NONPARAMETRIC ESTIMATION OF AN ARBITRARY NON-SMOOTH FUNCTIONAL BASED ON TESTING A PAIR OF COMPOSITE HYPOTHESES

MUKHWANA MOSES KOLOLI

DOCTOR OF PHILOSOPHY
(APPLIED STATISTICS)

JOMO KENYATTA UNIVERSITY
OF
AGRICULTURE AND TECHNOLOGY

2023

Nonparametric Estimation of an Arbitrary Non-smooth Functional Based on Testing a Pair of Composite Hypotheses

Mukhwana Moses Kololi

A Thesis Submitted in Partial Fulfillment of the Degree of Doctor of Philosophy in Applied Statistics of Jomo Kenyatta University of Agriculture and Technology

2023

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University

Signature..Date ....

Mukhwana Moses Kololi

This thesis has been submitted for examination with our approval as University Supervisors

Signature...Date.........................

Prof. Orwa O. George
JKUAT, Kenya

Signature...Date........................

Dr. Mung'atu K. Joseph
JKUAT, Kenya

Signature... Date........................

Prof. Romanus O. Odhiambo
JKUAT, Kenya

DEDICATION

To my family; my daughter Abigael, sons Wayne and Adrian, and wife Colleta.

ACKNOWLEDGMENTS

The development of this thesis has been a lengthy research process that involved hard work, valuable insights, dedication and contribution from various specialists in the field of Statistics. I am grateful to the following people and institutions for the academic assistance and support towards the success of this thesis.

First and foremost, I am grateful to Almighty God who through His infinite mercy and love gave me life and guided me throughout the writing of this thesis. My special gratitude goes to my supervisors Prof. George O. Orwa, Dr. Joseph K. Munga'tu and Prof. Romanus O. Odhiambo whose continuous support, guidance and assistance made it possible to complete this thesis. You directly and indirectly gave ideas which were incorporated towards the development of this thesis. I was deeply touched by your invaluable input, continued stewardship and scholarly contributions throughout the period of this study. God bless you so much.

I am also grateful to the entire staff of the Department of Statistics and Actuarial Sciences of JKUAT for both material and moral support during my course of study.

I also extend my appreciation to the entire members of Mathematics department of Kibabii university for giving me an ample time. My special thanks goes to my family members for their understanding and encouragement in the course of this study. May God bless you all.

TABLE OF CONTENTS

DECLARATION ii

DEDICATION iii

ACKNOWLEDGMENTS iv

TABLE OF CONTENTS v

LIST OF TABLES viii

LIST OF FIGURES ix

LIST OF APPENDICES x

ABBREVIATIONS AND SYMBOLS xi

ABSTRACT xii

CHAPTER ONE **1**

INTRODUCTION **1**

1.1 Background of the study 1

1.2 Statement of the problem 5

1.3 Objectives 6

1.3.1 General objective 6

1.3.2 Specific objectives 6

1.4 Justification 7

1.5 Scope of the study 7

1.5 Scope of the study 7

CHAPTER TWO **9**

LITERATURE REVIEW **9**

2.1 Introduction 9

2.2 Nonparametric Estimation of Non-smooth Functionals 9

2.3 MiniMax Lower Bound Techniques 12

2.4 Saddle-point Approximation 17

CHAPTER THREE **20**

METHODOLOGY **20**

3.1 Introduction 20

3.2 Nonparametric Estimation of Density Functions 21

3.2.1 Existing Nonparametric Estimators 23

3.3 Polynomial Approximation 25

3.3.1 Orthogonal Polynomials 27

3.3.2 Weierstrass Approximation Theorem 31

3.4 The Variance and the Bias 35

3.5 Test statistics of the Estimator in Hilbert Space 39

CHAPTER FOUR **41**

RESULTS AND DISCUSSION **41**

4.1 Introduction 41

4.2 The General MiniMax Lower Bound 41

4.3 The MiniMax Lower Bound for estimating the non-smooth functional 44

4.4 The Developed Estimator 47

4.5 Asymptotic Properties of the Developed Estimator 50

4.5.1 The Variance and Bias 50

CHAPTER FIVE **53**

EMPIRICAL STUDY 53

5.1 Introduction 53

5.2 The Mean functions simulated 53

CHAPTER SIX 57

CONCLUSIONS AND RECOMMENDATIONS 57

6.1 Introduction 57

6.2 Summary of Main Results 57

6.3 Conclusions 58

6.4 Recommendations 60

REFERENCES 63

APPENDICES 70

LIST OF TABLES

Table 5.1: Mean functions simulated 52

Table 5.2: Summary results for the bias 54

Table 5.3: Summary results for the mean squared errors 54

Table 5.4: Summary results for the confidence interval lengths 55

LIST OF FIGURES

Figure 3.1: The Boundary bias of Nadaraya-Watson estimator 24

Figure 3.2: Chebyshev Polynomials of first order 29

Figure 3.3: The weight function of orthogonal Chebyshev polynomial, $(1-x^2)^{-\frac{1}{2}}$ ... 30

Figure 3.4(a),(b)(c)(d): Plots of $f(x)$ and $T_n, n = 1, 4, 9, 16$ 32

LIST OF APPENDICES

Appendix I: Polynomial Approximation

Appendix II: Matlab-codes and R-codes for various Graphs

ABBREVIATIONS AND SYMBOLS

F	Probability distributions
$\mathbb{C}$	a set of complex numbers
$\mathbb{R}$	a set of real numbers
$\mathbb{R}^n$	vector space of n-tuples of real numbers
$\mathbb{C}^n$	vector space of n-tuples of complex numbers
ℓ^2	space of square summable sequences
ℓ^p	space of sequences summable with p-th power
MLB	MiniMax lower bound
MUB	MiniMax upper bound
MSE	Mean square error
NP	Nonparametric
i.i.d	Independently and identically distributed
$L < \ldots >$	Inner product space
L_1	Absolute error loss
L_2	Squared error loss
L_p	L_p loss
$\|\|.\|\|$	Norm (analytic distance)
L^2	Minimum norm
L^∞	Maximum norm
$\mathbb{E}[L(.,.)]$	Expected loss

ABSTRACT

In this research, an overall MiniMax lower bound (MLB) was derived for the development of the MiniMax Risk for estimating an arbitrary non-smooth functional, $\frac{1}{n}\sum_{i=1}^{n}|\lambda_i|$ from an observation $Y \sim N(\lambda, I_n)$ based on testing a pair of composite hypotheses. The Minimax lower bounds and upper bounds are used to quantify the fundamental limits and provide benchmarks for evaluating the performance of any estimator in statistical inference. In nonparametric estimation of statistical functionals, both the lower and upper bounds are constructed. In particular when working in the context of MiniMax estimation, the lower bounds are the most important. The problem of estimating non-smooth functionals shows some properties that are different from those that arise in estimating standard smooth functionals. For these reasons the standard methods fail to give sharp results when used to estimate non-smooth functionals. A pair of priors with a large difference in the expected values of the functional were constructed while making the Chi-square distance between two normal mixtures small. The estimator was developed using the best polynomial approximation, Hermite polynomials and Saddlepoint approximation, and it's asymptotic properties: bias, variance were derived. The developed estimator was compared with the Nadaraya-Watson and the Modified Nadaraya-Watson estimators. The MSE, biases and confidence interval lengths of the estimators were computed using simulated data. Smaller values of MSE and biases were obtained for the developed estimator. Short confidence interval lengths were also obtained for the developed estimator. The results developed in this research can also be used to solve excess mass.

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

In statistical inference, one of the fundamental problems is estimating statistical functionals. A functional denoted by $F(f, g, h, ...)$ is a mathematical relation that maps one (or more) function to a constant. Functionals just like functions, reach extremum values when their derivative is zero. The functionals are estimated by either parametric or nonparametric procedure. The parametric procedure involves making assumptions about the underlying distribution. The method of moments and method of maximum likelihood are some of the known examples of parametric methods of estimation (DiNardo and Tobias, 2001) and (Casella and Berger, 2002).

In this research, an overall MiniMax lower bound (MLB) was derived for the development of the MiniMax Risk for estimating an arbitrary non-smooth functional. The MiniMax lower bounds and MiniMax upper bounds play a key role in applied statistical inference (Goldenshluger and Lepski, 2020). They are used in fields of science, engineering and geosciences. For instance, they quantify the lower and upper bounds of estimation and testing, data compression and L_1 distance. They are also used to measure difficulties involved in the corresponding task and provide benchmarks for constructive algorithms (Cai and Low, 2011). Knowing their functional form allows further statistical quantities, such as estimation and construction of stochastic models for various applications to be handled.

Statistical researchers have put in a lot of effort to derive MiniMax lower bounds, upper bounds and optimal rate of convergence in the development of statistical inefficiencies. Although the estimation of smooth functionals is the most developed in literature, the non-smooth functionals estimation remain elusive despite their usefulness in real life application (Cai and Low, 2011) and (Goldenshluger and Lepski, 2020).

Smooth functionals are understood to be differentiable on the corresponding norm; the analytic distance induced by an inner product. For instance, estimating linear functionals is well developed and their optimal rate of convergence is derived based on testing a simple null hypothesis against a simple alternative hypothesis. See for example, (Le Cam, 1973), (Donoho and Liu, 1991) and (Juditsky and Nemirovski, 2020).

The situation is different for nonlinear functionals: the rate-optimal estimators are available only for particular functional classes in the problem of estimating quadratic functionals (Bickel and Ritov, 1988) and (Birǵe and Massart 1995). Bickel and Ritov, (1988) developed the theory of estimating quadratic functionals. They derived the optimal rate of convergence based on testing a simple null hypothesis against a composite alternative hypothesis in a large parameter space.

In the recent past, estimation of non-smooth functionals in the nonparametric set-up has gained popularity (Rockafella, 1994), (Lepski et al., 1999) and (Comminges and Dalalyan, 2013). For instance, progressive clustering, data ranking, irregularity detection, mapping environmental pollution, examining contour levels, excess mass and micro array analysis of genes. These functionals have different rates of convergence from the usual parametric rates that occur in the standard smooth functionals. Thus, their estimation involve techniques which are different from the ones used to estimate smooth functionals.

The nonparametric statistical procedure has several attractive properties. First, it does not require specific assumptions to be made about the data. Second, the information used generally relates to some functions of the actual magnitudes of the random variables in the sample. The field of nonparametric has widened its appeal with a number of new tools for statistical analysis. Some of the tools include: Nonparametric density estimation, Nonparametric regression and Gaussian white noise model. See for example (Tsybakov, 2009).

In nonparametric density estimation, random variables $Y_1, \ldots, Y_n$ are identically distributed from a continuous probability density function f, with respect to the Lebesgue measure on $\mathbb{R}$ where

$$f(y) \geq 0, \qquad \int\limits_{-\infty}^{\infty} f(y)dy = 1 \tag{1.1}$$

The Lebesgue measure is central on the interval $[0, 1]$ and on the real line in probability and also in statistical analysis (Billingsley, 1995). The density function f, is assumed to belong to a large family of densities so that it can be represented through an infinite number of parameters. To develop an optimal estimator, the "smoothness" conditions are imposed on f and its derivatives (Tsybakov, 2009).

Nonparametric regression involves n independent random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ such that

$$Y_j = f(X_j) + \epsilon_j, \quad X_j \in [0, 1], \quad j = 1, \ldots, n \tag{1.2}$$

where f is an unknown response function, Y_i is the variable of interest, X_j is the independent variable and ϵ_i is the error term satifying $E(\epsilon_j) = 0$ for all i. The function $f \in \mathcal{P}$. The class $\mathcal{P}$ can be the set of all continuous nonparametric functions on $[0, 1]$ or a set of all the convex functions and many others.

Statistical researchers have an interest in knowing the relationship between independent variable X and a dependent variable Y. For instance, a regression curve can be used to show the relationship and use equation (1.2) to obtain the value of Y after X is observed (Wolfgang, 1996). Special features of this function are, for instance, monotonicity or unimodality, location of zeros or the size of extrema.

The assumptions that are made on equation (1.2) are appropriate for a Gaussian white noise model. The Gaussian white noise model is a statistical model used to simulate the effect of random processes that occur in nature. Literature on a majority of the smoothing techniques that are highly adaptive have been developed for a Gaussian white noise model (Wang et al., 2008). Conversely, when the error, ϵ_j in equation (1.2) has a heavy-tailed distribution, these smoothing techniques are not readily applicable.

The Gaussian white noise is a model that is idealized to provide an estimate to equation (1.2). Lepski et al., (1999) and (Tsybakov, 2009), considered the idealized "signal and white noise" model of observations as follows: the observed data $Y(t)$, $t \in [0, 1]$ was a trajectory of the stochastic differential equation

$$dY(t) = f(t)dt + n^{-1/2}dW(t) \tag{1.3}$$

where f was the unknown function, W was the standard Weiner process on $[0, 1]$, and n was an integer that gave the "volume of observations". Equation (1.3) represented a realistic model with noisy observations of the unknown function at equidistant or random points. The a priori information on the unknown function,

f was that it possessed some smoothness; belonged to a Hölder class $\Sigma(\beta, L)$ where $\beta, L > 0$ were known parameters. The function f is smooth when it is r times differentiable on R^1, where r is a large integer but less than β, and the r^{th} derivative $f^{(r)}$ of f is Hölder continuous with the exponential $\beta - r$ and constant L.

$$|f^{(r)}(t_0) - f^{(r)}(t_1)| \leq |t_0 - t_1|^{\beta - r}, \qquad t_0, t_1 \in R^1 \tag{1.4}$$

In either of the three cases, the asymptotic behavior of the estimators as $n \to \infty$ is quite important. Asymptotic statistical literature deals most with the rates of convergence, particularly for problems involving infinite dimensional parameters. The rates seem to be caused by analytic properties of particular "smoothness" assumptions and other regularity conditions. However, these properties and conditions cannot explain fully the rates as consequences of geometric properties of models. See for example, (Le Cam, 1973).

The best rate of convergence depends on a requirement that an estimator performs well at a sequence of models that lie nearby. According to (Polland, 2005) the rate of convergence refers to uniform convergence of models in small neighborhoods of some specific model of interest. Deriving the best rate of convergence and the MiniMax lower bounds is important when developing the MiniMax theories in nonparametric functional estimation literature.

1.2 Statement of the problem

Several estimators of smooth functionals have been discussed in literature. The estimators are in simpler cases and their optimal rates of convergence are often parametric or algebraic rates (Lepski et al., 1999). However, the smoothness properties and other regularity conditions used to estimate smooth functionals cannot explain the rates as consequences of geometric properties of models. See for example, (Rockafella, 1994) and (Cai and Low, 2011).

The estimation of non-smooth functionals exhibit some properties that are different from those that occur in smooth functionals. Thus, the standard techniques used in estimating smooth functionals cannot give sharp results when applied on the non-smooth functionals. For instance, when a polynomial is used to smoothen the risk at the origin to obtain an optimal estimator, the polynomial factor can be worse in the tail regions, where the density might be negative and unable to be integrated to 1 (Lepski et al., 1999).

1.3 Objectives

1.3.1 General objective

To estimate an arbitrary non-smooth functional based on testing a pair of composite hypotheses in the nonparametric set-up.

1.3.2 Specific objectives

The specific objectives are:

1. To derive a general MiniMax lower bound of estimating an arbitrary functional based on testing a pair of composite hypotheses.

2. To develop the MiniMax Risk of estimating the non-smooth functional.

3. To develop an estimator for a non-smooth functional in nonparametric procedure.

4. To derive the asymptotic properties of the developed estimator: vis, bias and variance.

1.4 Justification

The MiniMax lower bounds and MiniMax upper bounds are constructed in statistical inference for assessing the quality of decision rules and the performance of any estimation method (Cai and low, 2011). These statistical efficiencies of estimators play a key role in advanced statistical analysis. Although estimators and their convergence rates for smooth functionals are well covered in statistical literature, those for non-smooth functionals are elusive even though they are important in application to real life. In this research, an overall MLB was derived for developing the MiniMax Risk to estimate an arbitrary non-smooth functional. The non-smooth functional is estimated in the nonparametric set-up since the standard techniques used to estimate smooth functionals cannot give sharp results (Cai and Low, 2011). Non-smooth functionals also exhibit rates of convergence that were different from those that occur while estimating smooth functionals.

The nonparametric approach is appropriate to a wide range of data which cuts a cross all the measurement scales. The method can be used even on a small sample size, which would require the distributions to be known precisely all together for parametric methods to be used. In addition, the approach can be used in many real life applications such as progressive clustering, data ranking, irregularity detection, mapping environmental pollution, excess mass and micro array analysis of genes.

1.5 Scope of the study

In estimating a statistical functional in the nonparametric set-up, both the lower and upper bounds are required. The lower bounds are the most important when working in the context of MiniMax estimation. In this research, the parameter space Λ will be split into two disjoint subsets Λ_0 and Λ_1 and an overall MLB derived for developing the MiniMax risk. The MiniMax Risk will be based on testing a pair of composite hypotheses, $H_0 : \lambda \in \Lambda_0$ and $H_1 : \lambda \in \Lambda_1$ where H_0 and H_1 are the null and alternative hypotheses respectively. Two priors will be constructed such that they have a large difference in the expected values of the functional $T(\lambda)$ and a small difference while making the Chi-square distance between the two mixture models.

The best polynomial approximation, Hermite polynomials and Saddlepoint approximation will be used to develop the estimatior. The developed estimator will be compared with the Nadaraya-Watson and the Modified Nadaraya-Watson estimators. The MSE, biases and confidence interval lengths of the estimators will be computed using simulated data.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter, literature review was done in areas related to nonparametric estimation of an arbitrary non-smooth functional based on testing a pair of composite hypotheses. The estimator developed rely on the best polynomial approximation, Hermite polynomials and Saddlepoint approximation of the function f, by using the an unbiased estimator. The entire chapter is organized as follows: Nonparametric estimation of non-smooth functionals is stated in section 2.2. The MiniMax lower bound techniques are presented in section 2.3, the saddlepoint approximation is in section 2.4.

2.2 Nonparametric Estimation of Non-smooth Functionals

In nonparametric statistical approach, no assumption on the underlying distribution is made. Predictions that are more robust in the sense that they do not rely on whether or not the underlying distribution is known are made. The minimax nonparametric problems of estimating density functionals have been extensively studied in the statistical literature. The case of linear functionals is particularly well understood.

Korostelev (1990) considered the problem of estimating the L_1 norm $||f||_1 = \int |f(t)|dt$ with the optimal rate of convergence given as $O(n^{-\beta/2\beta+1})$, where β is the order of smoothness of f, such that a plug-in estimator $\int |\hat{f}(t)|dt$ associated with an optimal estimate $\hat{f}$ of f, is optimal in order. Korostelev and Tsybakov (1994) also considered problems of estimating non-smooth functionals in the image model.

Lepski et al., (1999) observed a Hölder continuous function. The problem was considered in a nonparametric estimation of L_r norms of the continuous function. The study of the function was done in the standard asymptotic set-up, when the parameter $n \to \infty$. Their results largely depended on two circumstances: estimating a smooth functional and estimating a singular functional. The rate of convergence in the former case is $n^{-\frac{1}{2}}$, where n is the sample size while the rate of convergence in the latter case corresponded to estimating the function f itself in the corresponding norm. They noted that the rates of convergence differed from each other. The function and the choice of the norm determined the value of a norm.

They also found that estimating the L_r norm was a case in-between the above extreme cases. The optimal rate of convergence of $||f_r||$ was worse than $n^{-1/2}$ but better than the rate of convergence of the NP estimates of f. The results obtained were based on r; when the value of r was even, the rate was $n^{-\beta/(2\beta+1-1/r)}$ and, when r was odd, a logarithmic in n factor could be used to improve the NP rate $n^{-\beta/(2\beta+1)}$.

The ideas of (Lepski et al., 1999) were advanced by (Cai and Low, 2011). They developed a general MiniMax lower bound procedure. They constructed a pair of priors to obtain the lower bound which was based on the difference between $\mathbb{E}(T)$ on each of the priors μ_0, μ_1 and $Var(T)$ under the μ_0. The lower bound was based on the Chi-square distance between a pair of marginal distributions.

The studies by (Lepski et al., 1999) and (Cai and Low, 2011) are closely related to the work in this thesis by the methodology used and the problem of interest. These authors studied the problem of NP estimation of norms of a signal observed in Gaussian white noise. Both the rate and sharp asymptotics for the estimators in the MiniMax set-up were obtained. They frequently used the assumptions of continuity and normality to obtain estimators.

A part from (Cai and Low, 2011), (Keisuke and Jack, 2012) examined the challenges to statistical inference when the problem of interest is a nondifferentiable functional of the underlying distribution. The situation found in applications of lower and upper bounds analysis, models with moment inequality, and optimal dynamic treatment regimes estimation. They related the existence of unbiased and regular estimators to differentiability of the functional. Their results indicated that if the object of interest was not differentiable, then there exist no estimator sequences that were locally asymptotically unbiased. Strong limits were placed on estimators, bias correction methods, and statistical inference procedures, and provided motivation to consider other criteria for evaluating estimators and inference procedures.

Following the recommendation by (Keisuke and Jack, 2012), (Yu-Xiang Wang et al., 2016) proposed a MiniMax framework for adaptive data analysis. In adaptive data analysis, a sequence of queries were made on data and at each step the choice of the query may depend on the results of previous steps. By assuming that the queries were normally distributed, they obtained a sharp MiniMax lower bound on the squared error in the order $O(\frac{\sqrt{k}\sigma^2}{n})$, where k is the number of queries asked and, $\frac{\sigma^2}{n}$ is the ratio of the signal-to-noise for a single query. The lower bound was based on the construction of a least favorable adversary who picks a sequence of queries that are most probably to be affected by over fitting.

In the cases fore mentioned cases, the disparity between the actual value and its estimator was specified by a real valued loss function $L(\theta, \hat{\theta})$ which quantifies the amount by which prediction deviates from the actual values. The two commonly used loss functions are the quadratic error loss, $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ and the absolute error loss function, $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$. The former has a tendency to be dominated by outliers whereas the latter is not differentiable at $a = 0$ (Ramachanran and Chris, 2009). The statistical researchers prefer using a loss function that is globally continuous and differentiable in optimization algorithms.

2.3 MiniMax Lower Bound Techniques

The development of MiniMax theories in the nonparametric functional estimation literature hinges on statistical researchers' efforts to extract the MiniMax lower bounds and optimal rate of convergence. In the literature of statistical inference a variety of lower bound techniques were discussed. Le cam (1973) and, (Donoho and Liu, 1991) for example, derived the optimal rate of convergence by testing a simple null hypothesis against a simple alternative.

Quadratic functionals were estimated by (Bickel and Ritov, 1988). They tested a simple null hypothesis against a composite hypothesis to find the optimal lower bounds for a broad parameter space. Under the white noise model (Lepski et al., 1999) estimated the L_r norm of the drift function. Cai and Low (2011), improved on the ideas of (Lepski et al., 1999) where they considered testing a pair of composite hypotheses and inter wining the set of functional values on the pair of hypotheses.

According to (Donoho & Liu, 1991) and, (Ibragimov & Khasminski, 1991), the estimator that minimizes the maximum risk is the MiniMax estimator and expressed in equation form as

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}) \tag{2.1}$$

where the infimum is over all estimators $\hat{\theta}$. The right hand side of equation (2.1) is the MiniMax Risk

$$R \equiv R(\theta) = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}), \tag{2.2}$$

Cai and Low (2011), noted that computing an estimator that minimizes the maximum risk is not an easy task and even if the estimator is computed, it depends on an unknown distribution. They proposed the MiniMax rate-optimal and an asymptotically MiniMax estimator for the MiniMax estimator. When the MiniMax rate-optimal estimator with maximum risk (equal to MiniMax risk) is used, the MiniMax estimator is

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}) \asymp \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta}), \qquad n \to \infty \tag{2.3}$$

where (.) $\asymp$ (..) means that both (.)/(..) and (..)/(.) are both bounded as $n \to \infty$. When an asymptotically MiniMax estimator is used, the MiniMax estimator is

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}) \sim \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta}), \qquad n \to \infty \tag{2.4}$$

where (.) $\sim$ (..) means (.)/(..) $\to 1$.

The MiniMax rate-optimal estimator in equation (2.3) are closely related to best (MiniMax) polynomial approximation problem where the best polynomial approximation problem is a convex optimization problem. The connection between the two is important in the sense that the difficult-to solve MiniMax and convex problem of Minimax risk in $\sup_{f \in \mathcal{P}} \mathbb{E}_f [F(f) - \hat{F}]^2$ can be transformed into another efficiently solvable MiniMax and convex problem of minimizing the maximum deviation of the polynomial from a given function.

Cai & Low (2011) developed an estimator based on approximation theory and the Hermite polynomials. They considered the problem when the means were bounded by a given value, the estimator developed was shown to be asymptotically sharp MiniMax. The lower bound was calculated using the difference between the functionals expectation over each of the priors μ_0 and μ_1 and on the variance of the functional under the prior μ_0. The chi-square distance between two marginal distributions of the observations was also used to set the bound.

They applied the technique of transforming the difficult-to solve MiniMax and convex problem into efficiently solvable MiniMax and convex problem where the MiniMax polynomial was used to smoothen the risk at the origin to obtain an optimal estimator. However, the existence of the polynomial factor in the estimate of the density can be worse in the tail regions, where the density might be negative and unable to be integrated to 1.

Nicolas and Jean-Muller (2003) implemented "regular enough" functions on a computing system using the polynomial approximations. They noted that a polynomial that best estimates a function has coefficients that are not exactly representable with a finite number of bits. Nonetheless, polynomial estimations that were actually implemented had coefficients that were expressed by a finite-and often small - number of bits. Then they considered polynomial estimations with at most m_i fractional bits in the degree i coefficient. This enabled them to obtain the best polynomial estimation under this constraint.

The problem of oscillation at the edges of an interval that occurs while using polynomial interpolation with polynomial points was studied by (Nicolas & Jean-Muller, 2003) and (Cai & Low (2011). The best polynomial approximations were used for interpolation. The polynomials were useful for interpolation and at the interpolating points the error between the function and the interpolating polynomial was zero. However, the error was more between the interpolating points.

Lepski et al., (1999) estimated the function $|t|$ on $[-1, 1]$ by its truncated Fourier series where the estimate was based on unbiased estimates of each term in the approximation

$$|t| \approx \sum_{k=1}^{N} c_k \cos(\pi k t) \tag{2.5}$$

of smooth functionals. As a result, they were able to approximate the functional $\int |f(t)|dt$ by the finite sum

$$\sum_{k=1}^{N} c_k \int_{0}^{1} \cos(\pi k f(t))dt \tag{2.6}$$

They choose N in a way that combines the approximation error of equation (2.5) and the "stochastic error"- the one of estimating the smooth functional equation (2.6) through noisy observations. They did note, however that the estimator based on Fourier series has higher accuracy than estimator based on polynomials.

Polynomial approximation was used to approximate the function $f(t) = |t|$ with the best polynomial approximation $P^*(x)$ to a continuous function $f(t)$ having at least $(2k + 2)$ alternating points. In the derivation of the MiniMax lower bounds, the set of these points was used to construct the least favorable priors. The priors were constructed using the Hann-Banach theorem and the Riesz representation theorem.

The Hann-Banach theorem allows any continuous linear functional defined on a subspace of a normed space say X with a continuous extension to the entire subspace. This theorem allows the bounded linear functionals on a subspace of some vector space to be extended to the entire space. When considering vector spaces, the elements are multiplied and added by scalars. The theorem makes it possible to estimate length and angle using vector spaces with inner product structure. In this way, the inner product in infinite-dimensional abstract space leads to Hilbert space. Data in infinite-dimensions are defined as continuous functions, and when the form of the functions in the true model are unknown, the most efficient use of data is to allow the estimated functions to depend on sample size (Chichilnisky, 2009).

The Hilbert space abbreviated by L_2, is a complete vector space with an inner product space $(L \langle ., . \rangle)$. The L_2 is the set of square integrable functions in which the square integrable functions form a complete metric space under the metric brought by the inner product (Larry, 2006). The inner product space is defined by the set and the specific inner product $\langle ., . \rangle$. The significance of forming a complete metric space is to allow the sequences to converge and find a point to which they converge within space.

Bergstrom (1985), demonstrated that there is a limitation on the real line for the use of the Hilbert space. He pointed out that the standard Hilbert space like $L_2(\mathbb{R})$ requires that the unknown function tends to zero at infinity. This makes it unreasonable to be used on certain models like the financial model.

2.4 Saddle-point Approximation

The uniform asymptotic expansion of the distribution of random variables that depends on the sample size is known as saddle-point approximation (Daniels, 1987). The cumulant generating function is used to apply saddle-point approximation to a distribution of some random variables. The cumulant generating function (CGF) of the random variables is used in derivation of the saddle-point expansion. Cumulants such as $k_1 = \mu$ and $k_2 = \sigma^2$ are related to moments.

This approximation can be derived using two techniques: The saddle-point method and the Edgeworth expansion. These methods can be used to improve estimators that do not perform well in the tail areas. The saddle-point method is the method of selecting an integration path that passes through a saddle-point in such a way that the integral is centered in the small area a round the saddle-point. An asymptotic estimate of a complex integral is obtained using this approach.

$$I = \frac{1}{2\pi i} \int_{s} g(z)dz \tag{2.7}$$

where the function g is analytic on the open set $\Omega \subset \mathbb{C}$ and s is a piece-wise smooth path from a to b in Ω. The saddle-point density and distribution functions are obtained when the method is applied to the Fourier inversion formula of a probability density and distribution functions.

The saddle-point approximations are constructed by performing operations on the MGF or, equivalently, the CGF of a random variable. For instance, let $X_1, \ldots, X_n$ be independent, identically distributed random vectors from a density $f(x)$ with respect to the Lebesgue measure. The MGF, $M(t) = \mathbb{E}\exp(tX)$ and CGF, $K(t) = \log M(t)$. The MGF is assumed to exist in the an open neighborhood around the origin.

In statistics, the saddle-point method has been used by a number of researchers. For instance, Daniels (1954) used the saddle-point method to obtain a uniform asymptotic expansion of the sample mean using the Fourier inversion formula of a density function of a sample mean. The expansion was achieved by improving the Edgeworth expansion density function of a sample mean. To obtain a uniform asymptotic expansion (Lugannani and Rice, 1980) applied the method to the Fourier inversion formula of a distribution function to a sample mean.

The saddle-point method was used by (Petrova and Solov'ev, 1997) to derive asymptotic estimates of an integral which was related to the hypergeometric function. Flajolet and Sedgewick (2009) used the saddle-point method to estimate Cauchy coefficient integrals of a generating function, $G(z)$. When the function $G(z)$ was analytic a round the origin on the disc $D(0, r)$, they discovered that $G(z)$ had the Taylor series.

Gatto (2007, 2010) used the method to estimate the likelihood of "ruin" and the "discounted claim quantities" in the sense of a compound Poisson process. Spady (1991) used the method to estimate the distribution of regression estimators defined by a system of estimating equations with observations presumed to be independent. Estimating the distribution of the L_1 regression estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ demonstrated the method's accuracy.

Saddle-point methods provide approximations to densities and probabilities in a variety of settings. In particular their relative errors are bounded in the tail areas, a desired property which is not attained by most other types of approximations used in statistics. The problem of approximating the MGF of a truncated random variable in terms of the MGF of the original random variable was considered by

(Butler and Wood, 2004). They approximated the MGF to ensure the application of saddle-point approximation to certain distributions determined by truncated random variables.

The edgeworth expansion technique is not widely used as the saddle-point method. The technique gives significantly better estimators at the mean of a distribution. The expansions are easy to express in terms of moments. The saddle-point method use these attributes to get an improved estimator at the mean of the distribution by changing the original distribution. At the value of interest, a specific conjugate distribution is chosen such that its mean can be modified back to the original distribution.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter examined the methods that were used to develop composite hypotheses-based estimator of an arbitrary non-smooth functional. An overall MLB for estimating an arbitrary non-smooth functional based on testing a pair of composite hypotheses was derived. The parameter space Λ was split into two disjoint subsets Λ_0 and Λ_1. Such that $H_0 : \lambda \in \Lambda_0$ and $H_1 : \lambda \in \Lambda_1$ are the null and alternative hypotheses respectively. Two priors were constructed such that they had a large difference in the expected values of the functional $T(\lambda)$ and a small difference while making the Chi-square distance between the two mixture models.

The MiniMax lower bound of estimating the non-smooth functional $T(\lambda)$ was developed based on the general MiniMax lower bound derived. The MiniMax Risk developed evaluated the performance of the estimator obtained. The difficult-to solve MiniMax problem was transformed into a solvable MiniMax polynomial where the risk was smoothened at the origin using the best polynomial approximation. Hermite polynomials were used to construct an unbiased estimators for each term in the expansion and the saddle-point approximation techniques were used to modify it. The asymptotic properties of the developed estimator were also considered.

The general MiniMax lower bound was derived by dividing the parameter space Λ into two disjoint subsets Λ_0 and Λ_1 where $H_0 : \lambda \in \Lambda_0$ against $H_1 : \lambda \in \Lambda_1$ are the null and alternative hypotheses respectively. A pair of priors was constructed and the chi-square distance between two priors bounded. The priors were constructed with a large difference in the expected values of the functional $T(\lambda)$ and a small difference when calculating the Chi-square distance between the two mixture models. Nonparametric Estimation of Density Functions, Polynomial approximation, Hermite polynomials and the Hilbert space are among the mathematical ideas discussed.

3.2 Nonparametric Estimation of Density Functions

There are several estimators in literature that statistical researchers have used to estimate density functions in the nonparametric set-up. The estimators include the local average estimator, kernel density estimator, nearest neighborhood estimator, the series estimator, the penalized likelihood estimator. These estimators were discussed by (Pagan and Ullah, 1999), among which the kernel density estimator is the best known estimator, well developed and widely used than others.

A kernel is a mathematical function that returns a probability of a random variable for a given value. It is any smooth function k such that $k(y) \geq 0$ and $\int k(y)dy = 1$, $\int yk(y)dy = 0$ and $\sigma_k^2 = \int y^2 k(y)dy > 0$. A kernel function weights the contribution of observations from a data sample based on their distance to a given sample. A parameter called the "bandwidth" or "smoothing" parameter controls the scope of observations from the data sample that contributes to estimating the probability of a given sample.

In practice, there are a number of kernel functions to choose from, but three are most common choices:

1. the Gaussian kernel

$$k(y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2}), -\infty < y < \infty \tag{3.1}$$

2. the Epanechnikov kernel

$$k(y) = \frac{3}{4}(1 - y^2)I(|y| \leq 1) \tag{3.2}$$

3. the Biweight or Quartic kernel

$$k(y) = \frac{15}{16}(1 - y^2)^2 I(|y| \leq 1) \tag{3.3}$$

Three other kernels that aren't as common are:

4. the Uniform kernel

$$k(y) = \frac{1}{2}I(|y| \leq 1) \tag{3.4}$$

5. the Triangular kernel

$$k(y) = (1 - |y|)I(|y| \leq 1) \tag{3.5}$$

6. the Triweight kernel

$$k(y) = \frac{35}{32}(1 - y^2)^3 I(|y| \leq 1) \tag{3.6}$$

The choice of the kernel function determines the weight given to each observation. For instance, a uniform kernel function assigns equal weights to all points closest to the target and diminishes the weights to those points that are ”farthest” from the centre of the kernel.

3.2.1 Existing Nonparametric Estimators

Some of the existing nonparametric estimators were discussed in this section. They are the Nadaraya-Watson estimator, Local polynomial estimator, the reflection of data technique, the transformation of data technique and the pseudo data methods.

The Nadaraya-Watson estimator was used to estimate finite population totals based on a sample drawn from the population (Dorfman, 1992). A population consisting of N units was considered and an estimate of the finite population total was defined as

$$T = \sum_{N} y_i, \qquad i = 1, \ldots, N \tag{3.7}$$

The estimation of the finite population totals was carried out by first expressing T as the sum of sample component and non-sample component.

$$T = \sum_{i \in s} y_i + \sum_{j \in p-s} y_j \tag{3.8}$$

where s is the sample size and p is the population size. The non-sampled values of the second part in the equation above were estimated. This was done by assuming availability of auxiliary variables as in equation (1.2).

Using a symmetric density function, (Dorfman, 1992) defined the Nadaraya-Watson weights by

$$w_i(x) = \frac{k_b(x_i - x)}{\sum\limits_{i=1}^{n} k_b(x_i - x)}, \tag{3.9}$$

where b is the bandwidth to estimate $f(X_i)$ in equation (1.2) thus, giving the Nadaraya-Watson estimator as

$$\hat{f}(X_i) = \sum_{i} w_i(x) y_i \tag{3.10}$$

The finite population total estimate obtained using the Nadaraya-Watson estimator was biased. The bias was induced at the boundary by the weighting function. This was illustrated by simulating a cubic function $Y = 10 - X^3 + e$, where $X \sim U(1, 2)$ and $e \sim N(0, 0.5)$ and a sample of size $n = 100$. The model used showed the boundary problem clearly. Figure 3.1 was obtained from the data of the simulation done using R statistics.

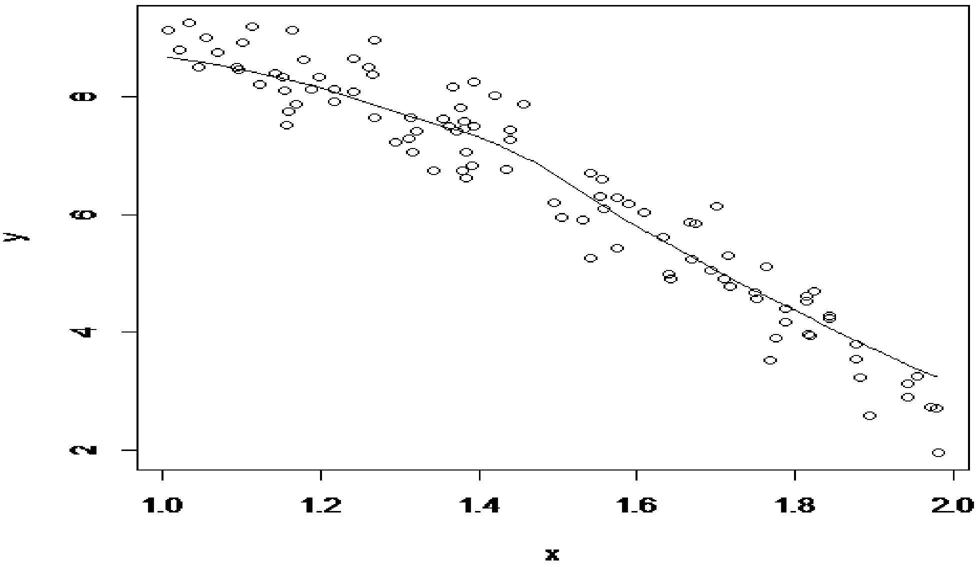

Figure 3.1: The Boundary bias of Nadaraya-Watson estimator

In Figure 3.1, most points are below the fitted line on the right boundary and above the line on the left. The fitted line illustrates how the Nadaraya-Watson estimator failed to capture the trend on the boundaries. In literature, many methods of minimizing the boundary effects have been proposed. For instance, the reflection of data technique, the transformation of data technique and the pseudo data methods.

The reflection of data technique was proposed by (Cline and Hart, 1991) and, (Silverman, 1986). In this method, $-X_1, \ldots, -X_n$ are added to the data set. The values are added since the kernel estimator is penalizing for lack of data on the negative axis. The estimator for this technique was defined by

$$\hat{m}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_i}{h}\right) \right\}, \qquad x \geq 0, \tag{3.11}$$

$\hat{m}(x) = 0$ for $x < 0$

Wand et al.(1991), and (Marron and Ruppert, 1994) studied the data transformation technique where a regular kernel estimator was used with the transformed data set $\{g(X_1), \ldots, g(X_n)\}$. The estimator was given as

$$\hat{m}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - g(X_i)}{h}\right) \right\} \tag{3.12}$$

The estimator was used to estimate the p.d.f. of $g(X)$ and not the p.d.f. of X.

The pseudo data methods, were studied by (Cowling and Hall, 1996). The technique involves generating data beyond the left end point of the support of the density. The technique transforms data into a new set and then puts it on the negative axis. The estimator obtained was defined as

$$\hat{m}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_{-i}}{h}\right) \right\} \tag{3.13}$$

where $m \leq n$ and $X_{(-i)} = -5X_{\frac{1}{3}} - 4X_{(\frac{2}{3}i)} + \frac{10}{3}X_i$

3.3 Polynomial Approximation

A function written in the form $p_k(x) = a_0x^k + a_1x^{k-1}+, \ldots, +a_k$ with some coefficients $a_0, ..., a_k$ is called a polynomial (Smyth, 1998). A linear function and a quadratic function are first and second degree polynomials respectively. Polynomial approximations are among the frequently used methods of evaluating a possibly different function in a small domain. For instance, if $f(x)$ and $p_k(x)$ are two continuous functions in the interval (a, b) and $p_k(x) = p_0x^k + p_1x^{k-1} +, \ldots, +p_k$. The values of $p_0, p_1, \ldots, p_k$ can be obtained such that the absolute value, $D_k = \max_{a \leq x \leq b} |f(x) - p_k(x)|$ between the polynomial $p_k(x)$ and $f(x)$ is small as possible for all x in (a, b).

The polynomial approximations are of two kinds: the least squares approximations and MiniMax approximations. MiniMax approximations are approximations that minimize the the worst-case error, while least squares approximations are those that minimize the "average risk". The distance $||f(x) - p(x)||$ is minimized in both cases, where $p(x)$ is the polynomial of a given degree.

The distance

$$||f(x) - p(x)||_{2[0,a]} = \left(\int_0^a w(x) \left[f(x) - p(x)\right]^2 dx \right)^{1/2} \tag{3.14}$$

is for least squares approximations, where w is a continuous weight function. For MiniMax approximations, the distance is

$$||f(x) - p(x)||_{\infty[0,a]} = \max_{0 \leq x \leq a} |f(x) - p(x)| \tag{3.15}$$

MiniMax approximation seeks for a polynomial of degree k that best estimates the given function in the interval while minimizing the absolute maximum error. Up to $k = 1$, the MiniMax polynomial can be computed analytically , but for $k > 1$ Remez's algorithm can be used. Remez is algorithm is an iterative algorithm based on known optimal approximation for certain $f(x)$. However, there is no general characterization-based algorithm to compute the minimax polynomial approximation.

3.3.1 Orthogonal Polynomials

Orthogonal polynomials are easy to use for finding the coefficients for approximating a function. A family of orthogonal polynomials have a recursive representation which make computations even faster. The Chebyshev polynomials and the Hermite polynomials are some of the orthogonal polynomials. If $P_i(x)$ and $P_j(x)$ are uncorrelated as x varies over $\mathbb{R}^n$, the polynomials P_i and P_j are said to be orthogonal. A recursively defined sequence of orthogonal polynomials can be used to improve control of the interpolation error on the interpolation interval (Sauer, 2006). They have the property of bounded variance, which means that their local maxima and minima on the interval $[-1, 1]$ are equal to 1 and -1 respectively, regardless of the polynomial's order.

The Chebyshev polynomials are of two kinds; the first and second kind denoted by $T_k(x)$ and $U_k(x)$ respectively. The subscript k is the degree of these polynomials (Levy, 2008). The Chebyshev polynomials of the first kind are defined as

$$T_k(x) = \sum_{j=0}^{[k/2]} (-1)^j \frac{k}{k-j} \binom{k-j}{j} 2^{k-2j-1} x^{k-2j} \tag{3.16}$$

The first kind Chebyshev polynomials are solutions to the Chebyshev differential equations;

$$(1-x^2)\frac{d^2y}{dx^2} - x\frac{dy}{dx} + k^2 y = 0, \tag{3.17}$$

for $|x| < 1$. (Rivlin, 1974), gave the expansion

$$|x| = \frac{2}{\pi} T_0(x) + \frac{4}{\pi} \sum_{k=1}^{\infty} (-1)^{k+1} \frac{T_{2k}(x)}{4k^2 - 1} \tag{3.18}$$

where $T_{2k}(x)$ is the Chebyshev polynomial of degree $2k$.

The second kind Chebyshev polynomials are solutions to the Chebyshev differential equations;

$$(1-x^2)\frac{d^2y}{dx^2} - 3x\frac{dy}{dx} + k(k+2)y = 0 \tag{3.19}$$

for $|x| < 1$.

Chebyshev approximation uses Chebyshev polynomials as the basis for polynomials. Let $T_k(x) = \cos[k \arccos x]$, $k \geq 0$ on $[-1, 1]$. Substituting $\alpha = \arccos x$;

$$T_k(x) = \cos(k\alpha), \quad 0 < \theta < \pi \tag{3.20}$$

$$T_{k+1}(x) = \cos(k+1)\alpha = \cos k\alpha \cos\alpha - \sin k\alpha \sin\alpha \tag{3.21}$$

$$T_{k-1}(x) = \cos(k-1)\alpha = \cos k\alpha \cos\alpha + \sin k\alpha \sin\alpha \tag{3.22}$$

Adding equation (3.21) and equation (3.22), we have;

$$T_{k+1}(x) + T_{k-1}(x) = 2\cos k\alpha \cos\alpha \tag{3.23}$$

Making right-hand side appear like a polynomial in x, $\cos\alpha = x$ is substituted in the above equation. Then,

$$\begin{aligned} T_{k+1}(x) &= 2\cos k\alpha \cos\alpha - T_{k-1}(x) \\ &= 2xT_k(x) - T_{k-1}(x) \end{aligned} \tag{3.24}$$

Equation (3.24) is a three-term recurrence relation to generate Chebyshev polynomials (Fox and Parker, 1968). For instance, the recursive relation satisfy:

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \end{aligned}$$

$$\begin{aligned} T_4(x) &= 8x^4 - 8x^2 + 1 \\ T_5(x) &= 16x^5 - 20x^3 + 5x \\ &\vdots \\ T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x), k \geq 1 \end{aligned} \tag{3.25}$$

From equation (3.25), $T_k(x)$ is a polynomial of order k.

The first five Chebyshev polynomials $T_k(x), k = 1, 2, ..., 5$ are shown in the figure below.

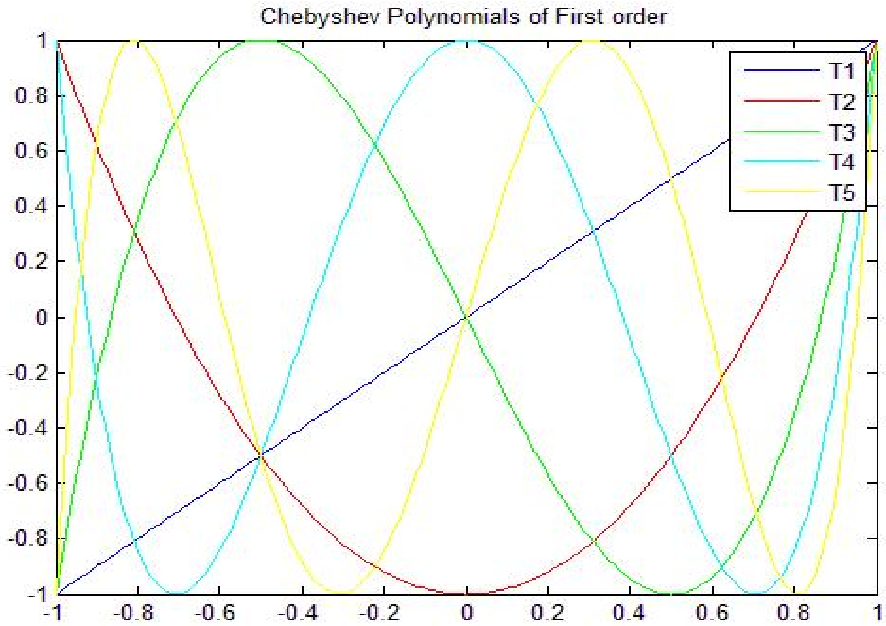

Figure 3.2: Chebyshev polynomials of first kind

Each graph in Figure 3.2 is symmetric to the y-axis or the origin, with a maximum value of 1 and a minimum value of -1 on the interval $[-1, 1]$. The zeros appear to be simple and real. The power of x is even for an even k for every non-zero term of $T_k(x)$. Similarly, the power of x is odd in every non-zero term.

As a result, $T_k(x)$ is an odd function for odd k. $T_k(x)$ is either an even or odd function for any k, which is an important fact about the zeros.

$T_k(x)$ is a Chebyshev polynomial with n roots termed the Chebyshev nodes. The nodes are determined by the formula $x_i = \cos\left(\frac{i-\frac{1}{2}}{n}\right)\pi$, for $i = 1, 2, ..., n$ given by an k^{th} degree polynomial $P_k(x)$ written in terms of $T_0, ..., T_k$

$$P_k(x) = C_0T_0(x) + C_1T_1(x) + ... + C_kT_k(x) - \frac{1}{2}C_0 \tag{3.26}$$

where

$$C_j = \frac{2}{k}\sum_{k=1}^{n+1} f(x_k)T_j(x_k), j = 0, 1, ..., n \tag{3.27}$$

and $x_k, k = 1, ..., n+1$ are zeros of T_{n+1} since $T_j(x) = \cos(j \quad \text{arccos} \quad x)$ and

$$\begin{aligned} T_j(x_k) &= \cos(j \quad \text{arccos} \quad x_k) \\ &= \cos\left(\frac{j(k-\frac{1}{2})}{n+1}\right)\pi \end{aligned} \tag{3.28}$$

The Chebyshev polynomials are a family of orthogonal polynomials on the interval $[-1, 1]$ with respect to a weight function $\frac{1}{\sqrt{1-x^2}}$ (Fox and Parker, 1968).

$$\int_{-1}^{1} T_i(x)T_j(x)\frac{1}{\sqrt{(1-x^2)}}dx = \left\{\begin{matrix} 0, & i \neq j \\ \scriptstyle\pi, & \scriptstyle i=j=0 \\ \scriptstyle\frac{\pi}{2}, & \scriptstyle i=j\neq 0 \end{matrix}\right\} \tag{3.29}$$

The weight function assigns varying degrees of importance to certain portions of the interval $[-1, 1]$. As seen in Figure 3.3, the weight function provides less emphasis in the interval's center and more emphasis around 1 and -1. On the interval, the weight function is positive almost everywhere and has a finite integral.

According to (Cai and Low, 2011), computing the Chebyshev polynomials is not easy. Instead, MiniMax polynomials or the Best polynomials are used. When the function $f(x)$ is continuous on an interval $[a, b]$ as stated by the Weierstrass

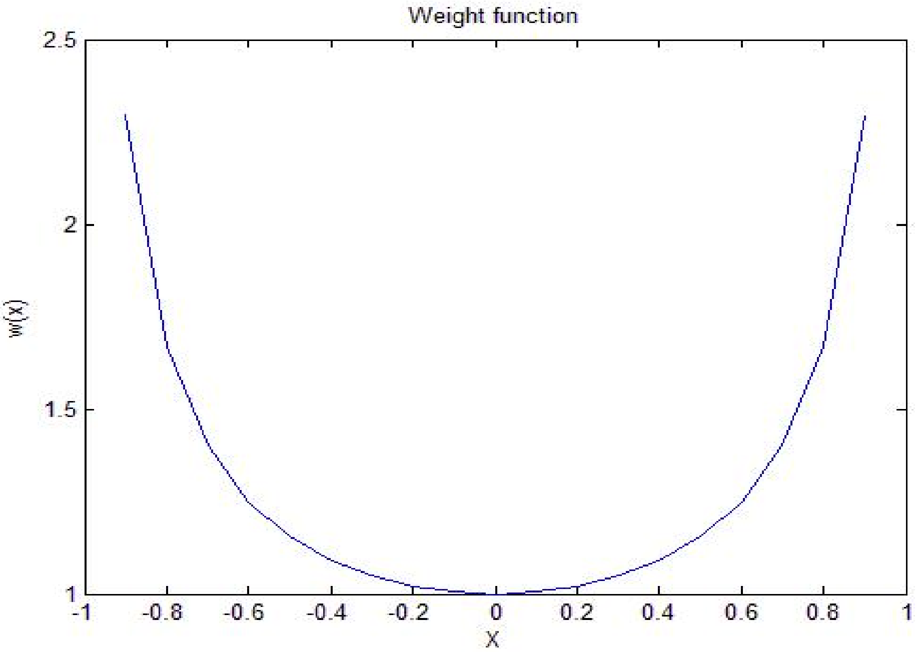

Figure 3.3: The weight function of orthogonal Chebyshev polynomial, $\boldsymbol{w}(x) = \frac{1}{\sqrt{1-x^2}}$

Approximation Theorem, the MiniMax polynomials exist and are unique. Any continuous function can be approximated as close as possible with polynomials, according to the Weierstrass Approximation Theorem, assuming that the polynomials can be of any degree. The theorem was formulated in L^∞ form, although it also holds in the L^2 form.

3.3.2 Weierstrass Approximation Theorem

The space of polynomials of degree $\leq n$ is denoted as $\mathcal{P}_n$. Let $f(x)$ be a continuous function on $[a, b]$. Then there are polynomials $P_n(x)$ that converges uniformly to $f(x)$ on $[a, b]$ that is, $\forall \epsilon > 0$, there exists an $N \in \mathbb{N}$ and polynomials $P_n(x) \in \mathcal{P}_n$, such that $\forall x \in [a, b]$

$||f(x) - P_n(x)||_\infty \leq \epsilon \quad \forall n \geq N$

where $||.||_\infty$ is the sup norm or L^∞ norm:

$$||f(x) - P_n(x)||_\infty = \max_{x\in[-1,1]} |f(x) - P_n(x)| \tag{3.30}$$

When $n \to \infty$, $P_n(x) \to f(x)$. In other words,

$$\lim_{n\to\infty} P_n(x) = f(x) \quad \forall x \in [a, b] \tag{3.31}$$

The theorem was based on the Bernstein polynomials defined on the interval $[0, 1]$. It shows the existence of a set of polynomial functions, but it does not provide a general method of finding one, and the polynomial functions are not guaranteed to converge uniformly.

A simulation experiment was performed to demonstrate the Weierstrass Approximation Theorem and Figure 3.4 below were obtained. Let $f(x) = x^2 \sin 10x$ be a continuous function on the interval $[-1, 1]$ and T_k be the Chebyshev polynomial where $k = 1, 4, 9, 16$. The following figures represent plots of $f(x)$ and T_k obtained by varying the degree k of the polynomial on the interval $[-1, 1]$.

There are polynomials $T_k \in \mathcal{P}k$ that converge uniformly to $f(x)$ on the interval $[-1, 1]$ as shown in Figure 3.4. The uniform norm, $||f - T_k|| = {}_{-1\leq x\leq 1}^{\max} |f(x) - T_k(x)|$ can be used to quantify how best the Chebyshev polynomials $T_k \in \mathcal{P}_k$ converges to $f(x)$. The norm gives the error of approximation as the largest distance between $f(x)$ and $T_n(x)$ (Rivlin, 1990).

In approximation theory, the Best polynomial approximation has been extensively researched (Rivlin, 1990). $P(x) \in \mathcal{P}_k$ is the closest polynomial to f for any $k \geq 0$, and $\mathcal{P}_k$ is the class of all real polynomials of degree at most k for every

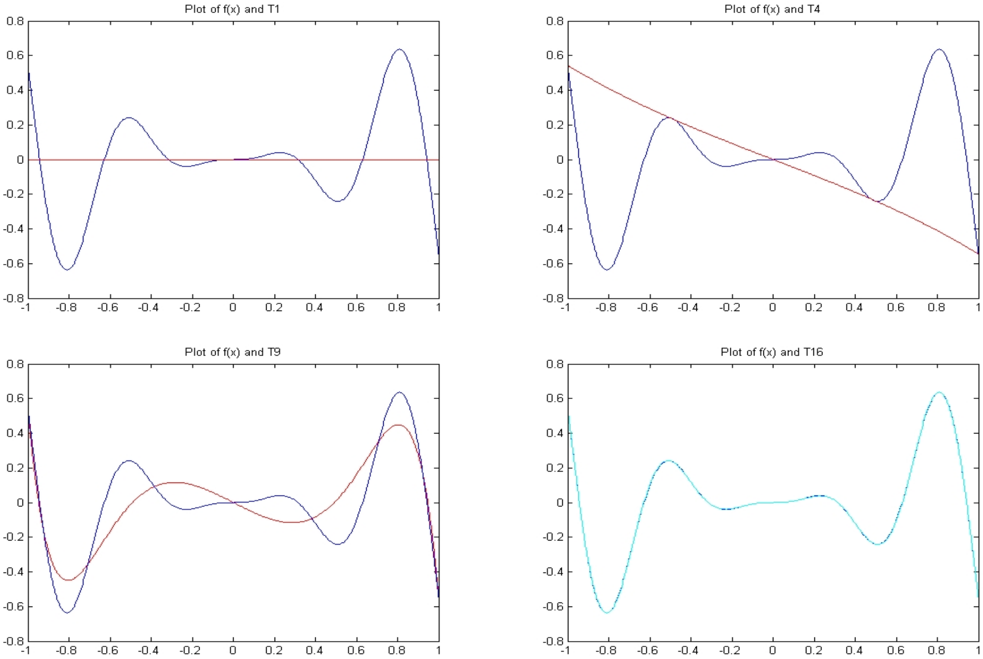

Figure 3.4: (a),(b),(c),(d) plots of $f(x)$ and $T_k, k = 1, 4, 9, 16$

continuous function f on $[-1, 1]$,

$$\delta_k(f) = \max_{x\in[-1,1]} |f(x) - P(x)| \tag{3.32}$$

where δ_k is the distance in the uniform norm on $[-1, 1]$ from the absolute value function $f(x) = |x|$ to the space $\mathcal{P}_k$, the class of all real polynomials of degree at most k.

The polynomials $P_k(x) \in \mathcal{P}_k$ exist and converges uniformly to the function $f(x)$ on $[-1, 1]$ that is, $\forall\epsilon > 0$, there exists polynomials $P_k(x) \in \mathcal{P}_k$, such that $\forall x \in [-1, 1]$
$||f(x) - P_k(x)||_\infty \leq \epsilon \quad \forall k \geq K$
where $||.||_\infty$ is the sup norm or L^∞ norm:

$$|f(x) - P_k(x)||_\infty = \max_{x\in[-1,1]} |f(x) - P_k(x)| \tag{3.33}$$

A polynomial $P_k^*(x)$ is said to be a best polynomial approximation of the function f if

$$\delta_k(f) = \max_{x \in [-1,1]} |f(x) - P_k^*(x)| \tag{3.34}$$

According to the classical Chebyshev alternation theorem, a polynomial $P_k^*(x) \in \mathcal{P}_k$ is the (unique) best polynomial that converges uniformly to a continuous function f if and only if the difference $f(x) - P_k^*(x)$ takes its maximal value with alternating signs at least $(k+2)$ times. This means that there exist $k+2$ points $-1 \leq x_0 < ... < x_{k+1} \leq 1$ such that

$$[f(x_i) - P_k^*(x_i)] = \pm(-1)^i \max_{x \in [-1,1]} |f(x) - P_k^*(x)|; \tag{3.35}$$

$i = 0, ..., k+1$

The polynomial $P_k(x)$ of degree at most k represents the error in polynomial interpolation of the function $f(x)$ with the nodes, $x_0, \ldots, x_k$. The error is given as

$$E = f(x) - P_k(x) = \frac{f^{k+1}(x)}{k+1!}\theta_i \tag{3.36}$$

Using the MiniMax property of the monic polynomial of degree $(k+1)$, the $(k+1)$ nodes are chosen nodes to minimize $|\theta_i|$ in $[-1, 1]$.

Using an estimator that satisfies equation (2.1) is difficult, an estimator that achieves equation (2.3) the MiniMax rate is used. The best rate of convergence is based on the requirement that an estimator works admirably on a fixed model and even at a sequence of models that are adjacent. The rate alludes to a model of interest achieving point-wise convergence uniformly over models in a small neighborhood.

The error in polynomial interpolation of a function $f(x)$ with the nodes, $x_0, x_1, ..., x_k$, by the polynomial $P_k(x)$ of degree at most k is given by

$$E = f(x) - P_k(x) = \frac{f^{k+1}(\xi)}{k+1!}\theta_i \tag{3.37}$$

where $\theta_i = (x - x_0)(x - x_1)...(x - x_k)$. The choice of these $(k + 1)$ nodes so that to minimize $|\theta_i|$ in $[-1, 1]$ is done using the MiniMax property of the monic polynomial of degree $(k + 1)$

$$\max_{x \in [-1,1]} |\widehat{T_{n+1}(x)}| \leq \max_{x \in [-1,1]} |\theta_i| \tag{3.38}$$

3.4 The Variance and the Bias

The variance and the bias are asymptotic properties of an estimator. These are estimators that hold as the sample size increases. They are important properties that a statistical researcher would be interested to check given an estimator.

The bias and the variance are used to measure accuracy and precision of an estimator respectively. These two components are incorporated in the mean squared error (Tsybakov, 2009). The variance and bias are controlled to find an estimator with good properties; small variance and bias (Douglas and George, 2008).

$$\begin{aligned} MSE(T) &= \mathbb{E}(T - \theta)^2 \\ &= E[\{T - \mathbb{E}\theta\} + \{\mathbb{E}(T) - \theta\}]^2 \\ &= Var(T) + \{\mathbb{E}(T) - \theta\}^2 \\ &= Var(T) + (bias)^2 \end{aligned} \tag{3.39}$$

where $(bias)^2 = \{\mathbb{E}(T) - \theta\}^2$ denotes the bias of T and $Var(T)$ denotes the variance of T.

The density function $f(x)$ of a distribution is the derivative of the cumulative distribution function $F(x) \equiv Pr\left\{x_i \leq x\right\}$, and the empirical c.d.f.

$$\hat{F}(x) \equiv \frac{1}{N}\sum_{i=1}^{n} 1\left\{x_i \leq x\right\} \tag{3.40}$$

is the natural NP estimator of the c.d.f., it seems reasonable to use the estimation of f on the empirical c.d.f. The empirical distribution is used to calculate a relative error of order $n^{-\frac{1}{2}}$ which decreases as the density in the distribution's tails increases. Although $\hat{F}$ is $\sqrt{n}$ - consistent and asymptotically normal, estimating f by differentiating $\hat{F}$ is difficult because its derivative is either zero or undefined.

By defining of the density f as the (right-) derivative of the c.d.f.

$$f(x) = \lim_{h\to 0}\frac{F(x+h)-F(x)}{h} \tag{3.41}$$

the estimator of the density f can be obtained by a corresponding difference ratio of $\hat{F}$

$$\begin{aligned} \hat{f}(x) &= \frac{\hat{F}(x+h)-\hat{F}(x)}{h} \\ &= \frac{1}{n}\sum_{i=1}^{n} I\left\{x < x_i \leq x+h\right\} \end{aligned} \tag{3.42}$$

where h is a smoothing parameter with a small positive value depending on the sample size.

Choosing the sequence h_n such that the mean bias and variance of $\hat{f}$ both tend to zero as the sample size increases is needed to show MSE consistency of $\hat{f}$. Since the empirical c.d.f., $\hat{F}$ is unbiased estimator of F the bias of $\hat{f}$ is clearly

$$\mathbb{E}[\hat{f}(x)] - f(x) = \frac{F(x+h)-F(x)}{h} - f(x) \to 0 \tag{3.43}$$

if $h = h_n \to 0$ as $n \to \infty$. The variance of $\hat{f}(x)$ is

$$\begin{aligned} Var(\hat{f}(x)) &= Var\left(\frac{1}{nh}\sum_{i=1}^{n} I\left\{x < x_i \leq x + h\right\}\right) \\ &= \frac{1}{nh^2}Var\left(1\left\{x < x_i \leq x + h\right\}\right) \\ &= \frac{1}{nh}\left[\frac{F(x+h) - F(x)}{h}(1 - F(x+h) + F(x))\right] \\ &= \frac{f(x)}{nh} + O\left(\frac{1}{n}\right) \end{aligned} \tag{3.44}$$

which will approach to zero if $nh = nh_n \to \infty$ as $n \to \infty$

The disparity between the density function $f(x)$ and its estimator $\hat{f}(x)$ can be measured using the loss function which is expressed as below

$$MSE = R(f(x), \hat{f}(x)) = \mathbb{E}(L(f(x) - \hat{f}(x))) \tag{3.45}$$

The expectation is calculated in terms of the distribution that produces the observation used by $\hat{f}(x)$.

Similarly, the integrated risk or integrated mean squared error defined by

$$R(f(x), \hat{f}_n(x)) = \int R(f(x) - \hat{f}_n(x))dx \tag{3.46}$$

can be used to summarize the risk over various values of x. The integrated MSE or the average squared error

$$R(r, \hat{r}_n) = \frac{1}{n}\sum_{i=1}^{n} R(r(x_i) - \hat{r}_n(x_i)) \tag{3.47}$$

is used to solve a regression problem. From equation (3.46), equation (3.47) can be written as;

$$L_2 = \int (\hat{f}_n(x) - f(x))^2 dx \tag{3.48}$$

where L_2 is the integrated squared error loss function.

The MSE is a function of both the bias and the variance. Controlling either the bias or the variance does not guarantee that MSE is controlled. For example, when the bias term is large and the variance is small, the data are oversmoothened; when the bias term is small and the variance is large, the data are undersmoothened. Minimizing the risk therefore, corresponds to balancing both the bias and variance.

From equation (3.42), h is a smoothing parameter and it is choosen to minimize an estimate of the risk. Writing equation (3.48) as a function of h,

$$\begin{aligned} L_2(h) &= \int (\hat{f}_n(x) - f(x))^2 dx \\ &= \int \hat{f}_n^2(x)dx - 2\int \hat{f}_n(x)f(x)dx + \int f^2(x)dx \end{aligned} \quad (3.49)$$

The last term in equation (3.49) does not depend on n, so minimizing the loss function is equivalent to minimizing the expected value of

$$L_2(h) = \int \hat{f}_n^2(x)dx - 2\int \hat{f}_n(x)f(x)dx \quad (3.50)$$

The error of estimations is minimized using the squared error risk. The risk relies on an unknown parameter thus, making it difficult to find the "best guess". For example, if $\hat{T}(X) = T(\xi)$ where $\hat{T}(X)$ denotes the estimator of $T(\xi)$, then $\hat{T}(X)$ denotes the minimum variance unbiased estimator and MLE. But if $\hat{T}(X) \neq T(\xi)$, then

$$R(\hat{T}(X), T(\xi)) = \int_{i=1}^{n} \mathbb{E}_\xi(\hat{T}(X) - T(\xi))^2 = n\sigma_n^2 \quad (3.51)$$

The problem of normal means, $y_i \sim N(\xi_i, 1)$, $i = 1, \ldots, n$ can be used to obtain estimators with relatively small risks than the risk in equation (3.51) which depends on n (Cai and Low, 2011).

3.5 Test statistics of the Estimator in Hilbert Space

An inner product generates an analytic distance (norm). A norm characterized by the inner product $\langle ., . \rangle$ will define the accompanying measure;

$$d(x, y) = ||x - y|| = \sqrt{\langle x - y, x - y \rangle}, \forall x, y \in X \tag{3.52}$$

The inner product norms fulfill properties that are not fulfilled by all norms. For instance, a complex vector space V is an inner product space (or a pre-Hilbert space) if there is a mapping $(., .) : V \times V \to \mathbb{C}$, called an inner product, that fulfills $\forall x, y, z \in V$, $\forall \alpha \in \mathbb{C}$:

1. $(x, x) \geq 0$
2. $(x, x) = 0 \Leftrightarrow x = 0$
3. $(x, y + z) = (x, y) + (x, z)$
4. $(x, \alpha y) = \alpha(x, y)$
5. $(x, y) = (y, x)^*$

When X is a standard normal on $(-\infty, +\infty)$, the Hermite polynomials which are uncorrelated and the orthogonal basis of the Hilbert space of functions satisfying

$$\int_{-\infty}^{\infty} |f(x)|^2 w(x) dx < \infty \tag{3.53}$$

are formed. The integral and the Gaussian weight function $w(x)$ in equation (3.53) gives an inner product.

Let ϕ be the density function of the standard normal variable, then the Hermite polynomials $H_k(x)$ with respect to ϕ for positive integers k are defined by the equation

$$\frac{d^k}{dx^k}\phi(x) = (-1)^k H_k(x)\phi(x) \tag{3.54}$$

or

$$H_k(x) = (-1)^k e^{x^2/2}\frac{d^k}{dx^k}e^{-x^2/2} \tag{3.55}$$

Thus we can obtain $H_0(x) = 1, H_1(x) = x, H_2(x) = x^2 - 1, \ldots, H_{n+1}(x) = xH_{n-1}(x)$. By differentiating (3.54), we obtain

$$\frac{d}{dx}[H_k(x)\phi(x)] = H_{k+1}(x)\phi(x) \tag{3.56}$$

For this version of the polynomial,

$$\int H_k^2(x)\phi(x)dx = k! \tag{3.57}$$

and

$$\int H_k(x)H_j(x)\phi(x)dx = 0 \tag{3.58}$$

when $k \neq j$.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

The findings of our research are presented in this chapter. The results on the general MiniMax lower bound derived and the MiniMax lower bound for estimating the non-smooth functional are shown in section 4.2 and 4.3 respectively. The developed estimator's asymptotic properties are highlighted in section 4.5.

4.2 The General MiniMax Lower Bound

One of the key components for the development of MiniMax lower bound is the general MiniMax lower bound for estimating an arbitrary functional $T(\lambda) = \frac{1}{n}\sum_{i=1}^{n}|\lambda_i|$. Let the estimator of $T(\lambda)$ based on X be $\hat{T}(X) = \hat{T}$, where X is a random sample with the probability distribution P_λ and $\lambda \in \Lambda$. The bias of $\hat{T}(X)$ is denoted $\mathbb{E}_\lambda \hat{T}(X) - T(X)$, and the prior distributions based on Λ_0 and Λ_1 are w_0 and w_1 respectively. The means and variances of $T(\lambda$ under the pair of priors are:

$$\mathbb{E}(T(\lambda)) = m_i = \int T(\lambda) w_i(d\lambda) \tag{4.1}$$

$$var(T(\lambda)) = v_i^2 = \int (T(\lambda) - m_i)^2 w_i(d\lambda) \tag{4.2}$$

When the prior is w_i, the marginal distribution of X is F_i, and the density of X is f_i. The chi-square distance, S between f_0 and f_1 is defined as

$$S = \left\{ \mathbb{E} f_0 \left(\frac{f_1(X) - f_0(X)}{f_0(X)} \right)^2 \right\}^{1/2} \tag{4.3}$$

The average risk for the estimator $\hat{T}(X)$ under any mixture prior $pw_0+(1-p)w_1$, where $0 \leq p \leq 1$ can be obtained as follows:

$$\mathbb{E}_{f0}\left\{\left(\hat{T}(X)-m_0\right)\left(\frac{f_1(X)-f_0(X)}{f_0(X)}\right)\right\} = \left(m_1+\int B(\lambda)w_1(d\xi)\right)-\left(m_0+\int B(\lambda)w_0(d\lambda)\right)$$

$$\begin{aligned}
\mathbb{E}_{f_0}\left(\hat{T}(X)-m_0\right)^2 &= \int \mathbb{E}_\lambda\left(\hat{T}(X)-m_0\right)^2 w_0(d\lambda)\\
&= \int \mathbb{E}_\lambda\left(\hat{T}(X)-T(\lambda)+T(\lambda)-m_0\right)^2 w_0(d\lambda)\\
&= \int \mathbb{E}_\lambda\left(\hat{T}(X)-T(\lambda)\right)^2 w_0(d\lambda)+\int\left(T(\lambda)-m_0\right)^2 w_0(d\lambda)\\
&+ 2\int\left(\hat{T}(X)-T(\lambda)\right)\left(T(\lambda)-m_0\right)w_0(d\lambda)\\
&= \epsilon^2+\upsilon_0^2+2\upsilon_0\epsilon = (\epsilon+\upsilon)^2 \qquad (4.4)
\end{aligned}$$

Using the Cauchy-Schwarz inequality to prove that the triangle inequality holds for equation (4.4), we obtain,

$$\mathbb{E}_{f_0}\left\{\left(\hat{T}(X)-m_0\right)\left(\frac{f_1(X)-f_0(X)}{f_0(X)}\right)\right\} \leq \left(\mathbb{E}_{f_0}(\hat{T}(X)-m_0)^2\right)^{1/2}.S \leq (\epsilon+\upsilon)S \qquad (4.5)$$

Hence,

$$\left(m_1+\int B(\xi)w_1(d\lambda)\right)-\left(m_0+\int B(\lambda)w_0(d\lambda)\right) \leq (\epsilon+\upsilon)S \qquad (4.6)$$

and it follows

$$\int B(\lambda)w_1(d\lambda)-\int B(\lambda)w_0(d\lambda) \leq m_0-m_1+(\epsilon+\upsilon)S \qquad (4.7)$$

The quadratic equation of the mixture prior can be given

$$Q(x) = px^2+(1-p)(a-bx)^2 \qquad (4.8)$$

where $0 < p < 1$, $a > 0$ and $b > 0$. Differentiating and equating equation (4.8) to zero,

$$\begin{aligned}
Q'(x) &= 2px-2b(1-p)(a-bx)\\
&= 2px-2b(1-p)(a-bx) \qquad (4.9)
\end{aligned}$$

to minimize Q'(x) then Q'(x)=0

$$\begin{aligned} 2px &= 2b(1-p)(a-bx) \\ px &= (1-p)(ab-b^2x) \\ px &= ab-b^2x-abp+b^2xp \\ px+b^2x-b^2xp &= ab-abp \\ x(p+b^2-b^2p) &= ab(1-p) \\ x &= \frac{ab(1-p)}{p+b^2-b^2p} \end{aligned} \tag{4.10}$$

The quadratic equation $Q(x)$ of the mixture prior is minimized when

$$x = x_{min} = \frac{ab(1-p)}{p+b^2(1-p)} \tag{4.11}$$

and that at this value, $a-bx>0$ and $Q(x_{min}) = \frac{a^2p(1-p)}{p+b^2(1-p)}$. At the same value, the quadratic $px^2-(1-p)\left(\max(a-bx,0)\right)^2$ is also minimized. From equation (4.7),

$$\int B^2(\lambda)w_1(d\lambda) \geq \left(\max\left[m_1-m_0-\upsilon_0 S-(S+1)\epsilon,0\right]\right)^2 \tag{4.12}$$

Let $a = m_1-m_0-\upsilon_0 S$ and $b=S+1$ for $0\leq p\leq 1$, then using equation (4.11) and equation (4.12) we have

$$\begin{aligned} p\epsilon^2+(1-p)\int B^2(\lambda)w_1(d\lambda) &\geq p\epsilon^2+(1-p)[\max(m_1-m_0-\upsilon_0 S-(S+1)\epsilon,0)]^2 \\ &\geq \frac{p(1-p)(|m_1-m_0|-\upsilon_0 S)^2}{p+(1-p)(S+1)^2} \end{aligned} \tag{4.13}$$

The average risk under the mixture prior is "large" according to equation (4.13). Analyzing the MiniMax risk measures the statistical complexity of the estimation problem where a real valued loss function $L(T(\lambda),\hat{T}(X))$ quantifies the amount by which the prediction of $T(\lambda)$ deviates from its estimator, $\hat{T}(X)$. Given an estimator $\hat{T}(X)$, the squared error loss

$$L\left(T(\lambda),\hat{T}(X)\right) = \int_{i=1}^{n}(T(\lambda)-\hat{T}(X))^2 = ||T(\lambda)-\hat{T}(X)||^2 \tag{4.14}$$

is utilized with risk function;

$$R\left(T(\lambda), \hat{T}(X)\right) = \mathbb{E}_\lambda \left(L(T(\lambda) - \hat{T}(X)\right) = \int\limits_{i=1}^{n} \mathbb{E}_\lambda \left(T(\lambda) - \hat{T}(X)\right)^2 (d\lambda) \quad (4.15)$$

According to (Brown and Low, 1996), the maximum risk is often at least as high as the average risk. As a result of equation (4.15), we have

$$\int \mathbb{E}_\lambda \left(\hat{T}(X) - T(\lambda)\right)^2 w_i(d\lambda) \geq \frac{|(m_1 - m_0| - \upsilon_0 S)^2}{(S+2)^2} \quad (4.16)$$

which yields the MiniMax risk's general lower bound.

This MiniMax Risk formalizes the possibility of the presence of the best rate of convergence and utilized as a benchmark for assessing the performance the estimation method.

4.3 The MiniMax Lower Bound for estimating the non-smooth functional

The MiniMax lower bound for estimating the non-smooth functional $T(\lambda) = \frac{1}{n}\sum_{i=1}^{n} |\lambda_i|$ was developed using the general lower bound from the previous section as a starting point. It was also necessary to find the least favourable prior distributions ω_0 and ω_1, as well as an effective upper bound for the Chi-square distance between the marginal distributions.

Let ω_0 and ω_1 be the two priors with special properties. A linear functional T can be extended to $C[-1, 1]$ using the Hahn-Banach Theorem without increasing the norm of the functional, the Riesz representation Theorem states that for each $g \in C[-1, 1]$

$$T(g) = \int\limits_{-1}^{1} g(t)\tau(dt) \quad (4.17)$$

where τ is a Borel signed measure with variance equal to 1. Taking $z = 2\tau$, then $\int\limits_{-1}^{1} |t|z(dt) = 2\delta_k$ and $\int\limits_{-1}^{1} t^l z(dt) = 0$, for $l = 0, \ldots, k$. Let z_1 and z_0 be the positive and the negative components of z. Then both z_1 and z_0 are symmetric. Since the variation of z is 2 and $\int\limits_{-1}^{1} z(dt) = 0$ it also follows that both z_1 and z_0 are probability measures which satisfy

$$\int\limits_{-1}^{1} t^l z_1(dt) = \int\limits_{-1}^{1} t^l z_0(dt) \tag{4.18}$$

for $l = 0, \ldots, k$ and

$$\int\limits_{-1}^{1} |t|z_1(dt) - \int\limits_{-1}^{1} |t|z_0(dt) = 2W\delta_k \tag{4.19}$$

where δ_k is the distance between the absolute value function $f(t) = |t|$ and the space $\mathcal{P}_k$ of polynomials of order k in the uniform norm on $[-1, 1]$.

For an even integer k_n, let z_0 and z_1 be two probability measures. Let $g(x) = Wx$ and let w_i be probability measures on $[-W, W]$ defined by $w_i(A) = z_i(g^{-1}(A))$ for $i = 0, 1$. As a consequence,

1. w_0 and w_1 are symmetric a round 0
2. $\int\limits_{-W}^{W} t^l w_1(dt) = \int\limits_{-W}^{W} t^l w_0(dt)$ for $l = 0, \ldots, k$
3. $\int\limits_{-W}^{W} |t|w_1(dt) - \int\limits_{-W}^{W} |t|w_0(dt) = 2\delta_k$

Let w_0^n and w_1^n be product priors with $w_i^n = \prod\limits_{j=1}^{n} w_i$. On the coordinates, these are n independent priors. We have

$$\mathbb{E}_{w_1^n} T(\lambda) - \mathbb{E}_{w_0^n} T(\lambda) = \mathbb{E}_{w_1}|\lambda| - \mathbb{E}_{w_0}|\lambda| = 2W\delta_{k_n} \tag{4.20}$$

and

$$\mathbb{E}_{w_1^n} \left(T(\lambda) - \mathbb{E}_{w_0^n} T(\lambda)\right)^2 = \frac{1}{n}\mathbb{E}_{w_1}(|\lambda| - \mathbb{E}_{w_0}|\lambda|)^2 \leq \frac{W^2}{n} \tag{4.21}$$

Put $f_{0,W}(y) = \int \phi(y-t)w_0(dt)$ and $f_{1,W}(y) = \int \phi(y-t)w_1(dt)$. Because $g(x) = e^{-x}$ is a convex function of x and w_0 is symmetric and $y_i \sim N(\lambda_i, 1), i = 1, \ldots, n$

$$\begin{aligned} f_{0,W}(y) &\geq \frac{1}{\sqrt{2\pi}} e^{(-\int \frac{(y-t)^2}{2} w_0(dt))} \\ &= \phi(y) e^{(-\frac{1}{2}W^2 \int t^2 z_0(dt))} \\ &\leq \phi(y) e^{(-\frac{1}{2}W^2)} \end{aligned} \tag{4.22}$$

The Hermite polynomial $H_k(y)$ is defined in (3.56). Then

$$\phi(y - \alpha t) = \sum_{k=0}^{\infty} H_k(y)\phi(y)\frac{\alpha^k t^k}{k!} \tag{4.23}$$

and it follows that

$$\int \frac{(f_{1,W}(y) - f_{0,W}(y))^2}{f_{0,W}(y)} dy \leq e^{\frac{W^2}{2}} \sum_{k=k_n+1}^{\infty} \frac{1}{k!} W^{2k} \tag{4.24}$$

Then the Chi-square distance, S_n^2 between $f_{0,W}$ and $f_{1,W}$ is given as

$$S_n^2 = \int \frac{(\prod\limits_{i=1}^{n} f_{1,W}(y_i) - \prod\limits_{i=1}^{n} f_{0,W}(y_i))^2}{\prod\limits_{i=1}^{n} f_{0,W}(y_i)} dy_1, \ldots, dy_n \tag{4.25}$$

Hence,

$$\begin{aligned} S_n^2 &= \int \frac{(\prod\limits_{i=1}^{n} f_{1,W}(y_i))^2}{\prod\limits_{i=1}^{n} f_{0,W}(y_i)} dy_1, \ldots, dy_n - 1 \\ &= (\prod_{i=1}^{n} \int \frac{(f_{1,W}(y_i))^2}{f_{0,W}(y_i)} dy_i) - 1 \\ &\leq (1 + e^{\frac{W^2}{2}} \sum_{k=k_n+1}^{\infty} \frac{1}{k!} W^{2k})^n - 1 \\ &\leq (1 + e^{\frac{3}{2}W^2} \frac{1}{k_n!} W^{2k-n})^n - 1 \end{aligned} \tag{4.26}$$

Since $k! > (\frac{k}{e})^k$, then

$$\leq \left(1 + e^{\frac{3}{2}W^2} \left(\frac{eW^2}{k_n}\right)^{k_n}\right)^n - 1 \tag{4.27}$$

Let k_n be the smallest positive integer that satisfies the condition $k_n \geq \frac{logn}{\log\log n} + \frac{\log n}{(\log\log n)^{3/2}}$ then $S_n \to 0$. Let $z_0 \leq \frac{W}{\sqrt{n}}$ and by equation (2.1) we obtain the MiniMax risk for estimating $T(\lambda) = \frac{1}{n}\sum\limits_{i=1}^{n} |\lambda_i|$ over $\Lambda_n(W) = \{\lambda_i \in \mathbb{R}n : |\lambda_i| \leq W\}$ and $W > 0$ bounded from below as

$$\begin{aligned}\inf_{\hat{T}} \sup_{\lambda\in\Lambda_n(W)} \mathbb{E}\left(\hat{T} - T(\lambda)\right)^2 &\geq \frac{(2W\delta_{k_n} - \frac{W}{\sqrt{n}}S_n)^2}{(S_n+2)^2} \\ &= \beta_*^2 W^2 \left(\frac{\log\log n}{\log n}\right)^2 (1+o(1)) \qquad (4.28)\end{aligned}$$

where β_* is a Bernstein constant.

4.4 The Developed Estimator

To develop an estimator of $T(\lambda) = \frac{1}{n}\sum\limits_{i=1}^{n} |\lambda_i|$ over the bounded parameter $\lambda \in \Lambda_n(W)$ where $\Lambda_n(W) = \{\xi \in \mathbb{R}^n : |\lambda_i| \leq W\}$, two cases were considered; $W = 1$ and the general case $W > 0$. When $W = 1$, truncating the expansion (3.18) and let

$$P_K(x) = \frac{2}{\pi}T_0(x) + \frac{4}{\pi}\sum_{k=1}^{K}(-1)^{k+1}\frac{T_{2k}(x)}{4k^2-1} \qquad (4.29)$$

Let $P_K(x)$ be written as

$$P_K(x) = \sum_{k=0}^{K} p_{2k}x^{2k} \qquad (4.30)$$

From equation (3.18) each $|x_i|$ can be estimated by the polynomial $P_K^*(x_i) = \sum\limits_{k=0}^{K} p_{2k}^* x_i^{2k}$ on the interval $[-1,1]$ and hence the functional $T(\lambda) = \frac{1}{n}\sum\limits_{i=1}^{n}|\lambda_i|$ can be estimated by the polynomial

$$\begin{aligned}\tilde{T}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} P_K^*(\lambda_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=0}^{K} p_{2k}^*\lambda_i^{2k}\right) \\ &= \sum_{k=0}^{K} p_{2k}^* a_{2k}(\lambda)\end{aligned} \qquad (4.31)$$

where $a_{2k}(\lambda) \equiv \frac{1}{n}\sum\limits_{i=1}^{n} \lambda_i^{2k}$ and p_{2k}^* are the coefficients of the best polynomial estimation of $|\lambda|$ over $[-1,1]$ up to degree K and each $|\lambda_i|$ can be estimated by the polynomial $P_K^*(\lambda_i) = \sum\limits_{k=0}^{K} p_{2k}^* \lambda_i^{2k}$ on the interval $[-1,1]$

Using the Hermite polynomials, the coefficients, $a_{2k}(\lambda)$ are estimated separately for each k. For each positive integer k, $X \sim N(\mu,1)$, $\mathbb{E}H_k(X) = \mu^k$. Since $\mathbb{E}H_k(x_i) = \lambda_i^k$ for each i when $x_i \sim N(\lambda,1)$, $\frac{1}{n}\sum\limits_{i=1}^{n} \lambda_i^k \equiv a_k(\lambda)$ which can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} H_k(x_i) = \bar{A}_k \tag{4.32}$$

and the estimator of $T(\lambda)$ defined by

$$\begin{aligned} \widehat{T_K(\lambda)} &= \sum_{k=0}^{K} p_{2k}^* \frac{1}{n}\sum_{i=1}^{n} H_{2k}(x_i) \\ &= \sum_{k=0}^{K} p_{2k}^* \bar{A}_{2k} \end{aligned} \tag{4.33}$$

where $H_k(x)$ is a Hermite polynomial with respect to ϕ; where ϕ is the density function of a standard normal variable as shown in equation (3.56).

For the general case $W > 0$, each λ_i is rescaled over the bounded parameters $\Lambda_n(W)$ before each absolute value $|\lambda_i|$ is estimated term by term when estimating the functional $T(\lambda)$ over the bounded parameter $\lambda \in \Lambda_n(W)$. Let $|\lambda_i^{'}| = \frac{1}{W}\lambda_i$ then $|\lambda_i^{'}| \leq 1$ for $i = 1, \ldots, n$ and

$$||\lambda_i^{'}| - P_K^*(\lambda_i^{'})| \leq \frac{\beta_*}{2K}(1+o(1)), \qquad \forall |\lambda_i^{'}| \leq 1 \tag{4.34}$$

Hence,

$$||\lambda_i^{'}| - \tilde{P}_K^*(\lambda_i^{'})| \leq \frac{\beta_* W}{2K}(1+o(1)), \qquad \forall |\lambda_i^{'}| \leq W \tag{4.35}$$

where $\tilde{P}^{*}_{K}(x) = \sum\limits_{k=0}^{K} \tilde{p}^{*}_{2k} x^{2k}$ with $\tilde{p}^{*}_{2k} = \tilde{p}^{*}_{2k} x^{2k} W^{-2k+1}$ and β_* is a Bernstein constant defined as

$$\beta_* = \lim_{k \to \infty} 2k\delta_{2k} f \tag{4.36}$$

Varga and Carpender (1987), computed $\beta_* = 0.2801694990....$

Taking $\mathbb{E}H_k(x_i) = \lambda_i^k$, $\frac{1}{n}\sum\limits_{i=1}^{n} \lambda_i^{2k} \equiv a_{2k}(\lambda)$ can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} H_{2k}(x_i) = \bar{A}_{2k} \tag{4.37}$$

and define the estimator, $\widehat{T_K(\lambda;W)}$ of $T(\lambda)$ by

$$\begin{aligned} \widehat{T_K(\lambda;W)} &= \sum_{k=0}^{K} \tilde{p}^{*}_{2k} \bar{A}_{2k} \\ &= \sum_{k=0}^{K} p^{*}_{2k} W^{-2k+1} \bar{A}_{2k} \end{aligned} \tag{4.38}$$

The choice of K, the cutoff value determines the performance of the estimator. Cai and Low, (2011) chose

$$K = \frac{\log n}{2 \log \log n} \tag{4.39}$$

where n is the sample size. If the cutoff value is K_*, then

$$K = K_* = \frac{\log n}{2 \log \log n} \tag{4.40}$$

and the final estimator of $T(\lambda)$ defined by

$$\widehat{T_*(\lambda)} = T_{K_*}\widehat{(\lambda;W)} = \sum_{k=0}^{K_*} \tilde{p}^{*}_{2k} \bar{A}_{2k} \tag{4.41}$$

The cutoff is really affected when the sample size is small. Properties of the Developed Estimator

4.5 Asymptotic Properties of the Developed Estimator

The properties of the developed estimator were considered in this section. The distribution of the estimator and its related statistics assuming that the sample size is adequately large were determined using the asymptotic theory. The assumptions were made based on the sample generated by the stochastic procedure. The asymptotic properties considered were: variance and bias.

4.5.1 The Variance and Bias

The variance and bias are important functions associated with the performance of any estimator. These functions are incorporated in the MSE. The MSE of the developed estimator is

$$\begin{aligned} MSE(\widehat{T_*(\lambda)}) &= MSE(\widehat{T_{K_*}(\lambda;W)}) \\ &= \mathbb{E}[\widehat{T_{K_*}(\lambda;W)} - T(\lambda)] \\ &= \mathbb{E}[(\widehat{T_{K_*}(\lambda;W)} - \mathbb{E}T(\lambda)) + \mathbb{E}(\widehat{T_{K_*}(\lambda;W)} - T(\lambda))]^2 \\ &= Var\left(\widehat{T_{K_*}(\lambda;W)}\right) + \left(\mathbb{E}\left(\widehat{T_{K_*}(\lambda;W)} - T(\lambda)\right)\right)^2 \\ &= Var(\widehat{T_{K_*}(\lambda;W)}) + [bias(\widehat{T_{K_*}(\lambda;W)})]^2 \qquad (4.42) \end{aligned}$$

The developed estimator achieves the MiniMax lower bound in equation (2.3) using the MSE of the estimator defined in equation (3.39). Thus,

$$\widehat{T_{K_*}(\lambda;W)} = \sum_{k=0}^{K_*} \tilde{p}_{2k}^* \bar{A}_{2k} \qquad (4.43)$$

Finding the expectation of both sides we obtain:

$$\mathbb{E}[\widehat{T_{K_*}(\lambda;W)}] = \mathbb{E}[\sum_{k=0}^{K_*} \tilde{p}_{2k}^* \bar{A}_{2k}] = \sum_{k=0}^{K_*} \tilde{p}_{2k}^* \mathbb{E}\bar{A}_{2k} \qquad (4.44)$$

But $\mathbb{E}\bar{A}_{2k} = a_{2k}(\lambda)$ for $k \geq 0$ and hence

$$\mathbb{E}[\widehat{T_K(\lambda;W)}] = \sum_{k=0}^{K} \tilde{p}_{2k}^{*} a_{2k}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \tilde{P}_K^{*}(\lambda_i) \tag{4.45}$$

The bias of the estimator, $\widehat{T_K(\lambda;W)}$, for any $\lambda \in \Lambda_n(W)$ was bounded as follows,

$$|\mathbb{E}(\widehat{T_K(\lambda;W)} - T(\lambda))| = |\frac{1}{n}\sum_{i=1}^{n} \tilde{P}_K^{*}(\lambda_i) - \frac{1}{n}\sum_{i=1}^{n} |\lambda_i|| \leq \frac{1}{n}|\tilde{P}_K^{*}(\lambda_i) - |\lambda_i|| \leq \frac{\beta_* W}{2K}(1+o(1)) \tag{4.46}$$

To find the variance of the estimator $\widehat{T_K(\lambda;W)}$, let the variance of a random variable X be defined as

$$Var(X) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x), \tag{4.47}$$

and for any random variables X_i, $i = 1, \ldots, n$

$$\mathbb{E}(\sum_{i=1}^{n} X_i)^2 \leq (\sum_{i=1}^{n}(\mathbb{E}(X_i^2)^{1/2})^2 \tag{4.48}$$

Let $X \sim N(\mu, 1)$, then $(H_K(X)) = \mu^k$ and let $X = \mu + z$ with $z \sim N(0,1)$. Then the expectation, $\mathbb{E}(H_k^2(z)) = k!$, $\mathbb{E}(H_i(z), H_j(z)) = 0$, for all $i \neq j$ and

$$H_k(\mu + z) = \sum_{j=0}^{k} \binom{k}{j} \mu^j H_{k-j}(z) \tag{4.49}$$

and

$$\mathbb{E}(H_k(\mu + z)) = \sum_{j=0}^{k} \binom{k}{j} \mu^j \mathbb{E}(H_{k-j}(z)) \tag{4.50}$$

Hence,

$$\begin{aligned} \mathbb{E}(H_k^2(X)) &= \mathbb{E}(H_k^2(\mu + z)) \\ &= \sum_{i=0}^{k}\sum_{j=0}^{k} \binom{k}{i}\binom{k}{j} \mu^{i+j} \mathbb{E}(H_{k-i}H_{k-j}(z)) \\ &= \sum_{j=0}^{k} \binom{k}{j}^2 \mu^{2j}(k-j)! \\ &= k!\sum_{j=0}^{k} \binom{k}{j}^2 \mu^{2j}\frac{1}{j!} \end{aligned} \tag{4.51}$$

Note that $k!/j! \leq k^{k-j}$ and hence by the Binomial theorem,

$$\mathbb{E}(H_k^2(X)) = k! \sum_{j=0}^{k} \binom{k}{j} \mu^{2j} \frac{1}{j!} \leq k^k \sum_{j=0}^{k} \binom{k}{j} \left(\frac{\mu^2}{k}\right)^j = k^k \left(1 + \frac{\mu^2}{k}\right)^k \leq e^{\mu^2} k^k \tag{4.52}$$

If $|\mu| \leq W$ and $W^2 \geq k$, for all $0 \leq j \leq k$, $\mu^{2j}\frac{1}{j!} \leq W^{2j}\frac{1}{j!} \leq W^{2k}\frac{1}{k!}$. Hence,

$$\mathbb{E}(H_k^2(X)) = k! \sum_{j=0}^{k} \binom{k}{j} \mu^{2j} \frac{1}{j!} \leq k^k \sum_{j=0}^{k} \binom{k}{j} W^{2k} \frac{1}{k!} = (2W^2)^k \tag{4.53}$$

Therefore,

$$\begin{aligned} Var(\bar{A}_{2k}) &= \frac{1}{n^2} \sum_{i=1}^{n} Var(H_{2k}(x_i)) \\ &= \frac{1}{n} Var(H_{2k}(x_i)) \\ &\leq n^{-1} e^{W^2} (2k)^{2k} \end{aligned} \tag{4.54}$$

Thus

$$\begin{aligned} Var(\widehat{T_K(\lambda; W)}) &\leq \left\{ \sum_{k=0}^{K} |\tilde{p}_{2k}^*| Var^{\frac{1}{2}} \bar{A}_{2k} \right\}^2 \\ &\leq \left\{ \sum_{k=0}^{K} |\tilde{p}_{2k}^*| W^{-2k+1} e^{\frac{W^2}{2}} (2k)^k n^{-2} \right\}^2 \\ &\leq \left\{ \sum_{k=0}^{K} |\tilde{p}_{2k}^*| W^{-2k+1} \right\}^2 e^{M^2} (2k)^{2k} n^{-1} \end{aligned} \tag{4.55}$$

Hence, the MSE of $\widehat{T_K(\lambda; W)}$ is bounded by

$$\mathbb{E}\left(\widehat{T_K(\lambda; W)} - T(\lambda)\right)^2 \leq \frac{(\beta_*)^2 W^2}{(2K)^2}(1 + o(1)) + \left\{ \sum_{k=0}^{K} |\tilde{p}_{2k}^*| W^{-2k+1} \right\}^2 e^{W^2} (2k)^{2k} n^{-1} \tag{4.56}$$

Set the cutoff as in equation (4.40). In equation (4.56), the second term is negligible in relation to the first term. The MSE for all $\lambda \in \Theta_n(W)$ is defined as

$$\mathbb{E}\left(\widehat{T_K(\lambda; W)} - T(\lambda)\right)^2 \leq \beta_*^2 W^2 \left(\frac{\log\log n}{\log n}\right)^2 (1 + o(1)) \tag{4.57}$$

CHAPTER FIVE

EMPIRICAL STUDY

5.1 Introduction

In this section, the mean functions in Opsomer et al. (2001) were used for the empirical study. Table 5.1 presents the six mean functions that were used for simulations in this section. The comparison between the developed estimator and the non-parametric techniques that exist in literature i.e. the Nadaraya-Watson estimator, $\hat{T}_{NW}$ and the Modified Nadaraya-Watson estimator, $\hat{T}_{MNW}$ was also done in this section. The data were simulated using R codes to compare the developed estimator, standard Nadaraya-Watson estimator and the modified Nadaraya-Watson estimator.

5.2 The Mean functions simulated

Table 5.1: Mean functions simulated

Mean function	**Equation**
Linear	$Y_1 = 1 + 2(x - 0.5)$
Quadratic	$Y_2 = 1 + 2(x - 0.5)^2$
Jump	$Y_3 = 1 + 2(x - 0.5)I_{x \leq 0.65} + 0.65I_{x>0.65}$
Bump	$Y_4 = 1 + 2(x - 0.5) + exp(-200(x - 0.5)^2)$
Sine	$Y_5 = 2 + sin(2\pi x)$
Exponential	$Y_6 = exp(-8x)$

The mean functions in Table 5.1 were chosen because they are often applicable to real life situations. For instance, time between two successive breakdowns of a machine after repair constitute an exponential distribution. The Bumps and the jumps are used in events with a high rate of occurrence such as disease outbreaks, floods or rainfall at a given region within a certain period, while the sine distributions are used in situations whose occurrences are periodic.

The first data set was obtained through simulation by use of a linear model with the relation

$$Y_1 = 1 + 2(x - 0.5) + e_i \tag{5.1}$$

The random variable X was simulated using a rectangular distribution that takes the values that are equally likely from 0 to 1 inclusive. It is assumed that $(x_i, y_i), \quad i = 1, \ldots, N$, are independent and identically distributed random variables. The error term e_i is a standard normal variable defined as $e_i \sim N(0, 1)$.

The second data set was obtained through simulation by use of a quadratic model which has the relation

$$Y_2 = 1 + 2(x - 0.5)^2 + e_i, \qquad i = 1, \ldots, N \tag{5.2}$$

The random variable X was simulated using a rectangular distribution that takes the values that are equally likely from 0 to 1 inclusive. It was also assumed that $(x_i, y_i), \quad i = 1, \ldots, N$, were independent and identically distributed random variables. The error term was a standard normal variable defined as $e_i \sim N(0, 1)$. This was done for all the mean functions in Table 5.1.

In all the mean functions, $Y_1, \ldots, Y_6$, a population of size $N = 1000$ was simulated using the R code. Five hundred samples of size $n = 250$ were generated using simple random sampling without replacement. In each selected sample, the estimate of the population total, the estimate of the mean squared error, the bias and confidence interval lengths were computed. The biases of the population totals were obtained using the relation

$$\left(\left[\sum_{i=1}^{2000}\frac{\hat{T}_i}{2000}\right]-T\right)/T \tag{5.3}$$

where T is the actual population total and T_i is one of the estimators of the population total from the i^{th} sample.

Table 5.2: Summary results for the bias

Mean function	$\widehat{T_K(\lambda;W)}$	$\hat{T}_{NW}$	$\hat{T}_{MNW}$
Linear	-0.1680	-0.4075	-0.2045
Quadratic	0.0300	0.0700	-0.1773
Jump	-0.2537	-0.3034	-0.1583
Bump	-0.1876	-0.5189	-0.2599
Sine	-0.2910	-1.1709	-0.5999
Exponential	-0.0123	0.5838	0.2901

In Table 5.2, a summary of the results of the bias simulated from the mean functions in Table 5.1 are presented. The negative and positive value imply underestimation and overestimation respectively. The developed estimator had smaller values compared to $\hat{T}_{NW}$ and $\hat{T}_{MNW}$. For the exponential mean function, the $\hat{T}_{NW}$ and $\hat{T}_{MNW}$ overestimates the population mean. The developed estimator only overestimates the population mean for the quadratic mean function. The modified Nadaraya-Watson underestimates the population mean in all the mean functions considered except in the exponential mean function.

Table 5.3: Summary results for the Mean squared errors

Mean function	$\widehat{T_K(\lambda;W)}$	$\hat{T}_{NW}$	$\hat{T}_{MNW}$
Linear	0.0490	0.1902	0.0478
Quadratic	0.0440	0.5621	0.1402
Jump	0.0879	0.9458	0.2384
Bump	0.0623	0.3209	0.0805
Sine	0.1370	1.4103	0.3698
Exponential	0.0031	1.1674	0.2908

Table 5.3 presents a summary of the mean squared error values. The $\hat{T}_{NW}$ estimator had the largest MSE values while the developed estimator had the smallest

MSE values. Thus, the developed estimator, $\widehat{T_K(\lambda;W)}$ was better than $\hat{T}_{NW}$ and $\hat{T}_{MNW}$ estimators.

The confidence intervals were normally constructed around the point estimators and obtained at 95% confidence level. The 95% confidence intervals for each of the estimators were computed using the formula

$$T = \hat{T} \pm z_{\alpha/2}\sqrt{var(\hat{T})} \tag{5.4}$$

and the confidence interval length obtained by subtracting the lower limit from the upper limit.

In Table 5.4, the developed estimator had shorter confidence interval lengths than the $\hat{T}_{NW}$ and $\hat{T}_{MNW}$ estimators. Shorter confidence interval lengths imply that the developed estimator was equal to the true parameter.

Table 5.4: Summary results for the confidence interval lengths

Mean function	$\widehat{T_K(\lambda;W)}$	$\hat{T}_{NW}$	$\hat{T}_{MNW}$
Linear	0.9270	1.7090	0.8573
Quadratic	0.7770	1.7090	1.4676
Jump	1.0944	3.8124	1.9138
Bump	0.9748	2.2051	1.9138
Sine	1.5270	4.890	2.5452
Exponential	0.2195	4.235	2.114

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

In this chapter, the summary of main results that lead to the conclusions and recommendations was covered. The problem specifically considered the estimation of an arbitrary non-smooth in the nonparametric set-up. The study set out to derive a general MiniMax lower bound, develop the MiniMax lower bound of estimating the non-smooth functional $T(\lambda)$ from the general MiniMax lower bound, develop an estimator and derive the asymptotic properties of the developed estimator: vis, bias, variance and normality.

6.2 Summary of Main Results

Statistical researchers have made considerable effort to derive MiniMax lower bounds, upper bounds and the optimal rate of convergence in the development of MiniMax theories in the nonparametric function estimation. Specifically, when working in the context of MiniMax estimation, the lower bounds are the most important. In the preceding applications the bounds used have been given in simpler cases and the optimal rates of convergence for estimating smooth functionals are often parametric rates.

The basis for developing of the MiniMax lower bound was formed by deriving the general MiniMax lower bound which is shown by equation (4.16). The general MiniMax lower bound was derived by dividing the parametric space Λ into two disjoint subsets Λ_0 and Λ_1 where $H_0 : \lambda \in \Lambda_0$ against $H_1 : \lambda \in \Lambda_1$ are the null and alternative hypotheses respectively. The two priors ω_0 and ω_1 with a large difference in the expected values of the functional were constructed while making the Chi-square distance between two normal mixtures small. The MiniMax risk for the developed estimator was given in equation (4.28), and the asymptotic properties: bias and variance were derived.

The developed estimator, attained the MiniMax lower bound in equation (2.3). It was also shown that the estimator had smaller bias and MSE values than the standard Nadaraya-Watson estimator and the modified Nadaraya-Watson estimator. Additionally, the confidence interval lengths of the developed estimator were shorter than confidence interval lengths of the standard Nadaraya-Watson estimator and the modified Nadaraya-Watson estimator which shows that the developed estimator is better.

6.3 Conclusions

Our perception drove us to the end that estimating non-smooth functionals display a few highlights that are altogether not the same as those in estimating smooth functionals. The absence of these properties features the motivation behind why standard methods fail to give sharp outcomes. Hence, the best polynomial approximation and the Hermite polynomial were utilized in the determination of lower bounds and development of an estimator. These orthogonal polynomials were used to develop the estimator in the NP set-up.

The problem of estimating the non-smooth functional $T(\lambda)$ was special, and the standard procedures for deriving the lower bounds in the problems of estimating the value of a functional seemingly do not work. The functional $\tilde{T}(\lambda)$ is "nearly smooth"- it looses smoothness at the origin. The parametric convergence rate was used to estimate the value of an "actually smooth" functional, while the nonparametric method was used to estimate the value of a non-smooth functional.

The nonparametric statistical approach offered an alternative set of statistical techniques that did not require any or only limited assumptions about the data. The hard-to solve problem was changed into an efficiently solvable MiniMax problem. This method of transfomation achieved a flexible distribution and gave a better estimator. A pair of probability distributions, each concentrated at its own small "r-sphere", were constructed in such a way that the chi-square distance between them was small.

When the sample size is finite, the properties of the estimator are similar to those when the sample size is arbitrarily large. Asymptotics are used to describe properties of the estimator when the sample sizes are arbitrarily large. Approximation to the distribution of the estimator and its associated statistics when the sample is assumed to be sufficiently large is derived by the asymptotic theory whose main tools are consistency and asymptotic normality.

The asymptotic MiniMax Risk was obtained by optimally tuning the polynomial approximation. The idea of the existence of the best rate of convergence was formalized by calculating the MiniMax Risk of the estimator. The rate was based on the requirement that the estimator performs well at a fixed model as well as at a sequence of models that lie nearby. The risk is additionally utilized as a benchmark for assessing the performance the estimation method.

6.4 Recommendations

The methodology and results developed in this thesis can be used to solve other related problems. For instance, the methodology can be used to estimate other non-smooth functionals such as excess mass. When estimating excess mass, the local scales are used rather than the global scales. This is possible since concentration of measure at a certain point allows estimation to be done locally. See for example (Wolfgang, 1994) and (Das Gupta, 2008).

Estimating excess mass is a general approach to statistical analysis that can be used in many practical applications. When determining the support of a function; a point where a function exceeds a certain level, exhibits discontinuity, or changes point is involved, the region of interest in the problem can be estimated on a finite number of observations.

The differences of excess-masses at different levels λ, where λ is a real number, are used in order to test the multimodality of a probability distribution in most applications. The dip-excess mass was introduced by (Hartigan and Hartigan, 1985) who proposed an estimator that could be used to test multimodality. In the literature, the dip-excess mass estimator was widely used. See for example, (Mullar and Sawitzki, 1991) and, (Fisher and Marron, 2001). They maintain on the fact that a procedure like this separates mode estimation from its location.

The estimation density level sets (or density contour clustering), which is the set of points $C(\lambda)$ on which the excess mass at level λ is calculated is another major use of the excess mass functional. This necessitates a good excess mass estimator as well as an optimization process. Polonik, (1995) proved consistency of such estimators of the density level set and found some rates of convergence. The author expected that the underlying distribution has density contour clusters lying in the class $\mathbb{C}$ under consideration.

Tsybakov (1997) proposed minimax rates for estimating smooth star-shaped level sets of a density. The level set estimation problem deals with reconstructing an unknown set $G(\lambda) = \{f \geq \lambda\}$ from a random sample of points $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ of a random variable X, where f denotes the density of X and λ is a positive threshold. Tsybakov's (1997) approaches were either difficult to implement or required assumptions which were difficult to check. They used a margin assumption quantifying the smoothness of the density p around the level λ as introduced by (Mammen and Tsybakov, 1999). Later, (Gayraud and Rousseau, 2005) used a Bayesian approach and (Rigollet and Vert, 2006) revisited the plug-in estimator. They may claim for computational feasibility as well as for strong theoretical properties. Klemela (2004), on the other hand studied and implemented a complexity penalized excess-mass criterion-based estimator of density support.

Excess mass estimation has also been used to estimate regression contour clusters by (Polonik and Wang, 2005), discrimination of locally stationary time series by (Chandler and Polonik, 2006) and anomaly detection and classification as described by (Rigollet and Vert, 2006) using level set estimation. The use of a nonparametric estimator of f was generally avoided in these methods. In practise, such an estimator may not be very attractive in higher dimensions d. In reality such an estimator is not appealing in higher dimensions d. Estimating the excess mass as an integrated functional of f at a fixed level say $\lambda > 0$

At any level λ, the excess mass is the total of contributions coming from the connectivity components $C_i(\lambda) \leq R^K$ of $\{f \geq \lambda\}$. The connectivity components $C_i(\lambda)$ of $\{f \geq \lambda\}$ are called λ clusters. These clusters are described as sets maximizing the distribution function.

REFERENCES

Bergstrom, R. (1985). The estimation of nonparametric functions in a Hilbert space. *Economic Theory,1*, 7-26.

Birǵe, L. and Massart, P. (1995). Estimation of the integral functionals of a density. *Ann. Statist.,23*, 11–29.

Bickel, P. J. & Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order convergence estimates. *Sankya Ser.*, *A 50*, 381-393.

Billingsley, P. (1995). *Probability and Measure.* (3rd ed.) New York: John Wiley and Sons.

Breidt, F. J. & Opsomer, J. D. (2000). Local Polynomial Regression Estimators in Survey Sampling. *The Annals of Statistics*, *28* (4), 1026-1053.

Brown, L. D. & Low, M. G. (1996). A constrained Risk Inequality with Applications to Nonparametric Functional Estimation. *Ann. Statist.*, *24*, 2524-2535.

Cai, T.T. & Low, M.G. (2011). Testing Composite, Hermite Polynomials, and Optimal Estimation of a non-smooth Functional. *Ann. statist.*, *39* (2), 1012-1041.

Casella, G. & Berger, R. L. (2002). *Statistical Inference.* (2nd ed.) New York: Wadworth Group.

Chandler, G. & Polonik, W. (2006). Discrimination of locally stationary time series based on the excess mass functional. *J. Amer. Statist. Assoc.*, *101*, 240–253.

Chichilnisky, G. (2009). The limits of Econometrics: Nonparametric functions in Hilbert spaces. *Economic Theory*, *25*, 1-17.

Cline, D. B. & Hart, J. D. (1991). Kernel Estimation of Densities of Discontinuous Derivatives. *Statistics*, *22* (1), 1-17.

Comminges, L. & Dalalyan, A. S. (2013). Minimax Testing of a Composite Null Hypothesis Defined via a Quadratic Functional in the Model of Regression. *Electronic Journal of Statistics*. *7*, 146-190.

Cowling, A. & Hall, P. (1996). On Pseudodata Methods for Removing Boundary Effects in Kernel Density Estimation. *Journal of the Royal Statistical Society ser.*B, 551-563.

Daniels, H. E. (1954). Saddlepoint Approximations in Statistics. *Ann. Math Statist.*, *25* (4), 631-650.

Daniels, H. E. (1987). Tail probability Approximations. *International Statistical Review.*, *55*, 37-48.

Das Gupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* New York: Springer.

Dekking, F. M., Kraaikamp, C., Lopuhaa, H. P., & Meester, L. E. (2005) *A Modern Introduction to Probability and Statistics.* New York: Springer-Verlag.

DiNardo, J. & Tobias, J. L. (2001). Nonparametric Density and Regression Estimation. *Journal of Economic Perspectives*, *15* (4), 11-28.

Dorfman, A. H. (1992). Nonparametric Regression for Estimating Totals in Finite Populations. In proceedings of the section on Survey Research Methods. *American Statistics Association*, *25*, 1-17.

Donoho, D. L. & Liu, R. C. (1991). Geometrizing Rates of Convergence II. *Ann. Statist*, 622-625.

Douglas, C. M. & George, C. R. (2008). *Applied Statistics and Probability for Engineers*. (3rd ed.) Delhi: Pashupati printers P. ltd.

Fisher, N. I. & Marron, J. S. (2001). Mode Testing via Excess Mass Estimate. *Biometrika. 88* (2), 499-517.

Flajolet, P. & Sedgewick, R. (2009). *Analytic Combinatorics.* Cambridge:Cambridge University Press.

Fox, L. & Parker, I. B. (1968). *Chebyshev Polynomials in Numerical Analysis.* Oxford: Oxford University Press.

Gatto, R. (2010). A Saddlepoint Approximation to the Distribution of Inhomogeneous Discounted Compound Poisson Processes. *Methodology and Computing in Applied Probability*, *12* (3), 533-551.

Gayraud, G. & Rousseau, J. (2005). Rates of convergence for a Bayesian level set estimation. *Scand. J. Statist.*, *32*, 639–660.

Hartigan, J. A. & Hartigan, P. M. (1985). The Dip Test of Unimodality. *Ann. Statist.*, *13* (1), 70-84.

Ibragimov, I. A. & Khasminski, R. (1991). Asymptotic Normal Families of Distributions and Effective Estimation Ann. Statist., *19*, 1681-1724.

Keisuke, H. & Jack, R. P. (2012). Impossibility Results for Nondifferentiable Functionals. *Econometrica, Economic Society.*, *80* (4), 1769-1790.

Juditsky, A. & Nemirovski, A. (2020).*Statistical Inference via Convex Optimization.* Princeton Series in Applied Mathematics, Princeton University Press.

Klemela, J. (2004). Complexity penalized support estimation. *J. Multivariate Anal. 88* (2), 274–297.

Korostelev, A. P. (1990). On the accuracy of estimation of non-smooth functionals of regression. *Theory Probab. Appl.*, *35*,768-770.

Korostelev, A. P. & Tsybakov, A. B. (1994). *Minimax Theory of Image Reconstruction.* New York:Springer-Verlag.

Lepski, O., Nemirovski, A. & Spokoiny, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory Relat. Fields*, *113*, 221-253.

Le Cam, L. (1973). Convergence of Estimates under Dimensionality Restrictions. *Ann.Statist.*, *1*, 38-53.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*, New York:Springer-Verlag.

Lugannani, R. & Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*, *12*, 475-490.

Mammen, E., J. S. Marron, & N. I. Fisher (1992). Some Asymptotics for Multimodality Tests Based on Kernel Density Estimates. *Probability Theory and Related Field*, *91*, 115-132.

Mammen, E. & Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.*, *27*, 1808–1829.

Marron, J. S. & Ruppert, D. (1994). Transformations to Reduce Boundary Bias in Kernel Density Estimation. *Journal of the Royal statistical Society*. Ser. B, *56*, 653-671.

Mullar, D. W. & Sawitzki, G. (1991). Excess mass estimates and tests of multimodality. *J, Amer.Statist.Assoc.*, *86*, 738-746.

Pagan, A. & Ullah, A. (1999). *Nonparametric Economics*, Cambridge:Cambridge University Press.

Petrova, S. S. & Solov'ev, A. D. (1997). The Origin of the Method of Steepest Descent. *Historia Mathematica*, *24*, 361-375.

Polonik, W. (1995). Measuring Mass Concentrations and estimating Density Contour Clusters-an excess mass approach. *Ann.Statist.*, *23* (3), 855-881.

Polonik, W. & Wang, Z. (2005). Estimation of Regression Contour Clusters-an application of the excess mass approach to regression. *Multivariate Anal.*, *94* (2), 227-249.

Ramachanran, K. M. & Chris, P. T. (2009). *Mathematical Statistics with Applications*. California: Elsevier Academic Press.

Rigollet, P. and Vert, R. (2006). Estimation of Regression Contour Clusters-an Application of Excess Mass Approach to Regression. *Multivariate Anal.*, *94* (2), 227-245.

Rivlin, T. J. (1974). *The Chebyshev Polynomials*. Second edition. New York: Wiley-Interscience.

Rivlin, T. J. (1990). *Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory.* (2nd ed.) New York: John Wiley and Sons.

Rockafellar, T. (1994). Mathematical Programming: State of the Art. (J. R. Birge and K.G. Murty, editors), University of Michigan Press, *Ann Arbor*, 248-248.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Spady, R. H. (1991). Saddlepoint Approximations for Regression Models. *Biometrika*, *78*(4), 879- 889.

Sauer, T. (2006). *Numerical Analysis.* New Yolk:Pearson Education Inc.(C).

Tsybakov, A. B. (1997). On parametric estimation of density level sets. *Ann.Statist.*, *25*(3), 948-969.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* New York:Springer-Verlag.

Varga, R. S. & Carpender, A. J.(1987). On a Conjecture of S. Bernstein in Approximation Theory. *Math. USSR Sbornik*, *57*, 547-560.

Wand, M. P., Marron, J.S. & Ruppert, D. (1991). Transformations in Density Estimation (with discussion). *Journal of the American Statistical Association*, *86*(414), 343-361.

Wang, L., Brown, L. D., Cai, T. & Levine, M. (2008). Effect of Mean and Variance Function Estimation in Nonparametric Regression. *Ann. Statist. 36*, 646-664.

Wasserman, L. (2006). *All Nonparametric Statistics.* New Yolk:Springer Science.

Wolfgang, H. & Schimile, M. G. (1996). *Statistical Theory and Computational Aspects of Smoothing*, New Yolk:Springer-Verlag.

Yu-Xiang, W., Jing, L. & Stephen, E. (2016). A Minimax Theory for Adaptive Data Analysis. *Journal of Machine Learning Research.*, *17*, 1-40.

Appendix I: Polynomial Approximation The Chebyshev polynomial T_{2m} can be written altenatively as

$$T_{2m}(x) = \sum_{l=0}^{m} [(-1)^{m-1} \sum_{j=m-l}^{m} \binom{2m}{2j} \binom{j}{m-1}] x^{2l}$$

Write $T_{2m}(x) = \sum_{l=0}^{m} t_{2l} x^{2l}$. Then

$$|t_{2l}| = \sum_{j=m-l}^{m} \binom{2m}{2j} \binom{j}{m-1} \leq \sum_{j=m-1}^{m} \binom{2m}{2j} \binom{m}{m-1} \leq 2^{2m} 2^{2m} = 2^{3m}$$

The coefficient for x^{2k} in the polynomial $Y_k(x)$ is bounded from above by

$$|y_{2k}| \leq \frac{4}{\pi} \sum_{j=k}^{K} \frac{2^{3j}}{4j^2 - 1} \leq 2^{3k}$$

.

Appendix II: MATLAB-codes and R-codes for various Graphs

MATLAB code for Fig.3.1

```
x = linspace(−1, 1, 201);
T1 = cos(a cos(x));
T2 = cos(2a cos(x));
T3 = cos(3a cos(x));
T4 = cos(4a cos(x));
T5 = cos(5a cos(x));
subplot(1,1,1)
plot(x,T1,‘b’)
hold on
plot(x,T2,‘r’)
plot(x,T3,‘g’)
plot(x,T4,‘c’)
plot(x,T5,‘y’)
```

MATLAB code for Fig.3.2

```
x = [−1 : 0.2 : 1];
y = [1./(√(1 − x^2))];
plot(x,y)
```

R code for Fig.2.1

```
par (mfrow = c(3, 2))
n = 5
p = .05
x = 0 : 5
plot (x, dbinom(x, n, p), ylab = ”p(x)”, main = ”n = 5, p = .05”)
```

...

MATLAB code for Fig.3.3 (a)

$x = chebyfun('x');$

$f = x.^2.sin(10 * x);$

```
subplot(1,1,1)

plot(f)

hold on

p=chebfun(f,1):hold on, plot(p,'r')

plot(f)

clear
```

MATLAB code for Fig.3.3 (b)

$x = chebyfun('x');$

$f = x.^2.sin(10 * x);$

```
subplot(1,1,1)

plot(f)

hold on

p=chebfun(f,4):hold on, plot(p,'r')

plot(f)

clear
```

MATLAB code for Fig.3.3 (c)

$x = chebyfun('x');$

$f = x.^2.sin(10 * x);$

```
subplot(1,1,1)

plot(f)

hold on

p=chebfun(f,9):hold on, plot(p,'r')
```

```
plot(f)

clear
```

MATLAB code for Fig.3.3 (d)

$x = chebyfun('x');$

$f = x.^2.sin(10 * x);$

```
subplot(1,1,1)

plot(f)

hold on

p=chebfun(f,16):hold on, plot(p,'r')

plot(f)

clear
```

R-codes for tabulated results

```
library(MASS)

require(sm)
```

`e=rnorm(1000,mean=0,sd=1)` **#random error**

`X=runif(1000,min=0,max=1)` **#explanatory variable**

$Y_1 = 1 + 2 * (X - .5)$

$Y_2 = 1 + 2 * (X - .5)^2$

$I_x = function(x, h)1 * (x \geq h)$

$Y_3 = 1 + 2 * (X - .5) * I_x(X, .65) * (1 - I_x(X, .65))$

$Y_4 = 1 + 2 * (X - .5) + \exp(-200 * (X - .5)^2)$

$Y_5 = 2 + \sin(2 * \pi * X)$

$Y_6 = \exp(-8 * X)$

$Y = Y_i + e$ **#regression function**

$mf = mf_2 = TT = 0$

$j = 0$

$BIASK = BIASNW = BIASMNW = TOTALS = 0$

$MSEK = 0$

$MSENW = MSEMNW = 0$

$TTK = TTNW = TTMNW = 0$

$means = 0$

$varK = varNW = varMNW = numeric()$

`while` $j \leq 10000$

$sindex = sample(1 : 1000, 500)$ # **selecting the sample**

$x.sample = X[sindex]$

$xreflect = c(x.sample, -x.sample)$

`xreflect`

$xnosample = setdiff(X, x.sample)$

$y.sampe = Y[sindex]$

$yreflect = c(y.sample, y.sample)$

`yreflect`

$ynosample = setdiff(Y, y.sample)$

$data1 = data.frame(xnosample)\ \#H = hcv(xnosample, ynosample)$

$H3 = (ucv(xreflect, nb = 1000, min(x.sample), max(xnosample)))$

$H1 = (ucv(x.sample, nb = 1000, min(x.sample), max(xnosample)))$

$\#H = (ucv(x.sample, nb = 1000, min(x.sample), max(xnosample))) * 100$

$\#H2 = (ucv(xnosample, nb = 1000, min(xnosample), max(xnosample)))$

$\#H2 = (ucv(xreflect, nb = 1000, min(xreflect), max(xrefrect), tot = 0.01)) * 100$

$\#H2 = (ucv(xreflect, nb = 1000)) * 100$

$nad1 = ksmooth(x.sample, y.sample, kernel = {}”normal”, bandwidth = 0.232, x.points = xnosample)$

$nad2 = ksmooth(xreflect, yreflect, kernel = {}”normal”, bandwidth = 0.03421, x.points = xnosample)$

$nad3 = loess(y.sample \ln x.sample, span = .5)$

$\#nad4 = locpoly(x.sample, y.sample, bandwidth = .25)$

$\#nad4 = locpoly(x.sample, y.sample, bandwidth = H)$

$model1 = npreg(xdat = x.sample, ydat = y.sample, bws = H, regtype = {}”ll”)$

$pred = predict(model1, newdata = data.frame(x.sample), exdat = xnosample)$

#Finding the variance ratio

$var = 0$

```
for(i in 1:length(y.sample))
uncond=c(mean(BIASK),mean(MSEK),mean(BIASNW),mean(MSENW),
mean(BIASMNW),mean(MSEMNW))
RESULTS=matrix(uncond,1,8)
colnames(RESULTS)=c("BIASK","MSEK","BIASNW","MSENW","BIASMNW","MSEMNW")
RESULTS
M=sum(Y)
V=mean(X)
```