M_1}{|s-t|} + \frac{M_1^2}{|st|}
\end{aligned}$$

Combine them, we have

$$\begin{aligned}
\operatorname{Var}\left[\hat{f}_X(x)\right] &\leq \frac{1}{n}\left[\sum_{|s|\leq N_h}\left(1+\frac{M_1^2}{s^2}\right) + \sum_{s,t:s\neq t,|s|,|t|\leq N_h}\left(\frac{M_1}{|s-t|}+\frac{M_1^2}{|st|}\right)\right] \qquad (2.8.5) \\
&= \frac{1}{n}\left[O(N_h) + \sum_{|s|\leq N_h}\frac{M_1^2}{s^2} + \sum_{s,t:s\neq t,|s|,|t|\leq N_h}\frac{M_1^2}{|st|} + \sum_{s,t:s\neq t,|s|,|t|\leq N_h}\frac{M_1}{|s-t|}\right] \\
&= \frac{1}{n}\left[O(N_h) + M_1^2\left(\sum_{|s|\leq N_h}\frac{1}{|s|}\right)^2 + \sum_{s,t:s\neq t,|s|,|t|\leq N_h}\frac{M_1}{|s-t|}\right] \\
&= \frac{1}{n}\left[O(N_h) + O\left[(\log N_h)^2\right] + O\left(N_h \log N_h\right)\right]
\end{aligned}$$

The last step is derived by integral test of series. Notice all the three terms here are independent from x, thus

$$\sup_{x\in[0,1]}\operatorname{Var}\left[\hat{f}_X(x)\right] = O\left(\frac{N_h\log N_h}{n}\right) = O\left[\frac{1}{nh}\log\left(\frac{1}{h}\right)\right]$$

□

Proof of Corollary 2.2.4.

Recall (2.8.1) in the previous proof, if $|f_X(1) - f_X(0)| = 0$, and $|f_X^{(1)}(1) - f_X^{(1)}(0)|$ is finite,

$$\begin{aligned}
\left|c_n(\tilde{f}_X)\right| &= \left|\int_0^1 f_X(x)e^{-i2\pi nx}dx\right| \\
&= \left|-\frac{1}{i2\pi n}\left[-\int_0^1 f_X^{(1)}(x)e^{-i2\pi nx}dx\right]\right| \\
&= \left|-\frac{1}{(i2\pi n)^2}\left[f_X^{(1)}(1) - f_X^{(1)}(0) - \int_0^1 f_X^{(2)}(x)e^{-i2\pi nx}dx\right]\right| \\
&\leq \frac{1}{(2\pi)^2 n^2}\left[|f_X^{(1)}(1) - f_X^{(1)}(0)| + ||f_X^{(2)}||_{L^1[0,1]}\right]
\end{aligned}$$

$$= \frac{M_2}{n^2}$$

where $M_2 = \frac{1}{(2\pi)^2}\left[|f_X^{(1)}(1) - f_X^{(1)}(0)| + ||f_X^{(2)}||_{L^1[0,1]}\right]$. Instead of (2.8.3) and (2.8.4), we have

$$\text{Var}\left[e^{-i2\pi s(x-X_1)}\right] \leq 1 + |c_s\tilde{f}_X|^2 \leq 1 + \frac{M_2^2}{s^4}$$

and

$$\text{Cov}\left[e^{-i2\pi s(x-X_1)}, e^{-2\pi t(x-X_1)}\right] \leq |c_{s-t}\tilde{f}_X| + |c_s\tilde{f}_X||c_t\tilde{f}_X| \leq \frac{M_2}{(s-t)^2} + \frac{M_2^2}{s^2t^2}$$

Recall (2.8.2),

$$\begin{aligned}
\text{Var}\left[\hat{f}_X(x)\right] &\leq \frac{1}{n}\left[\sum_{|s|\leq N_h}\left(1 + \frac{M_2^2}{s^4}\right) + \sum_{s,t:s\neq t,|s|,|t|\leq N_h}\left(\frac{M_2}{(s-t)^2} + \frac{M_2^2}{s^2t^2}\right)\right] \\
&= \frac{1}{n}\left(O(N_h) + \sum_{|s|\leq N_h}\frac{M_2^2}{s^4} + \sum_{s,t:s\neq t,|s|,|t|\leq N_h}\frac{M_2^2}{s^2t^2} + \sum_{s,t:s\neq t,|s|,|t|\leq N_h}\frac{M_2}{(s-t)^2}\right) \\
&= \frac{1}{n}\left[O(N_h) + M_2^2\left(\sum_{|s|\leq N_h}\frac{1}{s^2}\right)^2 + \sum_{s,t:s\neq t,|s|,|t|\leq N_h}\frac{M_2}{(s-t)^2}\right] \\
&= \frac{1}{n}\left[O(N_h) + O(1) + O(N_h)\right] \\
&= O\left(\frac{1}{nh}\right)
\end{aligned}$$

□

Proof of Theorem 2.2.6.

Because we have

$$\omega(s) = \begin{cases} 1, & |s| \leq 1 \\ 0, & |s| > 1 \end{cases}$$

and $\lambda(s) = \omega(hs)$, it implies

$$\text{E}\left[\hat{f}_X(x)\right] = \sum_{s\leq 1/h}\phi_X(s)e^{i2\pi sx} = S_{N_h}\tilde{f}_X(x)$$

which is partial sum of Fourier series, and $N_h = \lfloor 1/h \rfloor$. Since $\tilde{f}_X(x) \in C^{r-1}$, consider the n$^{\text{th}}$ Fourier coefficient $c_n(\tilde{f}_X)$ (is essentially characteristic function,

by replacing s by $-s$). Use integration by parts,

$$\begin{aligned}
c_n(\tilde{f}_X) &= \int_0^1 f_X(x)e^{-i2\pi nx}dx \\
&= -\frac{1}{i2\pi n}\left[f_X(1) - f_X(0) - \int_0^1 f_X^{(1)}(x)e^{-i2\pi nx}dx\right] \\
&= \frac{1}{i2\pi n}\int_0^1 f_X^{(1)}(x)e^{-i2\pi nx}dx \\
&= -\frac{1}{(i2\pi n)^2}\left[f_X^{(1)}(1) - f_X^{(1)}(0) - \int_0^1 f_X^{(2)}(x)e^{-i2\pi nx}dx\right] \\
&= \frac{1}{(i2\pi n)^2}\int_0^1 f_X^{(2)}(x)e^{-i2\pi nx}dx \\
&\cdots \\
&= -\frac{1}{(i2\pi n)^r}\left[f_X^{(r-1)}(1) - f_X^{(r-1)}(0)\right] + \frac{1}{(i2\pi n)^r}\int_0^1 f_X^{(r)}(x)e^{-i2\pi nx}dx \\
&= -\frac{1}{(i2\pi n)^{r+1}}\left[f_X^{(r)}(1) - f_X^{(r)}(0) - \int_0^1 f_X^{(r+1)}(x)e^{-i2\pi nx}dx\right]
\end{aligned} \tag{2.8.6}$$

If $n \le n' < \infty$, by inequalities of Schwarz and Bessel,

$$\begin{aligned}
&|S_n\tilde{f}_X(x) - S_{n'}\tilde{f}_X(x)| \le \sum_{|k|>n}|c_k(\tilde{f}_X)| \\
&= \sum_{|k|>n}\left|-\frac{1}{(i2\pi k)^{r+1}}\right|\left|f_X^{(r)}(1) - f_X^{(r)}(0) - \int_0^1 f_X^{(r+1)}(x)e^{-i2\pi nx}dx\right| \\
&\le \frac{1}{(2\pi)^{r+1}}\left(\left|f_X^{(r)}(1) - f_X^{(r)}(0)\right|\sum_{|k|>n}\frac{1}{k^{r+1}} + \sum_{|k|>n}\frac{c_k(\tilde{f}_X^{(r+1)})}{k^{r+1}}\right) \\
&\le \frac{1}{(2\pi)^{r+1}}\left(\frac{1}{n^r}|C_r|\cdot\left|f_X^{(r)}(1) - f_X^{(r)}(0)\right| + \left[\sum_{|k|>n}|c_k(\tilde{f}_X^{(r+1)})|^2\right]^{1/2}\cdot\left[\sum_{|k|>n}\frac{1}{k^{2(r+1)}}\right]^{1/2}\right) \\
&\le \frac{1}{(2\pi)^{r+1}}\left(O\left(\frac{1}{n^r}\right) + \frac{C_r'}{n^{r+1/2}}||f_X^{(r+1)}||_{L^2[0,1]}\right) \\
&= O\left(\frac{1}{n^r}\right)
\end{aligned} \tag{2.8.7}$$

C_r and C_r' are constants only depend on r. Note that $O\left(\frac{1}{n^r}\right)$ here is independent from x. Under the conditions of f_X, it is well known that partial sum of Fourier

series will converge to $\tilde{f}_X$ pointwise. Thus, let $n' \to \infty$, we have

$$\sup_{x\in[0,1]} |S_n\tilde{f}_X(x) - \tilde{f}_X(x)| = O\left(\frac{1}{n^r}\right)$$

Recall that $N_h = \lfloor 1/h \rfloor$,

$$\begin{aligned}\sup_{x\in[0,1]} \left|\mathrm{E}\left[\hat{f}_X(x)\right] - f_X(x)\right| &= \sup_{x\in[0,1]} |S_{N_h}\tilde{f}_X(x) - \tilde{f}_X(x)| \\ &= O\left(\frac{1}{N_h^r}\right) \\ &= O\left(h^r\right)\end{aligned}$$

Recall corollary 2.2.4, so for MSE, we still have (2.2.8). □

Proof of Theorem 2.3.3.

By Condition 2.3.1 and Condition 2.3.2, and (2.3.9), it is easy to see $f_{Y_n}(y; g_n)$ is also 2nd continuously differentiable and bounded on $[0, 1]$, and

$$|f_{Y_n}(1; g_n) - f_{Y_n}(0; g_n)| = O(\triangle_n) = o\left(\frac{1}{\log 1/h}\right)$$

Recall (2.8.1),

$$\begin{aligned}|c_n(\tilde{f}_{Y_n})| &= \frac{1}{2\pi|n|}\left[|f_{Y_n}(1; g_n) - f_{Y_n}(0; g_n)| + ||f_{Y_n}^{(1)}||_{L^1[0,1]}\right] \\ &\leq \frac{1}{2\pi|n|}\cdot o\left(\frac{1}{\log 1/h}\right) + \frac{1}{(2\pi)^2 n^2}\left[|f_{Y_n}^{(1)}(1; g_n) - f_{Y_n}^{(1)}(0; g_n)| + ||f_{Y_n}^{(2)}||_{L^1[0,1]}\right] \\ &= o\left(\frac{1}{n\log 1/h}\right)\end{aligned}$$

Recall (2.3.6), similarly let the variance of flat-top series estimator of f_{Y_n} given g_n be:

$$\mathrm{Var}\hat{f}_{Y_n}(y; g_n) = \left[\mathrm{Var}\hat{f}_Y(y; g)\right]\Big|_{g=g_n}$$

Then by exactly same calculation of (2.8.2), (2.8.3), (2.8.4) and (2.8.5), we have:

$$\begin{aligned}\sup_{y\in[0,1]} \mathrm{Var}\left[\hat{f}_{Y_n}(y; g_n)\right] &\leq \frac{1}{n}\left[O(N_h) + o\left(\frac{[\log N_h]^2}{[\log 1/h]^2}\right) + o\left(\frac{N_h \log N_h}{\log 1/h}\right)\right] \\ &= o\left(\frac{N_h}{n}\right)\end{aligned}$$

$$= o\left(\frac{1}{nh}\right)$$

It leads us to

$$\sup_{y\in[0,1]} \left|\hat{f}_{Y_n}(y;g_n) - \mathrm{E}\hat{f}_{Y_n}(y;g_n)\right| = o\left(\frac{1}{\sqrt{nh}}\right)$$

□

Proof of Lemma 2.3.4.

First, for $\phi_n(x)$, $x \in [0,1]$

$$\begin{aligned}
\phi_n(x) &= \int_0^x \phi_n'(t)dt \\
&= \int_0^x \sum_{k=1}^{n} \cos(2k\pi t)dt \\
&= \int_0^x \frac{\sin[(2n+1)\pi t]}{2\sin(\pi t)} - \frac{1}{2}dt \\
&= \int_0^x \left(\frac{\sin[(2n+1)\pi t]}{2\sin(\pi t)} - \frac{\sin[(2n+1)\pi t]}{2\pi t} + \frac{\sin[(2n+1)\pi t]}{2\pi t}\right)dt - \frac{x}{2} \\
&= \int_0^x \sin[(2n+1)\pi t]\left[\frac{1}{2\sin(\pi t)} - \frac{1}{2\pi t}\right]dt + \int_0^x \frac{\sin[(2n+1)\pi t]}{2\pi t}dt - \frac{x}{2}
\end{aligned}$$

Then for $x \in [0,1]$,

$$\begin{aligned}
|\phi_n(x) - \phi(x)| &= \left|\phi_n(x) - \left(\frac{1}{4} - \frac{x}{2}\right)\right| \\
&\leq \left|\int_0^x \sin[(2n+1)\pi t]\left[\frac{1}{2\sin(\pi t)} - \frac{1}{2\pi t}\right]dt\right| + \left|\int_0^x \frac{\sin[(2n+1)\pi t]}{2\pi t}dt - \frac{1}{4}\right|
\end{aligned} \tag{2.8.8}$$

Since it is well known that

$$\left|\int_0^A \frac{\sin(x)}{x}dx - \frac{\pi}{2}\right| \leq \frac{2}{A}$$

Then for the second term of (2.8.8), we have

$$\begin{aligned}
\left|\int_0^x \frac{\sin[(2n+1)\pi t]}{2\pi t}dt - \frac{1}{4}\right| &= \left|\int_0^{(2n+1)\pi x} \frac{\sin(u)}{2\pi u}du - \frac{1}{4}\right| \text{ with } u = (2n+1)\pi t \\
&= \frac{1}{2\pi}\left|\int_0^{(2n+1)\pi x} \frac{\sin(u)}{u}du - \frac{\pi}{2}\right| \\
&\leq \frac{1}{(2n+1)\pi x}
\end{aligned} \tag{2.8.9}$$

For the first term of (2.8.8), let

$$\begin{aligned}\kappa(t) &= \frac{1}{2\sin(\pi t)} - \frac{1}{2\pi t} \\ &= \frac{1}{2}\left[\frac{1}{\sin(u)} - \frac{1}{u}\right] \\ &= \frac{1}{2}\lambda(u)\end{aligned}$$

where $u = \pi t$ runs from 0 to x. Since $\sin(u) < u$ for $u > 0$, $\lambda(u)$ is positive:

$$\lambda(u) = \frac{u - \sin(u)}{u\sin(u)}$$

we find for all $u > 0$ that,

$$\begin{aligned}1 - \frac{1}{2}u^2 &< \cos(u) \\ u - \frac{1}{6}u^3 &< \sin(u) \\ \cos(u) &< 1 - \frac{1}{2}u^2 + \frac{1}{24}u^4\end{aligned}$$

It follows

$$\lim_{u\to 0}\lambda(u) = 0$$

and

$$\begin{aligned}\lambda'(u) &= \frac{\sin^2(u) - u^2\cos(u)}{u^2\sin^2(u)} \\ &> \frac{(u - \frac{1}{6}u^3)^2 - u^2(1 - \frac{1}{2}u^2 + \frac{1}{24}u^4)}{u^2\sin^2(u)} \\ &= \frac{\frac{1}{6}u^4 - \frac{1}{72}u^6}{u^2\sin^2(u)} \\ &> 0\end{aligned}$$

if $0 < u < 2\sqrt{3}$. Since $u = \pi t \le \pi x < \pi < 2\sqrt{3}$, $\lambda'(u) > 0$ for $0 < u \le \pi x$. Therefore, $\lambda(\cdot)$ is strictly increasing in $[0, \pi x]$. Then we have shown that $\kappa(t)$ is increasing in $[0, x]$ and has limit zero at zero. Now apply the second mean value theorem for integrals to the first term of (2.8.8), there is a number ξ such that $0 < \xi < x$, we have:

$$\left|\int_0^x \sin[(2n+1)\pi t]\left[\frac{1}{2\sin(\pi t)} - \frac{1}{2\pi t}\right]dt\right|$$

$$
\begin{aligned}
&= \left| \kappa(0) \int_0^{\xi} sin[(2n+1)\pi t]dt + \kappa(x) \int_{\xi}^{x} \sin[(2n+1)\pi t]dt \right| \\
&= \left| \frac{1}{2} \left(\frac{1}{\sin(\pi x)} - \frac{1}{\pi x} \right) \int_{\xi}^{x} \sin[(2n+1)\pi t]dt \right| \\
&\leq \left| \frac{1}{2} \left(\frac{1}{\sin(\pi x)} - \frac{1}{\pi x} \right) \frac{2}{2n+1} \right|
\end{aligned}
\tag{2.8.10}
$$

Then combine (2.8.8), (2.8.90 and (2.8.10), we obtain:

$$
|\phi_n(x) - \phi(x)| = O\left(\frac{1}{n}\right)
$$

For any $x \in (0,1)$. □

Proof of Lemma 2.3.10.

Under the conditions of $f(x)$, by the same method of (2.8.6), we have:

$$
\begin{aligned}
c_n(\tilde{f}) &= -\frac{1}{(i2\pi n)^2} \left[f^{(1)}(1) - f^{(1)}(0) \right] + \frac{1}{(i2\pi n)^2} \int_0^1 f^{(2)}(x) e^{-i2\pi n x} dx \\
&= \frac{1}{(2\pi n)^2} \left[f^{(1)}(1) - f^{(1)}(0) \right] - \frac{1}{(i2\pi n)^3} \left[f^{(2)}(1) - f^{(2)}(0) \right] \\
&\quad + \frac{1}{(i2\pi n)^3} \int_0^1 f^{(3)}(x) e^{-i2\pi n x} dx
\end{aligned}
$$

Then by exactly the same method of (2.8.7), we have:

$$
\begin{aligned}
|S_n \tilde{f}(x) - S_{n'} \tilde{f}(x)| &\leq \frac{1}{(2\pi)^2} \frac{C_1}{n} |f^{(1)}(1) - f^{(1)}(0)| + \frac{1}{(2\pi)^3} \Big(\frac{C_2}{n^2} |f^{(2)}(1) - f^{(2)}(0)| \\
&\quad + \frac{C_3}{n^{2+1/2}} ||f^{(3)}||_{L^2[0,1]} \Big) \\
&= \frac{K_1}{n} |f^{(1)}(1) - f^{(1)}(0)| + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

Note that K_1 and $O\left(\frac{1}{n^2}\right)$ here are all independent from x. Under the condition of $f(x)$, it is well known that partial sum of Fourier series will converge to $\tilde{f}$ pointwise. Thus let $n' \to \infty$, we have

$$
\sup_{x \in [0,1]} |S_n f(x) - f(x)| \leq \frac{K_1}{n} |f^{(1)}(1) - f^{(1)}(0)| + O\left(\frac{1}{n^2}\right)
$$

□

Proof of Theorem 2.3.11.

Since $|\theta_n - \theta_0| \to 0$ in probability, it is obviously sup $|g_n(x) - g_0(x)| \to 0$ in probability. That implies, for any $\epsilon > 0$, there exists N large enough that for all $n > N$, $[g_N(a), g_N(b)] \subset [g_0(a) - \epsilon, g_0(b) + \epsilon]$ in probability. Let $a' = g_0(a) - \epsilon$, $b' = g_0(b) + \epsilon$, and without loss of generality, let n in the following proof is larger than N.

$$
\begin{aligned}
&\sup_{x\in[a,b]} \left|\hat{f}_X(x;g_n) - f_X(x)\right| \\
&= \sup_{x\in[a,b]} \left|\hat{f}_{Y_n}[g_n(x);g_n]g_n'(x) - f_{Y_n}[g_n(x);g_n]g_n'(x)\right| \\
&\leq \sup_{x\in[a,b]} \Bigg(\left|\hat{f}_{Y_n}[g_n(x);g_n]g_n'(x) - \mathrm{E}\hat{f}_{Y_n}[g_n(x);g_n]\cdot g_n'(x)\right| \\
&+ \left|\mathrm{E}\hat{f}_{Y_n}[g_n(x);g_n]\cdot g_n'(x) - f_{Y_n}[g_n(x);g_n]g_n'(x)\right| \Bigg) \\
&\leq M \sup_{x\in[a,b]} \left(\left|\hat{f}_{Y_n}[g_n(x);g_n] - \mathrm{E}\hat{f}_{Y_n}[g_n(x);g_n]\right| + \left|\mathrm{E}\hat{f}_{Y_n}[g_n(x);g_n] - f_{Y_n}[g_n(x);g_n]\right| \right) \\
&= M \sup_{y\in[g_n(a),g_n(b)]} \left(\left|\hat{f}_{Y_n}(y;g_n) - \mathrm{E}\hat{f}_{Y_n}(y;g_n)\right| + \left|\mathrm{E}\hat{f}_{Y_n}(y;g_n) - f_{Y_n}(y;g_n)\right| \right) \\
&\leq M \sup_{y\in[a',b']} \left(\left|\hat{f}_{Y_n}(y;g_n) - \mathrm{E}\hat{f}_{Y_n}(y;g_n)\right| + \left|\mathrm{E}\hat{f}_{Y_n}(y;g_n) - f_{Y_n}(y;g_n)\right| \right) \quad (2.8.11)
\end{aligned}
$$

For the first term, by theorem 2.3.3, it is $o_p\left(\frac{1}{\sqrt{nh}}\right)$. For the second term, recall (2.3.11), let

$$\tilde{f}^*_{Y_n}(y;g_n) = \tilde{f}_{Y_n}(y;g_n) - 2[f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)]\phi(y)$$

and

$$S_{N_h}\tilde{f}_{Y_n}(y;g_n) = S^*_{N_h}\tilde{f}^*_{Y_n}(y;g_n) + 2\left[f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)\right]\cdot\phi_{N_h}(y)$$

Combine them, by Lemma 2.3.4 and Lemma 2.3.10:

$$
\begin{aligned}
&\sup_{y\in[a'b']} \left|S_{N_h}\tilde{f}_{Y_n}(y;g_n) - \tilde{f}_{Y_n}(y;g_n)\right| \\
&= \sup_{y\in[a'b']} \left|S^*_{N_h}\tilde{f}^*_{Y_n}(y;g_n) + [f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)]\cdot\phi_{N_h}(y) - \tilde{f}^*_{Y_n}(y;g_n) - [f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)]\cdot\phi(y)\right| \\
&\le \left|S^*_{N_h}\tilde{f}^*_{Y_n}(y;g_n) - \tilde{f}^*_{Y_n}(y)\right| + |f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)|\cdot \sup_{y\in[a'b']}|\phi_{N_h}(y) - \phi(y)| \\
&\le \left|\frac{K_1}{N_h}|f^{(1)}_{Y_n}(1;g_n) - f^{(1)}_{Y_n}(0;g_n)| + O\left(\frac{1}{N_h^2}\right)\right| + |f_{Y_n}(1;g_n) - f_{Y_n}(0;g_n)|\cdot O\left(\frac{1}{N_h}\right) \\
&= O_p(h^2)
\end{aligned}
$$

Put the term 1 and term 2 together, and go back to (2.8.11), we obtain:

$$\sup_{x\in[a,b]}\left|\hat{f}_X(x;g_n) - f_X(x)\right| = O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right)$$

□

2.9 Acknowledgements

Chapter 2, is currently being prepared for the following paper prepared for future submission for publication: Bias Reduction by Transformed Flat-top Kernel Estimator of Density on Compact Support. Liang Wang; Dimitris Politis. The dissertation author was the primary investigator and author of this material.

Chapter 3

Bayesian Method for Validation of Continuous Glucose Monitoring System Upgrade

3.1 Introduction

3.1.1 Continuous Glucose Monitoring (CGM)

Continuous Glucose Monitoring (CGM) Systems have had significant impact on the management of Type-1 diabetes over the last few years [69, 70]. CGM devices measure interstitial glucose continuously (providing glucose estimates about every 5 minutes) and display current and past glucose values and their trend to users to assist their diabetes management, adjunctive to their use of blood glucose meters. Currently there are two CGM devices approved for market in the US. Although differences exist in how they are designed, both of these devices use a glucose sensitive sensor inserted in the subcutaneous tissue that produces a current proportional to the interstitial fluid glucose concentration. Proprietary algorithms translate this current signal into glucose values after calibration with blood glucose measurements entered by a user. Accuracy of CGM devices depends on both the sensor probe and the algorithm that translates the sensor signal into glucose values.

CGM devices are classified as Class III medical devices in the USA and require Pre-market Approval (PMA) from the Food and Drug Adminstration (FDA) before they can be marketed under a prescription. Part of this approval process is validating the safety and effectiveness of the device in a clinical study where the device accuracy is compared with reference blood glucose measurements (e.g., from venous blood measured using Yellow Springs Instrument, Yellow Springs, OH) taken during an in-clinic monitoring session. In these studies, accuracy is measured over the period of use of the device (e.g. seven day use for Dexcom Gen 4 PLATINUM devices) and over the range of glucose spanning 40-400 mg/dL. An additional requirement of these validation studies is to evaluate the accuracy of the system at both very low and very high plasma glucose conditions (per ISO 15197). To achieve this, study volunteers are subjected to glucose excursions such that they spend a minimum required amount of time at these low and high glucose states, while their plasma glucose is monitored using BG meters and the YSI instrument (venous blood) in a clinical research center.

CGM accuracy is measured by its closeness to the YSI measurements, i.e. the proportion (or percentage) of CGM measurements that are within 20% of YSI measurements if BG $>$ 100 mg/dL or within 20 mg/dL for BG $\leq$ 100 mg/dL. This performance metric is often referred to as %20/20. YSI measurements are obtained by sampling venous blood once every 15 minutes. For the purpose of associating a YSI glucose value with a sensor glucose reading, the first CGM sensor reading immediately after the reference YSI value and within 5 minutes of obtaining the reference sample is currently matched with the YSI value. These corresponding results between CGM and YSI of matched pairs are then evaluated by %20/20. The accuracy of currently marketed Dexcom Gen 4 PLATINUM CGM system (with Software 505) is reported as %20/20 of 93.8% So far, traditional (Frequentist) statistical experimental design and methods using hypothesis testing were used to demonstrate that these systems have the minimum accuracy required in a clinical study so that they may be marketed in place of previous generation devices. Every time a change is made to the system, a

large clinical study is conducted to fully evaluate the system, each time assuming nothing is known about the system. Given the incremental evolution of the CGM systems over their previous generations, assuming each system as completely unknown is unwarranted and can result in potential delays in providing technologies that can significantly improve quality of life to the CGM users. In this chapter, we are proposing a modified Bayesian method of validation [FDA Guidance] which is able to be assessed by type I& II errors, that can help designing a least burdensome way to complete system validation without impacting effectiveness or safety. In addition we are also proposing Bayesian type I&II errors as alternative method of assessing other classical Bayesian methods of validation.

3.1.2 Bayesian Approach

Even though Bayesian Statistics have been in existence for over 300 years, Bayesian methods have only recently become popular because of improvements in computational techniques and availability of high performance computing facilities. Ashby [71] noted that Bayesian methods would be most used in new and rapidly developing areas, where flexibility and innovation are required.

The key difference between Frequentist statistics and Bayesian statistics is that Frequentist methods assume the parameter of interest is fixed but the value is unknown or uncertain, and sampling methods are used to estimate the parameter; Bayesian statistics considers all parameters as random or uncertain and uses probabilities to describe them informed by prior knowledge about the distribution of the parameter. For example, if one needs to make inferences on a parameter θ, Bayesian statistics starts with a prior probability distribution of θ, defined as $\pi(\theta)$, collects data X that has likelihood function model with parameter θ, denoted by $f(X|\theta)$. Once data are collected, $\pi(\theta)$, the prior probability distribution of θ is updated to a posterior distribution $\pi(\theta|X)$, which may be calculated using Bayes Theorem as:

$$\pi(\theta|X) \propto f(X|\theta)\pi(\theta)$$

Inferences are then made by the posterior distribution $\pi(\theta|X)$. Thus, the primary difference between Bayesian method and Frequentist approach for evaluation of the system performance is in the use of prior knowledge about the system. For the Frequentist approach, nothing is assumed about the system performance and the system performance is evaluated based on whether evidence from new data supports or refutes a hypothesis. On the contrary, the Bayesian approach starts with prior knowledge of the system and then uses observations from new data to modify this prior knowledge. Prior knowledge of the system may be based on previous studies using this system or from in-vitro or pre-clinical studies of the system. Once new observations are made using the system, prior knowledge of the system is updated using the newly generated data to obtain posterior distributions of the parameters in question. All decisions are then made on the posterior knowledge, as it includes both prior knowledge and observed new data.

Because of this reliance on prior knowledge of the system, prior information used takes the center stage in any Bayesian analysis. Thus the first step in conducting Bayesian analysis is establishing the appropriateness of using any existing data as representative of planned future use. This is specifically referred to as **exchangeability** [72]. Selecting how much weight to give prior information also plays an important role in the outcome of the analysis. In order to use Bayesian methods, FDA guidance [72] specifically asks for assessing the risks associated with misusing or inappropriately using prior information. On the other hand, if used appropriately, prior information could lead to smaller clinical studies for system validation.

3.1.3 Motivation and FDA's Guidance

In 2010 FDA released "Guidance for the use of Bayesian Statistics in Medical Device Clinical Trials (February 5, 2010)" to promote the use of Bayesian statistical methods for medical device approvals. This guidance provides a clear process to follow for using Bayesian methods for medical device validation. The guidance establishes the minimum requirements and simulations that have to

be conducted to minimize risk to users. The steps involved in this analysis are as follows:

1. Defining the primary clinical outcome or performance metrics;
2. Identifying and establishing basis for prior information (including exchangeability);
3. Defining hypotheses and decision rules;
4. Defining clinical study, data requirements, and success criteria;
5. Conducting simulations for minimizing Type I errors, including sensitivity analyses on any assumptions made to finalize minimum sample size;
6. Defining the statistical parameters for post-study analysis methods (prior distributions for analysis and decision rules).

Application of Bayesian methods for approval of devices is particularly suitable when improvements are made to devices that are already being used. For example, the FDA recently granted approval of an Investigational Device Exemption (IDE) Supplement to conduct an additional clinical trial utilizing the lead device product for the global wound care market, the dermaPACE device, in the treatment of diabetic foot ulcers(Sanuwave, NCT01824407). In this study, Bayesian analysis was used where previous device performance data was included as prior evidence of device safety and effectiveness.

The use of prior device data to support safety and effectiveness is plausible when the primary device and its interface to the body do not change, which satisfies the exchangeability requirement. By incorporating prior positive information (i.e. the prior distribution believes the treatment is better than control or placebo), the study required fewer subjects than would otherwise be the case while still ensuring adequate statistical power. Other examples of the use of these methods also exist [73, 74].

In this chapter, Bayesian analysis is applied to validate software (algorithm) updates to an existing CGM system. In this case, the sensor used to

measure glucose and its interface with the tissue are unchanged, justifying exchangeability and the use of Bayesian approach. The only change to the system is how the sensor signal is converted to glucose by software. Knowledge of expected changes in software and impact on estimated glucose values constituted the prior information. This was generated using post processing of existing raw sensor signals using the new software. Using Bayesian techniques described in [72, 75] is a small clinical study may then be designed to collect prospective information on the performance of the new software update so that prior information can be combined with new clinical study to form posterior information. Inferences on safety and effectiveness of the new system may then be made using the posterior distributions.

To summarize, in the Bayesian approach one is borrowing strength from existing knowledge of the system, adding to this existing knowledge by conducting a clinical study and then making inferences on the combined pool of data. Because of use of data from past studies, the prospective study sample size is reduced.

Effective Study Size = Borrowed Size + New Clinical Study Size

Because the new clinical study size may be small, one of the risks of Bayesian analysis is too much reliance on prior information, which may mask the true behavior of the system. This is referred to as the prior being too strong, which might result in inaccurate inferences in cases where exchangeability considerations are not straightforward. The FDA specifically asks for detailed analysis on the impact of prior information on analysis outcome. They recommend using Frequentist techniques such as Type I and Type II errors to avoid incorrect conclusions being drawn from Bayesian analysis.

- Type I Error: Type I error is defined as the incorrect rejection of the null hypothesis. For the purpose of this analysis Type I error translates to the probability that the new system is actually no better than the current system, but is incorrectly inferred to be better. This error is evaluated by assuming the true performance of the new system is same as the cur-

rent system and then evaluating the probability of inferring that the new system is better.

- Type II Error: Type II error is defined as the failure to reject null hypothesis. For the purpose of this analysis Type II error translates to the probability that the new system is actually better than the current system, but is incorrectly inferred to be equivalent or worse. This error is evaluated by assuming the true performance of new system is better and then evaluating the probability of inferring that the new system is worse.

In fact, several authors worked on the Bayesian method for sample size determination, see [76, 77, 78, 79, 80, 81, 82]. Unfortunately, none of them evaluate Type I&II error rate. The reason is that the classical Bayesian decision theory will lead to high Frequentist Type I error rate. The difficulty in application of Frequentist Type I&II error analysis to Bayesian methods arises because of the conflict between the strength of the prior information and one of those two errors, and the different perspectives on parameters between Frequentist method and Bayesian method.

Berry *et al.* [83] proposed Indifferent Zone Method which could control Type I error in an ideal level with Bayesian method. However, this method defines six decision zones rather than two decisions in Frequentist's method, and produces many no-decisions in our simulation to avoid the occurrence of Type I error. Gelman *et al.* [84] proposed Type S error. Although Bayesian method has excellent performance along with Type S error, it also has the same problem that makes numerous no-decisions to avoid high error rate.

We present alternative approaches to evaluate type I&II errors: In section 2 the mathematical model based on logit transformation of system update of CGM is introduced. In section 3, a modified Bayesian approach is adapted for Frequentist type I&II errors, that addresses the validation of safety and effectiveness of the system when small changes are made to an already approved system, for which a large knowledge base exists including its in-field use. This method leverages existing knowledge of a CGM system in order to reduce the complexity of the clinical study process, without impacting the assessment of

safety and effectiveness. Section 4 proposes the definition and performance of Bayesian Type I&II errors, that is the assessment of classical Bayesian hypothesis testing method with same perspective of Frequentist type I&II errors. Section 5 presents some discussions and conclusions.

3.2 Model of System Upgrade of CGM

3.2.1 Transformation and Performance Metrics

As noted before, CGM system performance is evaluated using the proportion of the CGM glucose values that are close to time matched YSI glucose values, i.e., the %20/20 metric. The goal of the analysis is to demonstrate statistically that the new system results in better performance than the old system. For the purpose of this analysis, let p_0 be the %20/20 of the current system and p_1 be the %20/20 of the new system. Then the odds of good performance for each system are defined as:

$$\mathrm{O}_0 = \frac{p_0}{1-p_0} \text{ and } \mathrm{O}_1 = \frac{p_1}{1-p_1}$$

Thus we may define Odds Ratio (OR) that provides a relative improvement of the performance of the new system when compared with the current system as:

$$\mathrm{OR} = \frac{\mathrm{O}_1}{\mathrm{O}_0}$$

If the performance of the new system is identical to the old system, we expect the OR to have a value of 1. Because of variability of performance across individuals, OR takes a distribution centered at 1. In practice, it is common to work with log(OR) instead of with OR directly, since log(OR) is approximately normally distributed and allows easier calculation of measures such as confidence intervals and also arriving at the posterior distribution because of the closed-form solutions (i.e., using conjugate distributions). Notice that log(OR) here is essentially a logit transformation.

Recall that we are validating the software algorithm update to an existing system, which means in the experiment, the current and the new systems

are used by the same subjects providing the same sensor signals. In order to demonstrate that the new system is safe and effective by posterior distribution, ideally we expect to see a posterior distribution of the log(OR) is to the right side of zero in a log odds ratio plot, as shown in Figure 3.1, that visually supports new system is better.

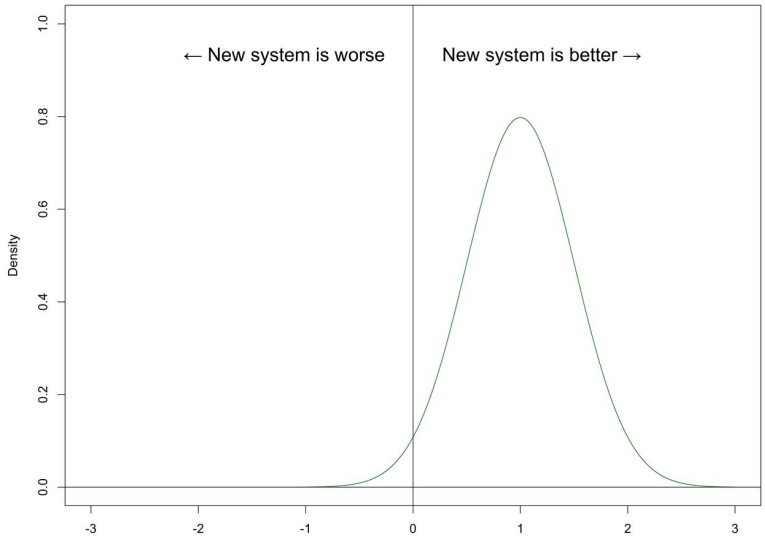


Figure 3.1: Example of Posterior distribution of Log Odds Ration

3.2.2 Data Model

In order to calculate %20/20 of old and new systems (p_0 & p_1), observed data (denoted by X_0 and X_1, that respectively are to corresponding number of %20/20 time matched YSI glucose values) are modeled as a binomial distribution.

$$X_0 \sim \text{Bin}(n, p_0), X_1 \sim \text{Bin}(n, p_1) \tag{3.2.1}$$

where n is the number of matched CGM-YSI pairs. In this exercise, we are comparing two algorithms operating on the same sensor signal. Each subject carry one sensor probe and consequently, X_0 and X_1 are highly correlated because

they are calculated by two different algorithms but with same sensor signal. Let us suppose the correlation of each matched pair from the same subject at the same time is ρ, and denote

$$\hat{p}_0 = \frac{X_0}{n}, \hat{p}_1 = \frac{X_1}{n}, \log(\widehat{\mathrm{OR}}) = \log\left(\frac{\hat{p}_1}{1-\hat{p}_1} / \frac{\hat{p}_0}{1-\hat{p}_0}\right)$$

and $\hat{\rho}$ be the sample correlation. Then by multi-dimensional central limit theorem, we have

$$\log(\widehat{\mathrm{OR}}) | \log(\mathrm{OR}) \sim N(\log(\mathrm{OR}), s^2) \tag{3.2.2}$$

where

$$s^2 = \frac{1}{n\hat{p}_1(1-\hat{p}_1)} + \frac{1}{n\hat{p}_0(1-\hat{p}_0)} + \frac{2\hat{\rho}}{n\sqrt{\hat{p}_0(1-\hat{p}_0)\hat{p}_1(1-\hat{p}_1)}}$$

In the previous section, we explained that in practice the prior of $\log(\mathrm{OR})$ usually is assumed to follow normal distribution. Now suppose the prior of log Odds Ratio is

$$\log(\mathrm{OR}) \sim N(\mu, \sigma^2) \tag{3.2.3}$$

The posterior distribution of log Odds Ratio is

$$\log(\mathrm{OR}) | \log(\widehat{\mathrm{OR}}) \sim N\left(\frac{\frac{\mu}{\sigma^2} + \frac{\log(\widehat{\mathrm{OR}})}{s^2}}{\frac{1}{\sigma^2} + \frac{1}{s^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{s^2}}\right) \tag{3.2.4}$$

In this chapter, for simplicity we only employ a conjugate prior distribution. For other forms of prior distribution, MCMC[1] method could be applied. Before discussing on the definition of decision rule, we need to look into the prior distribution selection first.

3.2.3 Prior Distribution of Log Odds Ratio

One of the primary differences between Bayesian method and Frequentist method statistical analysis is the utilization of prior knowledge. Frequen-

[1] Markov Chain Monte Carlo, is an algorithm to simulate a complicated posterior distribution without conjugate prior distribution.

tist approach always assumes the parameter we are interested is fixed and unknown. On the other hand, Bayesian approach uses a prior distribution to describe the variability of the unknown parameter. Prior distribution could be of two types, informative prior and non-informative prior. Informative prior includes skeptic prior (i.e. we believe equivalence or inferiority of the new system when compared to the old system) and enthusiastic prior (i.e. we believe the new system is superior to the old system). Non-informative prior usually refers to a vague/diffuse prior (for example, $N(0, 10^4)$) and a flat prior (for instance, uniform distribution).

Assumptions made regarding prior distributions have significant impact on the posterior distribution and the analysis outcome. The risk of using a prior distribution that is too informative is that prospective data may be ignored. On the other hand, using too little prior information usually gives us a result similar to the Frequentist approach. While there is no unique way to determine the correct prior, many methods have been developed to assess the impact of prior that help in determining the right distribution (see [72, 83, 85, 86]). FDA Guidance recommends conducting a sensitivity analysis on the prior distribution and using a prior distribution that does not result in the new system meeting the success criteria without a clinical study. That means if making statistical inference based on prior distribution, we will always reject the hypothesis that the new system is better. This requirement will guarantee very low Type I error, and allows for any new clinical data to modify the prior appropriately so that accurate inferences may be made on the posterior distributions.

Sensitivity analysis on the strength of the prior distribution is also very necessary and included in our chapter. The prior distribution could be treated as extra data points and its strength could be explained as the sample size of those extra data points. This borrowed sample size is recommended no larger than some proportion of sample size of new clinical dataset, and can be adjusted by varying the variance or precision of the prior distribution. In this chapter, recall (3.2.4), if we employ non-informative prior distribution $N(0, 10^4)$, and use the posterior distribution of existing data as prior distribution for new clinical

study, then this new prior distribution has variance approximately $\frac{1}{s^2} = \frac{K}{n}$, i.e. be inversely proportional to the borrowed sample size.

3.3 Classical Assessment: Type I&II Errors

3.3.1 Hypothesis and Decision Rule

We are interested in the hypothesis

$$\begin{aligned} &H_0 : \text{New system is same as or worse than old system.} \\ &\qquad\qquad \text{versus} \\ &H_1 : \text{New system is better than the old system.} \end{aligned} \tag{3.3.1}$$

A simple decision rule to test superiority may be

$$H_0 : \log(\text{OR}) \leq 0 \text{ vs. } H_1 : \log(\text{OR}) > 0 \tag{3.3.2}$$

Using classical Bayesian decision theory, a decision rule may be

$$\begin{aligned} &\text{Reject } H_0 \text{ if } \text{P}(H_0|\text{data}) < \text{P}(H_1|\text{data}) \\ &\text{i.e. } \text{P}\left[\log(\text{OR}) \leq 0|\text{data}\right] < \text{P}\left[\log(\text{OR}) > 0|\text{data}\right] \end{aligned} \tag{3.3.3}$$

When H_0 is true, Frequentist approach considers the probability $\text{P}[\text{Reject } H_0|\, \log(\text{OR}) = 0]$ because

$$\begin{aligned} \text{Type I error} &= \text{P}(\text{Reject } H_0|H_0 \text{ is true}) \\ &= \text{P}(\text{Reject } H_0|\, \log(\text{OR}) \leq 0) \\ &\leq \text{P}(\text{Reject } H_0|\, \log(\text{OR}) = 0) \end{aligned} \tag{3.3.4}$$

Thus they can obtain a Type I error less than 5% by making

$$\text{P}[\text{Reject } H_0|\, \log(\text{OR}) = 0] = 5\%$$

Then Type I error of (3.3.3) is

$$\text{P}\Big(\text{P}\left[\log(\text{OR}) \leq 0|\text{data}\right] < \text{P}\left[\log(\text{OR}) > 0|\text{data}\right] \Big| \text{truth is } \log(\text{OR}) = 0\Big)$$

$$\Leftrightarrow \mathrm{P}\Big(\mathrm{E}\left[\log(\mathrm{OR})|\mathrm{data}\right] < 0 \Big| \text{truth is } \log(\mathrm{OR}) = 0\Big)$$

This decision rule, however, may not be appropriate for Frequentist Type I&II error because, when the true log(OR)=0, the mean of posterior distribution, $\mathrm{E}\left[\log(\mathrm{OR})|\mathrm{data}\right]$ is 50% less than 0 and 50% greater than 0. Consequently, this decision rule makes Type I error rate large, converging to 50% as sample size increases. Some other classical Bayesian methods, like bayes factor and posterior p-value, also fail to work with Frequentist type I&II errors as very similar reasons.

A small modification of the decision rule would help to alleviate this issue:

$$\text{Reject } H_0 \text{ if } \pi(0|\mathrm{data}) < \pi(c_0|\mathrm{data})$$

where c_0 represent the value of log(OR) that new system is better. Compared to previous decision rule, we consider posterior density function instead of cumulative distribution function. Since our posterior distribution is still a normal distribution and thus is symmetric, this decision rule is also equivalent to

$$\text{Reject } H_0 \text{ if } \mathrm{E}[\log(\mathrm{OR})|\mathrm{data}] > \frac{c_0}{2}$$

or

$$\text{Reject } H_0 \text{ if } \mathrm{E}[\log(\mathrm{OR})|\log(\widehat{\mathrm{OR}})] > \frac{c_0}{2}$$

The selection of c_0 needs to be very carefully considered. Frequentist method sets a value to be alternative hypothesis for the calculation of power, and that value may be chosen based on experience or prior knowledge of how the new system performs. We recommend selecting the same value for c_0. A small c_0 is essentially a low criterion and might cause risk of high type I error. On the contrary, if c_0 is too large, we might fail to reject H_0 even if almost all the weight of posterior distribution is positive (e.g., when 95% of the posterior distribution is greater than zero but the mean is less than $\frac{c_0}{2}$), that leads to high type II error.

3.3.2 Selection of Prior Distribution

In the simulation study of this chapter, we constructed prior distribution based on a large dataset obtained from Dexcom Gen 4 PLATINUM pivotal study. As noted in section 3.2.2, %20/20 of two algorithms are calculated based on same sensor signal, and the observations therefore follow two binomial distributions and are strongly correlated. This dataset contained two columns of glucose values, each pair of them was observed from exactly same subject at same time, and each column had 13461 matched pairs, and data outside of $[40, 400]$ was excluded in this analysis. Recall p_0 is the %20/20 of old algorithm, p_1 is the %20/20 of new algorithm, then the sample proportion and sample correlation estimated from this study are $p_0 = 0.8156, p_1 = 0.8560, \rho = 0.69$, and the estimate of log(OR) was 0.2956.

With the information above, recall (3.2.3), we employed normal distribution as enthusiastic prior distribution with mean $\mu = \frac{0.2956}{2}$ according to FDA criterion, that would not reject H_0 without extra data, and skeptic prior distribution with mean $\mu = 0$, and reference prior distribution $N(0, 10^4)$. For the variance, we compared the performances when σ^2 are 0.1, 0.2, 0.5, and 1, of skeptic and enthusiastic prior distribution. By previous analysis, $\sigma^2 = 0.1$ is approximately equivalent to 200 extra data, $\sigma^2 = 0.2$ is approximately equivalent to 100 extra data, $\sigma^2 = 0.5$ is approximately equivalent to 40 extra data, and $\sigma^2 = 1$ implies about 20 extra data.

In summary, the skeptic prior distributions were

$$N(0, 0.1), \quad N(0, 0.2), \quad N(0, 0.5), \quad N(0, 1)$$

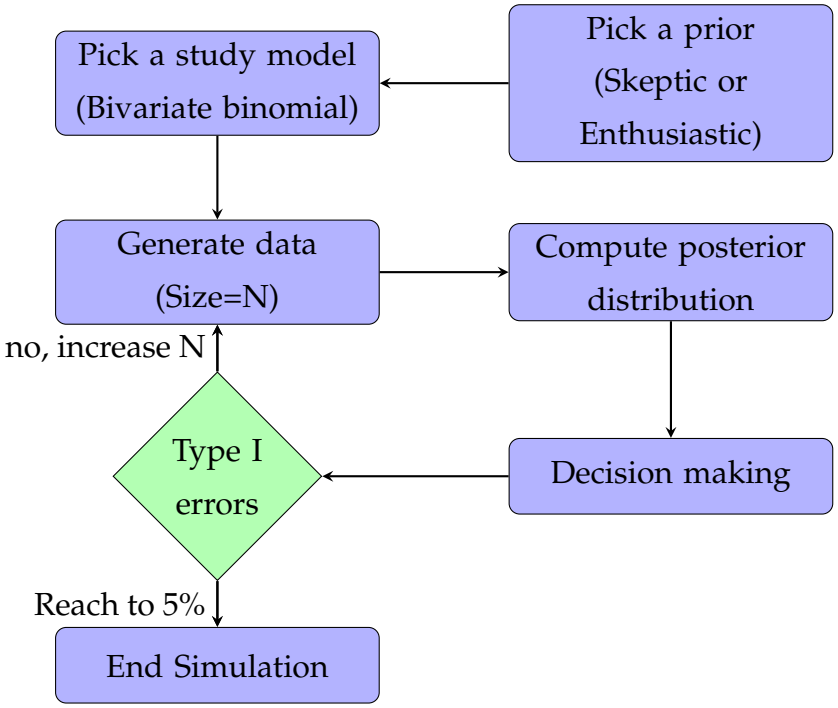
The enthusiastic prior distributions were

$$N\left(\frac{0.2956}{2}, 0.1\right), \quad N\left(\frac{0.2956}{2}, 0.2\right), \quad N\left(\frac{0.2956}{2}, 0.5\right), \quad N\left(\frac{0.2956}{2}, 1\right)$$

And the reference prior distribution was $N(0, 10^4)$.

3.3.3 Study Design

Our simulation is based on the model (3.2.1), and we tried to generate data exactly same as the dataset from Dexcom Gen 4 PLATINUM pivotal study. Thus the parameters in (3.2.1) were $p_0 = 0.8156, p_1 = 0.8560, \rho = 0.69$, and log(OR) was 0.2956. Consequently, we set $c_0 = 0.2956$ as our decision rule. For comparison, we also chose the paired t-test as the Frequentist method which could test the difference of two parameters taking account of high correlation. For the simulation procedure, please see the following flow chart:



In this analysis we assume that adjacent matched pairs are mutually independent and the sample size is up to 1000. In reality, this may not be true since glucose signal has high autocorrelation over short periods, and matched-pairs and %20/20 separated in time by 15-20 minutes are highly correlated [87]. Based on our analysis of CGM data, this correlation becomes insignificant for glucose values separated by more than about 30 minutes. Accordingly, a correction factor of 3 supported by power analysis is used for the final sample size numbers shown in the figures. Consequently, all the sample size in the following plots are already multiplied by this correction factor. Notice the equivalent

sizes of extra data from prior distribution analyzed above are supposed to be multiplied by the same factor.

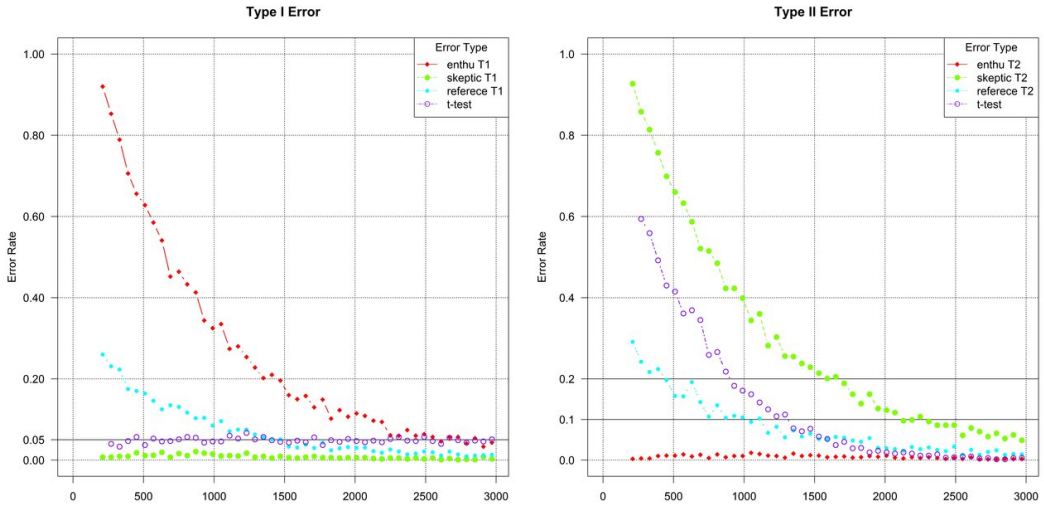


Figure 3.2: Prior $N(0, 0.1), N(0.1405, 0.1), N(0.10^4)$

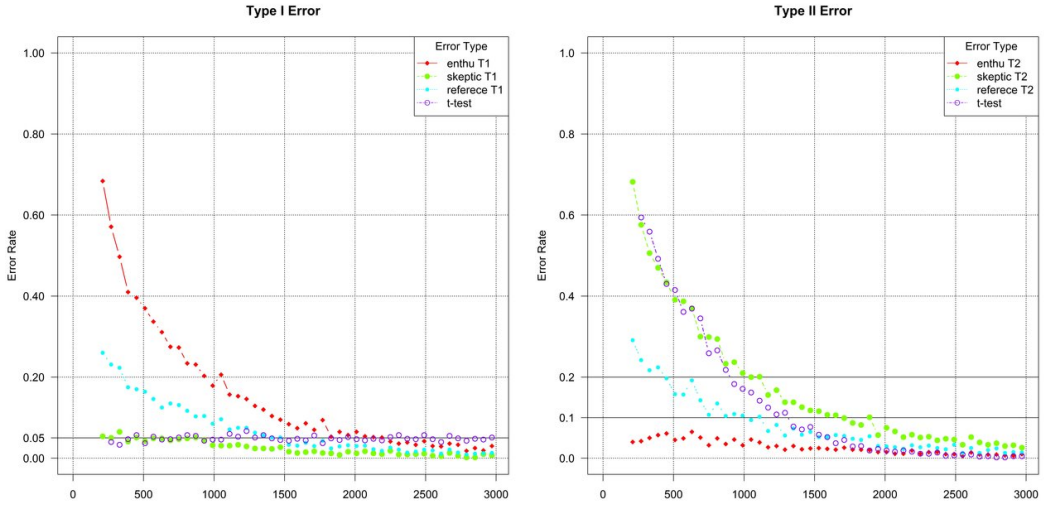


Figure 3.3: Prior $N(0, 0.2), N(0.1405, 0.2), N(0.10^4)$

Figures 3.2-3.5 show the approximate type I&II errors, estimated by Monte Carlo method, and are plotted as the study sample size increases. The decision

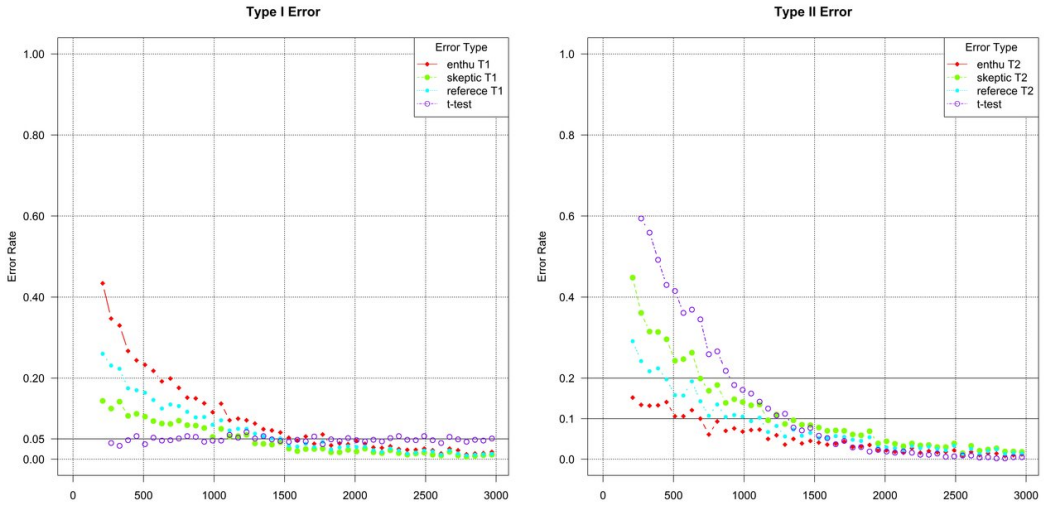


Figure 3.4: Prior $N(0, 0.5), N(0.1405, 0.5), N(0.10^4)$

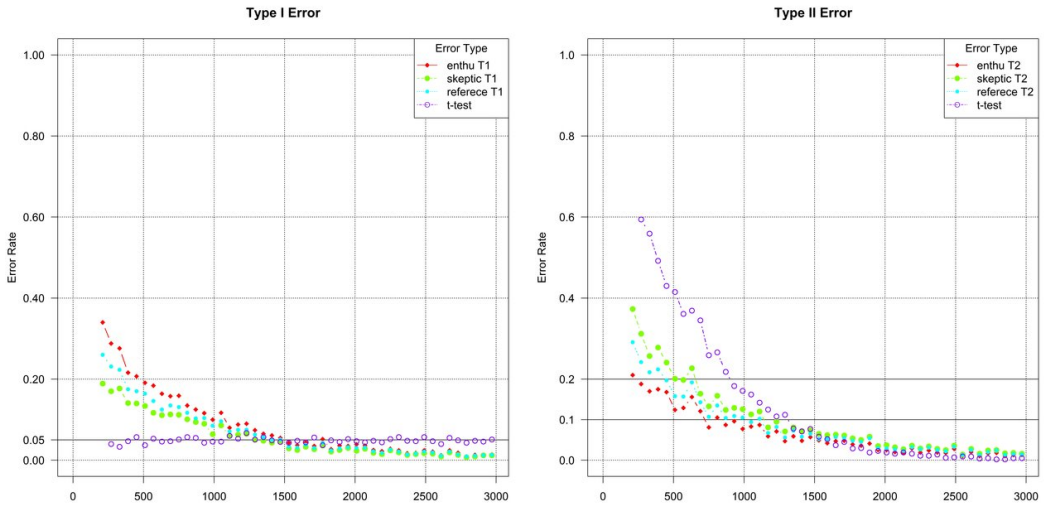


Figure 3.5: Prior $N(0, 1), N(0.1405, 1), N(0.10^4)$

rule of T-test was designed to achieve a fixed Type I error rate and the analysis is conducted to find the minimum sample size that results in pre-determined type II error rate (e.g. 10%). On the other hand, the decision rule of Bayesian method made type I & II errors decease at the same time as the sample size increases, which is also the reason that reference prior method is not visually same as T-test. Thus if n_1 is the minimal sample size for required type I error, and n_2 is the

minimal sample size for desired type II error, we recommend that the minimal sample size should be $n = \max(n_1, n_2)$. Nevertheless, if practitioner are very confident that new system is extremely better and pursue smaller sample size, then they could select skeptic prior distribution, get better type I error and risk higher type II error.

3.3.4 Post Study Analysis

In this section we compared our modified Bayesian method and Frequentist method with real data from Gen 4 PLATINUM pivotal study, where the statistical conclusion is the new algorithm is significantly better than the old algorithm. In this study, for sample size n, we randomly sampled n data points from the whole pool, and mean of enthusiastic prior distribution θ_1 were based on rest of them, the reference prior distribution, skeptic prior distribution and variance of enthusiastic prior distribution were still same as previous section. Sample size n increased from 100 to 3000. The simulation was run 1000 times and the proportion of right decision (i.e. reject H_0) was estimated.

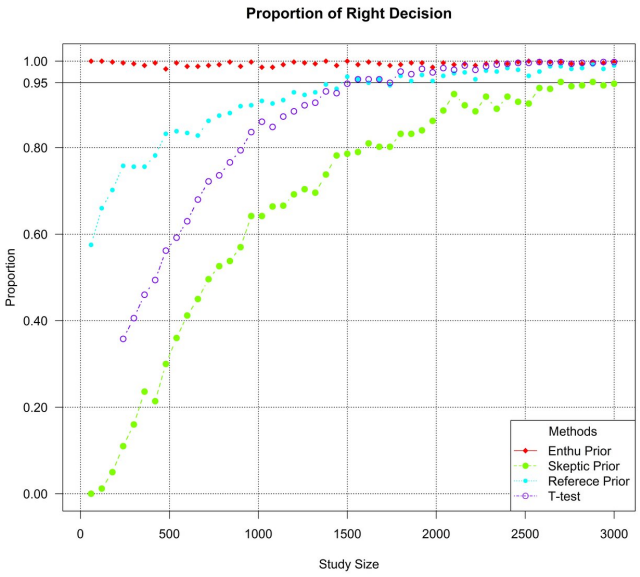


Figure 3.6: Prior $N(0, 0.1), N(\theta_1, 0.1), N(0.10^4)$

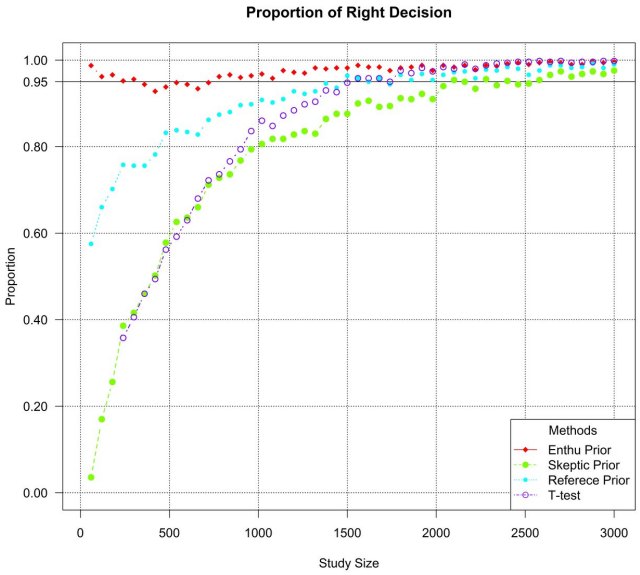


Figure 3.7: Prior $N(0, 0.2), N(\theta_1, 0.2), N(0.10^4)$

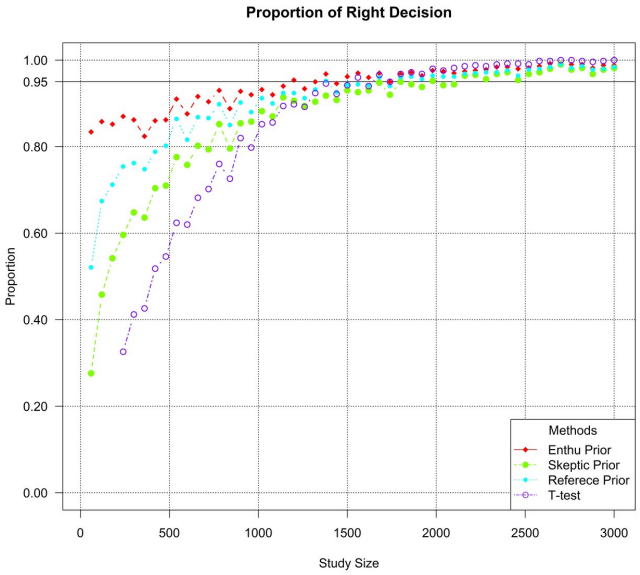


Figure 3.8: Prior $N(0, 0.5), N(\theta_1, 0.5), N(0.10^4)$

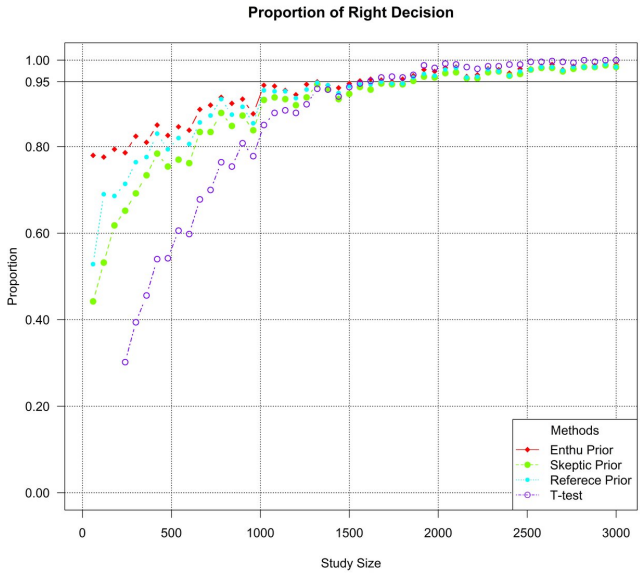


Figure 3.9: Prior $N(0,1), N(\theta_1,1), N(0.10^4)$

3.3.5 Sensitivity Analysis

In the study design, as the sample size increased to very large, all these four methods were consistent even if the prior distributions were informative, that is because prior distribution is down-weighted in front of a large prospective dataset. But different prior distributions still made significant difference when we only had a small sample. We can see skeptic prior distribution gave the best type I error rate, while enthusiastic prior distribution gave outstanding type II error.

From the figure 3.6-3.9, it is clear that the modified Bayesian method with enthusiastic prior distribution always performed better than t-test on making the correct decision. This is reasonable because enthusiastic prior distribution believed the truth. The ability making the right decision of T-test improved with sample size, and was almost same as the Bayesian method with reference prior distribution. But the skeptic prior distribution was not as effective at reducing the sample size as the others, and showed the risk of using wrong prior distribution. Reference prior is a compromise between enthusiastic and skeptic

priors, but still performs better than T-test when sample size is small. Modified Bayesian method could be helpful for reducing sample size significantly, but requires significant amount of pre-study analysis and simulation to thoroughly assess the impact of prior and in choosing the correct decision rules, i.e. choice of c_0. Our recommendation is to use a skeptic prior distribution to pass FDA's criterion with very small sample size if the product is extremely good, or to use a weak enthusiastic prior supported by evidence of how software changes impact data and simulations. These simulation and sensitivity analyses on all critical parameters would significantly mitigate the risk of Type I errors, while keeping the sample size low.

3.4 New Assessment: Bayesian Type I&II Errors

In the previous section, we successfully assessed a modified Bayesian method by Frequentist type I&II errors, but it did not make full use of the posterior distribution and Bayesian framework. As we analyzed in section 3.2.1, Frequentist method calculates type I&II errors only at two fixed points. For the modified Bayesian method to work, we had to only utilize the information of posterior distribution at the same two points. In this section, we propose Bayesian type I&II errors, that are able to assess classical Bayesian method by considering the probability of wrong decision under different hypothesises.

3.4.1 Bayesian Hypothesis Testing

In general, the classical hypothesis testing usually has the form:

$$H_0 : \theta \in \Theta_0; \text{ versus } H_1 : \theta \in \Theta_1$$

Recall (3.3.1) (3.3.2), for simplicity, we consider the problem:

$$H_0 : \theta \leq 0; \text{ versus } H_1 : \theta > 0 \tag{3.4.1}$$

where θ is log(OR). Frequentist method compute type I error under H_0 is true, in fact, recall (3.3.4), they only consider its upper bound $\mathrm{P}(\text{Reject } H_0 | \theta \leq 0)$

because

$$\mathrm{P}(\text{Reject } H_0|\theta \leq 0) \leq \mathrm{P}(\text{Reject } H_0|\theta = 0)$$

Similarly, for type II error or power, Frequentist analysis uses fixed $\theta_1 > 0$ (for example, $\theta_1 = 0.5$ and assume $\theta = \theta_1$) as the alternative hypothesis, instead of the true alternative hypothesis $H_1 : \theta > 0$. Essentially, although the hypothesis testing considers two different areas, Frequentist method only focuses on two fixed points within these regions. Suppose now we keep computing the Frequentist type I error, with the truth $\theta = 0$, but make inference based on a prior distribution and posterior distribution that θ extends the whole real line, then it might result in inaccurate outcomes.

On the contrary, hypothesis testing in the Bayesian paradigm requires specifying distributions on θ under each hypothesis, i.e.

$$\begin{aligned} H_0 : \theta \leq 0, &\quad \theta \text{ follows distribution } p_0(\theta) \\ &\quad \text{versus} \\ H_1 : \theta > 0, &\quad \theta \text{ follows distribution } p_1(\theta) \end{aligned}$$

where

$$p_0(\theta) \begin{cases} \geq 0, & \theta \leq 0 \\ = 0, & \theta > 0 \end{cases}, \quad p_1(\theta) \begin{cases} = 0, & \theta \leq 0 \\ \geq 0, & \theta > 0 \end{cases} \tag{3.4.2}$$

Inference is still made based on posterior distributions. Consider Type I error, suppose we use $\pi(\theta)$ as prior distribution in analysis, $x|\theta$ is the data from the specifying model, θ follows distribution $p_0(\cdot)$, and the posterior distribution is $\pi(\theta|x)$. One classical Bayesian decision rule is using the posterior confidence interval, that we will reject H_0 only when the left sided posterior confidence interval is completely positive, or:

$$\text{Reject } H_0 \text{ when } \int_{-\infty}^{0} \pi(\theta|x)d\theta < 5\% \tag{3.4.3}$$

The Frequentist type I error rate is

$$\mathrm{P}(\text{Reject } H_0|H_0 \text{ is true}) = \mathrm{P}[\text{Reject } H_0|\theta \text{ follows distribution } p_0(\theta)]$$

We will find that type I error rate is not a fixed value but essentially a function of random variable θ. To eliminate the uncertainty of θ, Bayesian method usually integrates θ over its own distribution. With same spirit, we define Bayesian type I error rate:

$$\mathrm{P}(\text{Bayesian Type I error}) = \int \mathrm{P}(\text{Reject } H_0|\theta)p_0(\theta)d\theta \tag{3.4.4}$$

Bayesian type II error has similar definition but integrated by $p_1(\theta)$. This kind of type I error rate is essentially averaging the Frequentist type I error over distribution $p_0(\theta)$. Unlike only focusing on 2 fixed points in Frequentist Type I error, Bayesian type I&II errors could take account of all possible value of θ under each hypothesis.

3.4.2 The True Distributions, $p_0(\cdot)$ and $p_1(\cdot)$

Bayesian method sometimes formulated hypothesis question (3.4.1) as a model selection problem [88], where we compare the model:

$$M_1 : f(x|\theta), p_0(\theta) = k_0\pi(\theta)\mathbb{1}_{\{(-\infty,0]\}}(\theta)$$

and

$$M_2 : f(x|\theta), p_1(\theta) = k_1\pi(\theta)\mathbb{1}_{\{(0,\infty)\}}(\theta)$$

Where $f(x|\theta)$ is the density function of observation X given θ, $\pi(\theta$ is the prior distribution of θ, and k_0, k_1 are normalized constants. In another word, $p_0(\theta)$ and $p_1(\theta_0)$ are the truncated density function of $\pi(\theta)$ under $(\infty, 0]$ and $(0, \infty)$.

$$p_0(\theta) = \pi(\theta|\theta \leq 0) = \frac{\pi(\theta)}{\int_{-\infty}^{0}\pi(\theta)d\theta}$$

and

$$p_1(\theta) = \pi(\theta|\theta > 0) = \frac{\pi(\theta)}{\int_{0}^{\infty}\pi(\theta)d\theta}$$

It implied the true distribution was totally dependent on prior distribution. Therefore, for the selection of prior distribution, It is not acceptable to employ a weak prior distribution or non-informative prior distribution. It could not

be the posterior distribution obtained by the whole dataset from Gen 4 Pivotal study neither, because such prior was too strong and would violate FDA's criterion (product could not pass the testing based on prior distribution without extra data). In fact, such posterior distribution asymptotically converged to a very small positive neighbourhood of the true parameter.

The true distribution and prior distribution need to be carefully analyzed and supported by evidence. In this chapter, our source of prior information was Gen 4 Pivotal study and concluded new algorithm is significantly better, thus skeptic prior distribution is also inappropriate. As a result, our prior distribution was enthusiastic prior distribution, with mean $\mu = 0.2956$, and variance $0.18 \leq \sigma^2 \leq 0.5$. 0.18 was chosen for always failing to reject H_0 without extra data by decision rule (3.4.3), 0.5 was picked to maintain the strength of prior distribution.

3.4.3 Study Design

In this section, we followed a very similar procedure as section 3.3.3, and estimated Bayesian type I&II errors (3.4.4) by Monte Carlo method.

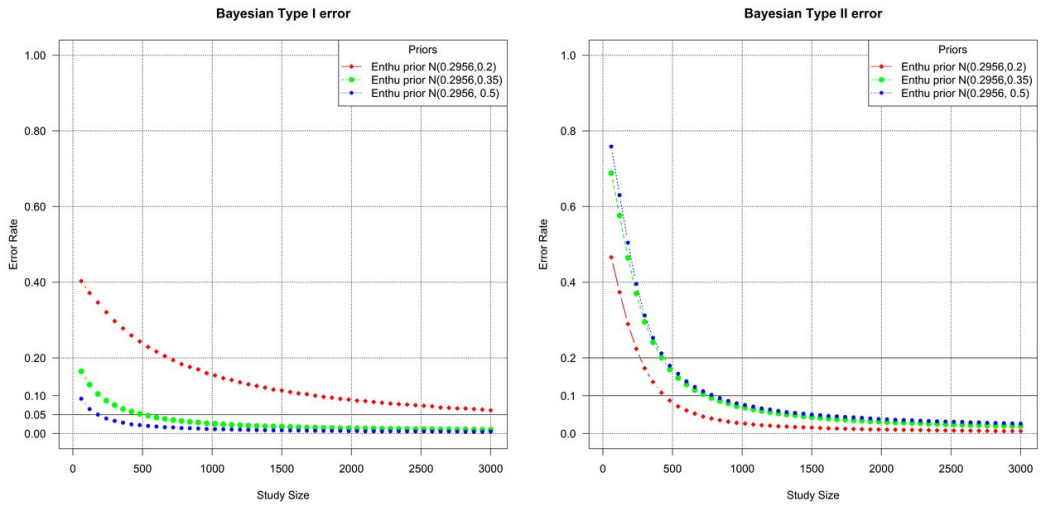


Figure 3.10: Truncated Normal as the true distribution

Take Bayesian type I error as example, first, we select prior distribution $\pi(\theta)$ and

true distribution $p_0(\theta)$, $p_1(\theta)$, then generated 500 samples of θ from distribution $p_0(\theta)$, and for each different θ we generated 500 datasets to estimate Frequentist type I error as in section 3.2.5 with decision rule (3.4.3). Then we average all the Frequentist type I error over those 500 samples of θ to obtain the Bayesian type I error. For the prior distribution, we used enthusiastic prior distributions $N(0.2956, 0.2)$, $N(0.2956, 0.35)$, $N(0.2956, 0.5)$. As noted in the previous section, they truncated distributions were the true distributions corresponding to different situation.

Recall section 3.3.3, T-test with Frequentist type I&II errors had 90% type II error when sample size increased to around 450. However, the Bayesian type I error rate (except enthusiastic prior $N(0.2956, 0.2)$) was almost 0 around 450, and Bayesian type II error was also much lower than 10%, where Frequentist type I error of T-test is still 5%. Note that posterior confidence interval as decision rule performed bad with Frequentist type I&II errors, but was nicely assessed by Bayesian type I&II errors. This new assessment also reflected the advantage of Bayesian method when using a correct prior distribution.

3.4.4 Post Study Analysis

With the exactly same procedure as section 3.3.4, we compared the proportion of right decision of Bayesian method (3.4.3) to paired T-test, see the following plot. Notice this proportion was neither Frequentist type error nor Bayesian type error, and therefore was more objective. It was apparently all three Bayesian methods (3.4.3) had much better performance. They were the best except the modified Bayesian method with enthusiastic prior distribution, that however had high risk of worse Frequentist type I error.

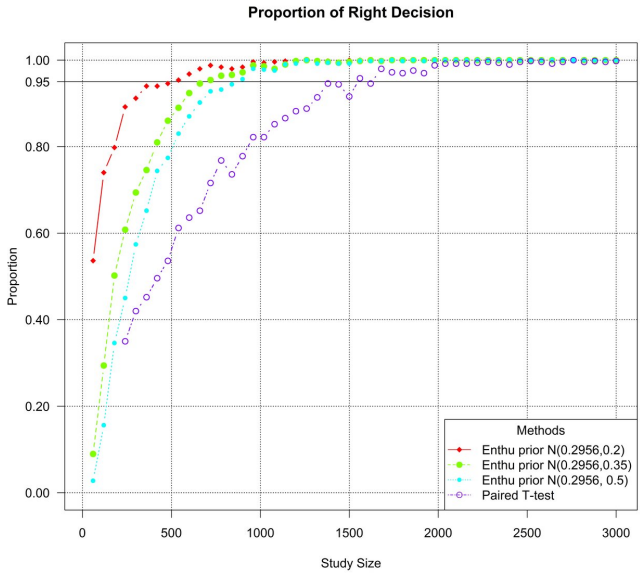


Figure 3.11: Proportion of right decisions

3.4.5 Sensitivity Analysis

Although the enthusiastic prior $N(0.2956, 0.2)$ had excellent Bayesian type II error that slightly better than the other priors, its Bayesian type I error was much worse and not recommended even considering to the benefit from Bayesian type II error. According to the previous analysis, we also did not favor a weak prior distribution. Thus, the strength of prior is neither too strong nor too weak. We recommend determining the acceptable weakest prior distribution first, and the acceptable strongest prior distribution that not violates FDA requirement, then the prior distribution around middle with evidence supported might be a good option.

3.5 Conclusions

Even though Bayesian methods have existed for more than a century, only recently have they been used extensively in the medical fields. Use of these methods for CGM, and the Bayesian type I&II errors of these methods have not

been presented before. The approach presented in this chapter follows many papers on the use of Bayesian methods and FDA guidance on using Bayesian statistics.

As noted in the FDA guidance, exchangeability and choice of prior distribution are two critical aspects of Bayesian analysis. Because strength of prior distribution plays an important role in the outcome of the analysis, FDA guidance recommends using Type I&II errors to evaluate its impact. In this chapter a modified Bayesian method evaluated by Type I & II errors are presented with emphasis on decision rules using Odds ratios. Some Bayesian statisticians [83] suggested using enthusiastic prior in the design phase of the study but non-informative prior distribution in the real application in the analysis phase. Based on the analysis of our modified Bayesian method that was conducted, it was found that choice of prior is essentially a tradeoff between Type I and Type II errors. FDA prefers skeptic prior distribution, which assumes the treatment has no difference from the control, because it essentially produces the lowest Type I error rate. However, using a skeptic prior results in lower power, which means we can not pass the criteria (determined by FDA) unless the new system is significantly superior to the old system. On the other hand, enthusiastic prior is equivalent to adding more data points that assume new system is better, which helps with a smaller study, but with a higher Type I error risk. The analysis also highlighted the importance of the decision rule used for rejecting or accepting the hypothesis. An appropriate choice depends on prior data and knowledge of changes made to the system. This is where exchangeability becomes the key factor. The application presented here, i.e., software update on exist system, is an ideal choice for using Bayesian approach because demonstrating exchangeability is significantly easier. The primary advantage of using this method is reducing the clinical study size so that key improvements are made available to the CGM users with clinical validations that are cheaper and least burdensome, while preserving the safety and effectiveness.

In addition to methods described in the FDA guidance, we also proposed Bayesian type I&II errors as alternative methods of assessing the decision rules.

We believe these methods based on classical Bayesian theory are more suitable for Bayesian analysis than Frequentist Type I and Type II errors, because hypothesis testing uses random parameters as opposed to fixed parameters. More importantly, they are able to assess two class decision problem nicely without producing too many no-decisions? Our future work will focus on further assessing these and other classical Bayesian methods, such as Bayes factor [89] and posterior p-value [90, 91, 92]. More discussion and work on the selection of true distribution $p_0(\theta)$ and $p_1(\theta)$ is also necessary and need to be developed.

3.6 Acknowledgements

Chapter 3, is currently being prepared for the following paper prepared for future submission for publication: Bayesian Method for Validation of Continuous Glucose Monitoring System Upgrade. Liang Wang; Naresh Bhavaraju. The dissertation author was the primary investigator and author of this material.

Bibliography

[1] E.A. Nadaraya. On estimating regression. *Teor. Veroyatnost. i Primenen.*, 9(1):157–159, 1964.

[2] L.P. Devroye. The uniform convergence of the nadaraya-watson regression function estimate. *The Canadian Journal of Statistics*, 6(2):179–191, 1978.

[3] E. Kong, O. Linton, and Y. Xia. Uniform bahadur representation for local polynomial estimates of m-regression and its application to the additive model. *Econometric Theory*, 26(5):1529–1564, 2010.

[4] E. Masry. Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Series Analysis*, 17(6):571–599, 1996.

[5] Q. Li and J.S. Racine. *Nonparametric Econometrics*. Princeton University Press, 2007.

[6] Q. Li and J.S. Racine. Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, 26:423–434, 2008.

[7] Z. Qu and J. Yoon. Nonparametric estimation and inference on conditional quantile processes. *Journal of Econometrics*, 185(1):1–19, 2015.

[8] B. Efron. Bootstrap methods: Another look at the jacknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[9] D.N. Politis. Computer-intensive methods in statistical analysis. *IEEE Singal Proc. Mag.*, pages 39–55, 1998.

[10] W. Hardle and A.W. Bowman. Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *Journal of the American Statistical Association*, 83(401):102–110, 1988.

[11] P. Hall. On bootstrap confidence intervals in nonparametric regression. *The Annals of Statistics*, 20(2):695–711, 1992.

[12] S.G. Shi. Local bootstrap. *Ann. Inst. Statist. Math.*, 43(4):667–676, 1991.

[13] M.B. Priestley and M.T. Chao. Non-parametric function fitting. *Journal of the Royal Statistical Society*, 34(3):385–392, 1972.

[14] Efstathios Paparoditis and Dimitris N. Politis. The local bootstrap for kernel estimators under general dependence conditions. *Annals of the Institute of Statistical Mathematics*, 52(1):139–159, 2000.

[15] A. Dowla, E. Paparoditis, and D.N. Politis. Local block bootstrap inference for trending time series. *Metrika*, 76(6):733–764, 2013.

[16] E. Paparoditis and D.N. Politis. The local bootstrap for periodogram statistics. *Journal of Time Series Analysis*, 20(2):193–222, 1999.

[17] D.N. Politis. Model-free model-fitting and predictive distributions. *TEST*, 22(2):183–221, 2013.

[18] Stefan Sperlich. Comments on: Model-free model-fitting and predictive distributions. *TEST*, 22(2):227–233, 2013.

[19] D.N. Politis. Bootstrap confidence intervals in nonparametric regression without an additive model. In *Topics in NonParametric Statistics: Proceedings of the First Conference of the International Society for NonParametric Statistics*, pages 271–282, New York, USA, 2014. Springer.

[20] Dimitris N. Politis. Model-free prediction with application to functional data analysis. In *Contributions in Infinite-Dimensional Statistics and Related Topics*, pages 221–226, Bologna, 2014. Societa Editrice Esculapio.

[21] Dimitris N. Politis. Model-free vs. model-based volatility prediction. *J. Financial Econometrics*, 5(3):358–389, 2007.

[22] Li Pan and Dimitris N. Politis. Model-free bootstrap for markov processes. In *Proceedings of the 60th World Statistics Congress–ISI2015*, pages 26–31, Rio de Janeiro, Brazil, 2015.

[23] Timothy L. McMurry and Dimitris N. Politis. Nonparametric regression with infinite order flat-top kernels. *Journal of Nonparametric Statistics*, 16(3-4):549–562, 2004.

[24] Dimitris N. Politis. *Model-Free Prediction and Regression: A Transformation-based Approach to Inference*. Springer, New York (to appear).

[25] O. Hossjer and D. Ruppert. Asymptotics for the transformation kernel density estimator. *The Annals of Statistics*, 23(4):1198–1222, 1995.

[26] D. Ruppert and D.B.H. Cline. Bias reduction in kernel density estimation by smoothed empirical transformations. *The Annals of Statistics*, 22(1):185–210, 1994.

[27] D.N. Politis and J.P. Ramano. On subsampling estimators with unkown rate of convergence. *Journal of the American Statistical Association*, 94(446):569–579, 1999.

[28] Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap*. Springer, 1995.

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

[30] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[31] Daren B.H. Cline and Jeffery D. Hart. Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics*, 22(1), 1991.

[32] Theo Gasser and Hans-Georg Muller. Kernel estimation of regression functions. In Th. Gasser and Rosenblatt M., editors, *Smoothing Techniques for Curve Estimation, lecture Notes in Mathematics*, pages 23–68. Springer Berlin Heidelberg, 1979.

[33] T. Gasser, H-G Muller, and V. Mammitzsch. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society. Series B*, 47(2):238–252, 1985.

[34] M.C. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

[35] Hans-Georg Muller. Smooth optimum kernel estimators near endpoints. *Biometrika*, 78(3):521–530, 1991.

[36] Shunpu Zhang and Rohana J. Karunamuni. Smooth optimum kernel estimators near endpoints. *Journal of Nonparametric Statistics*, 12(2):197–221, 2000.

[37] Ming-Yen Cheng, Jianqing Fan, and J.S. Marron. On authomatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708, 1997.

[38] Ming-Yen Cheng. Boundary aware estimators of integrated density derivative products. *Journal of the Royal Statistical Society, Series B*, 59(1):191–203, 1997.

[39] Shunpu Zhang and Rohana J. Karunamuni. On kernel density estimation near endpoints. *Journal of Statistical Planning and Inference*, 70(2):301–316, 1998.

[40] Ann Cowling and Peter Hall. On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society. Series B*, 58(3):551–563, 1996.

[41] M.P. Wand, Marron J.S., and D. Ruppert. Transformations in density estimation. *Journal of the American Statistical Association*, 86(414):343–353, 1991.

[42] Kuangyu Wen and Ximing Wu. An improved transformation-based kernel estimator of densities on the unit interval. *to appear in Journal of the American Statistical Association*, 2014.

[43] Peter Hall and Byeong U. Park. New methods for bias correction at endpoints and boundaries. *The Annals of Statistics*, 30(5):1460–1479, 2002.

[44] Songxi Chen. Beta kernel estimators for density function. *Computational Statistics and Data Analysis*, 31(2):131–145, 1999.

[45] M.C. Jones and D.A. Henderson. Kernel-type density estimation on the unit interval. *Biometrika*, 94(4):977–984, 2007.

[46] S. Zhang, R.J. Karunamuni, and Jones M.C. An improved estimator of the density function at the boundary. *Journal of the American Statistical Association*, 94(448):1231–1241, 1999.

[47] J.S. Marron and D. Ruppert. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B*, 56(4):653–671, 1994.

[48] Dimitris N. Politis and Joseph P. Romano. On flat-top kernel spectral density estimators for homogeneous random fields. *Journal of Statistical Planning and Inference*, 51(1):41–53, 1996.

[49] Dimitris N. Politis and Joseph P. Romano. Multivariate density estimation with general flat-top kernels of infinite order. *Journal of Multivariate Analysis*, 68(1):1–25, 1999.

[50] Dimitris N. Politis. On nonparametric function estimation with infinite-order flat-top kernels. In Ch. A. Charalambides, Markos V. Koutras, and N. Balakrishnan, editors, *Probability and Statistical Models with Applications*, chapter 30. Chapman and Hall, 2000.

[51] N. Politis, Dimitris. Adaptive bandwidth choice. *Journal of Nonparametric Statistics*, 15(4-5):517–533, 2003.

[52] Arthur Berg and Dimitris N. Politis. Higher-order polyspectral estimation with flat-top lag-windows. *Annals of the Institute of Statistical Mathematics*, 61(2):477–498, 2009.

[53] Dimitris N. Politis, Vasiliev V.A., and P.F. Tarassenko. Adaptive estimation of density function derivative. In *Proceedings of the Third International Workshop on Applied Methods of Statistical Analysis—Nonparametric Approach*, Novosibirsk, Russia, 2015.

[54] Bruce E. Hansen. Exact mean integrated squared error of higher order kernel estimators. *Econometric Theory*, 21(6):1031–1057, 2005.

[55] Ivan A. Canay. Simultaneous selection and weighting of moments in gmm using a trapezoidal kernel. *Journal of Econometrics*, 156(2):284–303, 2010.

[56] Emanuel Parzen. On estimation of a probability density function and its mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[57] M.S. Bartlett. Statistical estimation of density functions. *The Indian Journal of Statistics, Series A*, 25(3):245–254, 1963.

[58] Theophilos Cacoullos. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18(1):179–189, 1966.

[59] Luc Devroye. *A Course in Density Estimation*. Birkhauser Boston.

[60] Boris L. Granovsky and Muller. Optimal kernel methods: A unifying variational principle. *International Statistical Review*, 59(3):373–388, 1991.

[61] M.C. Jones. On higher order kernels. *Journal of Nonparametric Statistics*, 5(2):215–221, 1995.

[62] M.C. Jones and P.J. Foster. Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics*, 3(3):81–94, 1993.

[63] J.S. Marron. Visual understanding of higher-order kernels. *Journal of Computational and Graphical Statistics*, 3(4):447–458, 1994.

[64] J.S. Marron and M.P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736, 1992.

[65] Elizbar A. Nadaraya. *Nonparametric Estimation of Probability Densities and Regression Curves*.

[66] Edwin Hewitt and Robert E. Hewitt. The gibbs-wilbraham phenomenon: An episode in fourier analysis. *Archive for History of Exact Sciences*, 21(2):129–160, 1979.

[67] David Gottlieb and Chi-wang Shu. On the gibbs phenomenon and its resolution. *SIAM Rev.*, 39(4):644–668, 1997.

[68] E. Paparoditis and D.N. Politis. On the behavior of nonparametric density and spectral density estima tors at zero points of their support. *to appear in Journal of Time Series Analysis*, 2015.

[69] Roy W. Beck. The effect of continuous glucose monitoring in well-controlled type 1 diabetes. *Diabetes Care*, 32(8):1378–1383, 2009.

[70] Arturo Garcia, Anna Leigh Rack-Gomer, Naresh C. Bhavaraju, Haripriyan Hampapuram, Apurv Kamath, Thomas Peyser, Andrea Facchinetti, Chiara Zecchin, Giovanni Sparacino, and Claudio Cobelli. Dexcom g4ap: An advanced continuous glucose monitor for the artificial pancreas. *J Diabetes Sci Technol*, 7(6):1436–1445, 2013.

[71] Deborah Ashby. Bayesian statistics in medicine: A 25 year review. *Statist. Med.*, 25:3589–3631, 2006.

[72] Food and Drug Administration. Guidance for the use of bayesian statistics in medical device clinical trials. *U.S. Department of Health and Human Services, Food and Drug Administration and Center for Devices and Radiological Health*, 2010.

[73] Gene Pennollo and Laura Thompson. Experience with reviewing bayesian medical device trials. *Journal of Biopharmaceutical Statistics*, 18:81–115, 2008.

[74] Brian P. Hobbs and Carlin Bradley P. Practical bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*, 18:54–80, 2008.

[75] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition, 2004.

[76] Lurdes Y.T. Inoue, Donald A. Berry, and Giovanni Parmigiani. Relationship between bayesian and frequentist sample size determination. *The American Statistician*, 59(1):79–87, 2005.

[77] Lawrence Joseph, David B. Wolfson, and Roxane Du Berger. Sample size calculation for binomial proportions via highest posterior density intervals. *The Statistician*, 44(2):143–154, 1995.

[78] Athanassios Katsis and Blaza Toman. Bayesian sample size calculations for binomial experiments. *Journal of Statistical Planning and Inference*, 81:349–362, 1999.

[79] Dennis V. Lindley. The choice of sample size. *The Statistician*, 46(2):129–138, 1997.

[80] Valeria Sambucini. Sample size determination for inference on the odds ratio. *Journal of the Italian Statistical Society*, 9(1-3):219–243, 2000.

[81] Andrew D Moore and Lawrence Joseph. Sample size considerations for superiority trials in systemic lupus erythematosus (sle). *Lupus*, 8(8):612–619, 1999.

[82] Lawrence Joseph, David B. Wolfson, and Roxane Du Berger. Some comments on bayesian sample size determination. *The Statistician*, 44(2):167–171, 1995.

[83] Scott M. Berry, Bradley P. Carlin, J.Jack Lee, and Peter Muller. *Bayesian Adaptive Methods for Clinical Trials*. CRC Press, 2011.

[84] Andrew Gelman and Francis Tuerlinckx. Type s error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, 2000.

[85] Robert E. Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

[86] Valen E. Johnson and David Rossell. On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B*, 72(2):143–170, 2010.

[87] Troy Bremer and David A. Gough. Is blood glucose predictable from previous values? *Diabetes*, 48(3):445–451, 1999.

[88] Elias Moreno. Objective bayesian methods for one-sided testing. *Test*, 14(1):181–198, 2005.

[89] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[90] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

[91] Andrew Gelman. Understanding posterior p-value. *Electronic Journal of Statistics*.

[92] Gunnhildur H. Steinbakk and Geir O. Storvik. Posterior preditive p-values in bayeisan hierarchical models. *Scandinavian Journal of Statistics*, 36(2):320–336, 2009.

[93] P. Hall and J. Horowitz. A simple bootstrap method for constructing nonparametric confidence bands for functions. *Ann. Statist.*, 41(4):1693–2262, 2013.