# On Transformation Based Circular Density Estimators

Yuhan Cao

A Thesis
in the Department of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science (Mathematics) at
Concordia University
Montréal, Québec, Canada

May 2018

# CONCORDIA UNIVERSITY

## School of Graduate Studies

This is to certify that the thesis prepared

By: **Yuhan Cao**

Entitled: **On Transformation Based Circular Density Estimators** and submitted in partial fulfillment of the requirements for the degree of

### Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____Examiner

Dr. Arusharka Sen

_____Examiner

Dr. Wei Sun

_____ Thesis Supervisor

Dr. Yogendra P. Chaubey

Approved by _____

Chair of Department or Graduate Program Director

_____

Dean of Faculty

Date _____

# Abstract

Estimation of the probability density function for circular data is an important topic in statistical inference. In this thesis, I would like to introduce two transformation based methods for estimating probability density function in this context. One is derived from traditional kernel density estimator and the other one comes from the Bernstein polynomial estimator (Chaubey, 2017). We know both of the kernel density estimator (Silverman, 1986) and Bernstein polynomial estimator (Babu, Canty and Chaubey, 2002) are appropriate for the case of linear data, transformation of circular data to linear data would bring extreme simplicity to estimation of probability density function in the case of circular data by back transformation. I will conduct a simulation study to compare these methods with respect to their global and local errors. We find through our simulation study that transformed kernel density estimator has a stronger ability to alleviate the boundary problems than transformed Bernstein polynomial estimator, however, their overall performance is pretty much similar in the central part of the distribution. Therefore, in general we can say transformed kernel density estimator leads to a better method as compared to the transformed Bernstein polynomial estimator, however further research may be needed to study other transformations.

# Acknowledgements

I would like to thank my supervisor Dr. Yogendra P. Chaubey. During my graduate school, he gave me advice and suggestion when I feel confused, he helped me to overcome my language problems and academic problems, he supported me all the time and I can't finish my study and this thesis without his guidance. I was really impressed with his knowledge, energy and his serious attitude when it comes to academic problems.

I would like to thank all my instructors: Dr. A. Sen, Dr. D. Sen, Dr. W. Sun, Dr. L. Popovic and Dr. L. Kakinami, for their excellent lectures and their help when I took these lectures. Besides, I would like to thank Ms. Marie-France Leclere, Ms. Judy Thykootathil and Ms. Debbie Arless for supporting me all the time during my study in Canada.

Last but not the least, I would like to express my sincere gratitude to my parents for their unconditional love and support, specially my husband Honghao Zhang for providing his continuous support and company during my tough time. I can't go through anything without you.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Background

Circular statistics (sometimes referred to as directional statistics) is a very important subfield of statistics, where the observed values represent directions. Such data are important in many fields, for instance, a biologist may be interested in the movement of migrating animals while a geologist may be interested in study the circulation of underground water. Circular statistics began to apply in more and more areas (such as biology, geology, meteorology) and is gaining much more attention from statisticians. In circular statistics, circular data is a group of data that occurs around a circle, and we usually measure it from $0^\circ$ to $180^\circ$ in degrees or $0$ to $2\pi$ radians. Obviously, it's different from the linear data and special methods maybe required to handle such data.

As the specialized tools and technology for analyzing circular data are not currently widely used, many statisticians are still looking for a better way to do statistical analysis for circular statistics. As we know, statistical analysis includes data analysis, estimation, modeling, fitting, and so on. Estimation is a very important step and we want to focus on the estimators based on both circular data and linear data. Since the transformation between circular data and linear data exists, we can also find some effective methods to convert a linear density estimation to a circular density estimator, that is the main subject matter of this thesis.

## 1.2 Circular data and some descriptive statistics

Circular data consists observations that occur around a circle and it can be represented as angles or as points on the round surface. If we create an origin and two axes and draw a circle with the origin as the center, we could use the trigonometric functions of $X$ and $Y$ to express the coordinates of the point $P$ on the circle. In this way, we could convert the polar co-ordinates into the rectangular co-ordinates easily. Let $r$ be the distance from $P$ to the origin and $\theta$ be the direction, we could write down the co-ordinates of the point $P$

$$(x = r\cos\theta, y = r\sin\theta),$$

in this thesis the range of $\theta$ is fixed to $[-\pi, \pi)$.

When the distance $r$ equals to 1, the conversion between polar co-ordinates and rectangular co-ordinates is

$$(1, \theta) \rightarrow (\cos\theta, \sin\theta).$$

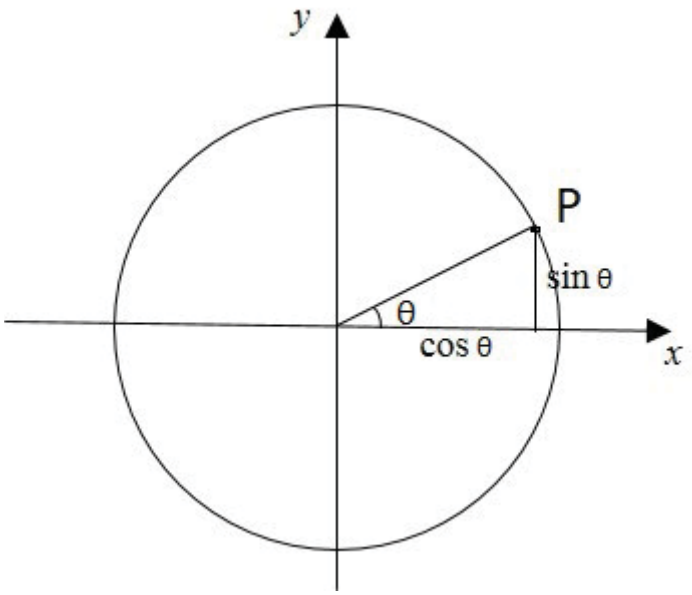
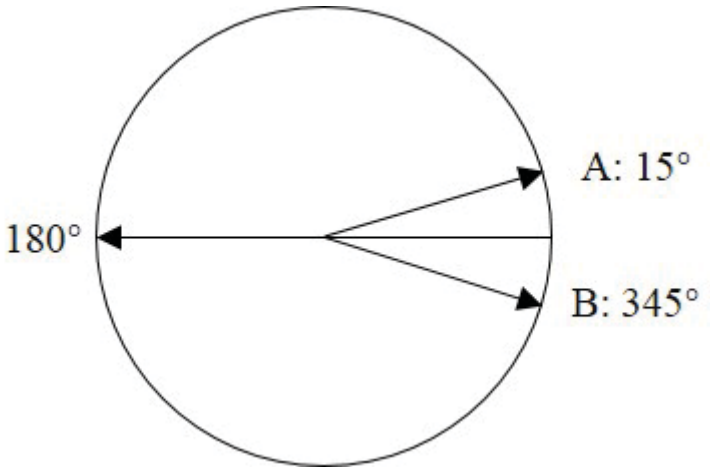



Figure 1.1: Relation between rectangular and polar co-ordinates

After understanding the basic concepts of the circular data, we begin to consider some descriptive statistics of it. The first one we want to know is the circular mean. As we know that the circular data is different from the linear data, we cannot use the normal way to calculate the mean. For example, in the figure below, suppose we have two points $A$ and $B$, which direction are $15°$ and $345°$ respectively (suppose the North be the zero direction and clockwise as the positive sense of rotation). Here we can see, the arithmetic mean is $180°$ and it points due South whereas the true mean should be towards North.




From the example, it is obvious that the arithmetic mean is not suitable for the circular data. Besides, the sample variance, which is related with the sample mean, is also not suitable for the circular data and we should find some alternative measures for center and dispersion if we want to deal with the circular data.

One way to calculate the circular mean is to treat the data composed of unit vectors and use the direction of their resultant vector. Let $\alpha_1$, $\alpha_2$,..., and $\alpha_n$ be a set of the circular observation given in terms of angles. Recall the conversion between polar and rectangular co-ordinates, which is

$$(1, \theta) \rightarrow (x = \cos\theta, y = \sin\theta).$$

We could obtain the resultant vector of these $n$ unit vectors by summing them component-

wise, and get

$$R = (\sum_{i=1}^{n} \cos\theta_i, \sum_{i=1}^{n} \sin\theta_i) = (COS, SIN) \tag{1.1}$$

Let $\overline{R}$ represents the length of the resultant vector $R$, where

$$\overline{R} = ||R|| = \sqrt{COS^2 + SIN^2}. \tag{1.2}$$

Then we could get the direction of $R$, which we also call the circular mean direction, $\overline{\theta_0}$.

There are two facts we know about $\overline{\theta_0}$ is that

$$\cos\overline{\theta_0} = \frac{COS}{R}, \sin\overline{\theta_0} = \frac{SIN}{R}.$$

To find $\overline{\theta_0}$, we can do the following calculation.
Since

$$\tan\overline{\theta_0} = \frac{\sin\overline{\theta_0}}{\cos\overline{\theta_0}},$$

we can get

$$\overline{\theta_0} = \arctan\frac{SIN}{COS}.$$

According to the property of trigonometric function, we can conclude

$$\overline{\theta_0} = \arctan\frac{SIN}{COS} = \begin{cases} \arctan\frac{SIN}{COS}, & \text{if } COS > 0, SIN \geq 0, \\ \frac{\pi}{2}, & \text{if } COS = 0, SIN > 0, \\ \arctan\frac{SIN}{COS} + \pi, & \text{if } COS < 0, \\ \arctan\frac{SIN}{COS} + 2\pi, & \text{if } COS \geq 0, SIN = 0, \\ undefined, & \text{if} COS = SIN = 0. \end{cases} \tag{1.3}$$

(Jammalamadaka and Sengupta, 2001)

## 1.3 Circular probability distribution

A circular probability distribution, literally, is a distribution with its total probability concentrated on a circumference of an unit circle. Every point on this circle has a direction, as we already know, the data is directional. If we measure a circular random variable $\theta$ in radians, then the range of it may be taken as $[0, 2\pi)$ or $[-\pi, \pi)$.

### 1.3.1 Generating circular distribution

To generate a circular distribution, many different models can be used. Here we give a brief introduction for a few such general methods. (Jammalamadaka and Sengupta, 2001)

#### Wrapped Distribution

Literally, wrapped distribution means we could wrap a linear distribution around a unit circle to conduct a circular distribution. Given a linear random variable X on the real line, we could transform it to a circular random variable by reducing its module $2\pi$, e.g. $\theta = X(mod\ 2\pi)$. This operation is nothing but wrap the real line around the circle of unit radius, and accumulate probability over all the overlapping points $x = \theta, \theta \pm 2\pi, \theta \pm 4\pi, ...$, so if we have the linear density function $p(x)$ and the corresponding circular density function $f(\theta)$, it can be described as:

$$f(\theta) = \sum_{m=-\infty}^{\infty} p(\theta + 2\pi m),\ -\pi \leq \theta < \pi.$$

By using this technique, we could conduct both discrete and continuous wrapped distributions.

#### Characterization properties

Generating a circular distribution through some characterize properties such as maximum entropy is the most common method for normal distribution (Jammalamadaka and Sengupta, 2001 and von Mises, 1918). The circular normal distribution has the maximum entropy property and also, its mean direction is estimated with maximum likelihood by the direction of the resultant vector. This characterization goes back to von Mises (1918) which is why we also call circular normal distribution as the von Mises distribution.

#### Offset distributions

By transforming a linear random vector to its directional component, we can obtain the offset distributions (Jammalamadaka and Sengupta, 2001). We transform the bivariate random vector $(X, Y)$ into polar co-ordinates $(r, \theta)$ and integrate over $r$ for a given $\theta$. If the joint distribution of a bivariate distribution is $p(x, y)$, then the corresponding circular offset distribution $f(x, y)$ is given by

$$g(\theta) = \int_0^\infty f(r\cos\theta, r\sin\theta) r dr. \tag{1.4}$$

In the following part, we will describe some standard circular distributions for a general reference (see Fisher, 1995).

### 1.3.2 Some standard circular distribution

#### Uniform circular distribution

**Definition 1.3.1** *If the total probability is spread out uniformly on the circumference of a circle, we call it a Circular Uniform Distribution with the constant density*

$$f(\theta) = \frac{1}{2\pi}, -\pi \leq \theta < \pi.$$

It's not hard to find that, in a uniform circular distribution, all the directions have equal probability and the mean direction is undefined. From the probability function we can

also write the cumulative distribution function $F(\theta)$ as

$$F(\theta) = \frac{\theta}{2\pi}, -\pi \leq \theta < \pi$$

**Wrapped Cauchy distribution**

By wrapping the Cauchy distribution on the real line we can get the wrapped Cauchy distribution. Since we know the density function of a Cauchy distribution is given by:

$$p(x) = (\frac{1}{\pi})\frac{\sigma}{\sigma^2 + (x-\mu)^2}, \quad x \in (-\infty, \infty). \tag{1.5}$$

We could get the probability density function for a wrapped cauchy distribution:

$$f(\theta) = \frac{1}{2\pi}(1 + 2\sum_{k=1}^{\infty} \rho^k \cos k(\theta - \mu)), \tag{1.6}$$

$$= \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2 - 2\rho\cos(\theta-\mu)}, \tag{1.7}$$

where $\rho = e^{-\sigma}$. (see Jammalamadaka and Sengupta, 2001)

This distribution is unimodal and symmetric. As $\rho \to 0$, the distribution tends to the uniform circular distribution and as $\rho \to 1$, the distribution goes to a point distribution with the concentrated direction $\mu$.

**Circular normal distribution**

Circular normal distribution was proposed as a statistical model by von Mises (1918) and we also call it von Mises distribution. The density function of a circular random variable with this distribution is shown as below

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_o(\kappa)} e^{\kappa\cos(\theta-\mu)}, \quad \theta \in [-\pi, \pi), \tag{1.8}$$

where $-\pi \leq \mu < \pi$ and $\kappa \geq 0$. Here $I_0(\kappa)$ is the modified Bessel function of the first kind

and order zero, and it's given by

$$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} exp(\kappa\cos\theta)d\theta = \sum_{r=0}^{\infty}(\frac{\kappa}{2})^2(\frac{1}{r!})^2. \tag{1.9}$$

Also, we could get the cumulative distribution function of the von Mises distribution based on the density function given in

$$F(\theta) = \frac{1}{2\pi I_0(\kappa)}(\theta I_0(\kappa) + 2\sum_{p=1}^{\infty}\frac{I_p(\kappa)\sin p(\theta-\mu)}{p}), \quad \theta \in [-\pi, \pi). \tag{1.10}$$

(Jammalamadaka and Sengupta, 2001)

## 1.4 From parametric density estimation to nonparametric density estimation

### 1.4.1 Parametric density estimation

Density estimation, basically is a method to conduct an estimator for the density function $f(x)$ from the random sample $X$. In statistics, it can be broadly divided into two types: parametric density estimation and nonparametric density estimation. We firstly give a brief introduction about the parametric one.

Consider a random sample $X$ that has the probability density function $f(x)$, and the function $f(x)$ gives a natural description of the distribution of $X$, we know

$$P(a < X < b) = \int_a^b f(x)dx.$$

Parametric density estimation is based on a specific theoretical distribution. Suppose the distribution of the sample data is known, for example, the normal distribution with mean $\mu$ and variance $\sigma^2$. We could firstly use the available data to find the estimator of $\mu$ and $\sigma^2$, and then the density function $f$ can be estimated according to the formula of

normal distribution.

Here we would like to describe one of the popular methods of parametric density estimation, namely, the *Maximum Likelihood* method.

Suppose we have the data set $D = \{x_1, x_2, ..., x_n\}$ corresponding to random sample $\{X_1, X_2, ..., X_n\}$ from a distribution with density function $f$, where $(x_1, x_2, ..., x_n)$ are $n$ independent and identically distributed observations. The density function $f$ is from a known family of distributions. Here we suppose $f(x) = N(\mu, \sigma^2)$. Then the parameters we want to find will be $\theta = (\mu, \sigma^2)$. (Johnson and Wichern, 2007)

Now we consider the joint density function with respect to the data set $D$:

$$f(x_1, x_2, ..., x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta). \tag{1.11}$$

We also call the above function as likelihood function of $\theta$,

$$L(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta). \tag{1.12}$$

Then we can do some algebra to find a value of $\theta$ that maximizes the above function. But instead of maximizing this function, it's usually easier to maximize the natural logarithm of it:

$$\begin{aligned} ln(L(\theta|x_1, ..., x_n)) &= \ln(\prod_{i=1}^{n} p(x_i|\theta)) \\ &= \sum_{i=1}^{n} \ln p(x_i|\theta). \end{aligned}$$

We call it the log-likelihood function and one method to maximize the log-likelihood function is the standard method from Calculus.

Since this density estimation is based on the specific distribution, it has a high statis-

tical efficiency if the model is correct. Otherwise a method not dependent on a model may be more appropriate.

### 1.4.2 Importance of nonparametric density estimation

With the increasing application of large database and the rise of data mining, the nonparametric density estimation becomes more attractive. In a nonparametric density estimation, we don't need to know the probability density function and we just use the data itself. There are not so many restrictions such as the parametric density estimation - a parametric family of distribution or something else - we can explore the information we need just from the data set itself. It's apparently more practical. Another attraction of nonparametric density estimation is that it's more comprehensible for people who don't have much statistical knowledge.

There are many kinds of nonparametric density estimation methods such as histogram estimation, nearest neighbor estimation, kernel density estimation, and so on. We will give some brief introduction for two of them in the following chapter.

# Chapter 2

# Nonparametric Density Estimation

## 2.1 General method

In this section we will give some brief introduction for some main univariate nonparametric density estimation.

### 2.1.1 Histograms

Histogram estimation is the earliest and most widely used density estimation method. Histogram is a kind of representation of the sample data. In a one-dimensional case, we divide real lines into cells that have equal size-which we also call "bin". Suppose $x_0$ is the origin and $h$ is the bandwidth, we define the bins of the histogram to be the intervals $[x_0 + mh, x_0 + (m+1)h)$ for any integer $m$, and here we choose the left of the interval is closed and the right to be an open one for the definiteness. (Silverman, 1986)

Now we could write the histogram as:

$$\hat{f}(x) = \frac{1}{nh}(\text{number of } X_i \text{ in the same bin as } x).$$

Suppose that we have divided the real lines into bins, then the histogram estimate can be defined as:

$$\hat{f}(x) = \frac{1}{n} \times \frac{(\text{number of } X_i \text{ in the same bin as } x)}{(\text{width of bin containing } x)}.$$

Recall that, if we want to construct a histogram, firstly we should choose a bin width and an origin. Obviously, the bin width affects much more on our estimation and it raised some problems like when we choose a big bin width and highlight the role of averaging, the potential details of the estimation may not be fully demonstrated. On the other hand, if we choose a small bin width, the random may affect too much on the histogram and we might get an irregular shape, which could lead to a not so correct conclusion.

Although histogram density estimation has some weakness, it's still an excellent tool for data analyzing and I look forward to further study it in the future.

### 2.1.2 The naive estimator

As we already know, the histogram estimation has a serious problem - the bin width selection, and sometimes some extreme cases may happen. Due to the uncertainty of data, some bins might be empty while some others have more than 100 occurrences. If that so, the probability density between two adjacent bins - one has 100 occurrences and another one is empty - would vary greatly. Considering this situation, the naive estimator raised.

In this estimator, Silverman (1986) considered the definition of a probability density:

**Definition 2.1.1** *If the random variable X has a density f, then*

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h \ < \ X \ < \ x + h)$$

*for any given* $x$.

Based on the above definition, we can choose a small $h$ and get the estimator by calculating the proportion of all the $X_i$ falling in the interval $(x - h, x + h)$, then the naive

estimator is given by

$$\hat{f}(x) = \frac{1}{2hn}(\text{number of } X_1, ..., X_n \text{ falling in } (x-h, x+h)). \tag{2.1}$$

To express this estimator more transparently, Silverman (1986) defined a weight function $w$ by

$$w(x) = \begin{cases} \frac{1}{2}, & \text{if } |x| < 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

Then the naive estimator can be written as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} w(\frac{x - X_i}{h}). \tag{2.3}$$

The generalization of the above estimator to generate weight functions gives, what is known as the kernel density estimator described below.

## 2.2 Kernel density estimation

### 2.2.1 Definition

Kernel density estimation was originally put forward by Parzen (1962) and Rosenblatt (1956), and it was popularized in many subsequent papers. The basic form of the common kernel density estimator (Silverman, 1986) is described below.

Suppose $(X_1, X_2, ..., X_n)$ is an independent and identically distributed sample of a random variable X with an unknown density function $p(x)$, then the kernel density estimator of $p(x)$ is given by

$$\hat{p}(x; h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{2.4}$$

where $K$ is the kernel function, a symmetric function that integrates to 1 and its mean is 0. $h$ is the smoothing parameter, usually called the $bandwidth$ or $windowwidth$. Generally, we choose the kernel function $K$ to be symmetric around zero but it's not necessary to be a positive function. Typically, the bandwidth $h$ tends to 0 ad the sample size $n$ tends to infinity.

There are many types of kernel functions such like:

Uniform kernel function

$$K(u) = \frac{1}{2} I(|u| \leq 1),$$

Triangular kernel function

$$K(u) = (1 - |u|) I(|u| \leq 1),$$

Quartic kernel function

$$K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1),$$

Gaussian kernel function

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

It has been proved that the kernel density function is not the crucial part in a kernel density estimator, any kernel function can guarantee the consistency of the density estimation (Wand and Jones, 1995). However, choosing an optimal bandwidth is a much more important problem. That is because, a big bandwidth may lead to a over-smoothed estimator, and a small bandwidth may yield a density estimator which is spiky and very hard to explain. So we will give more details in the following part about the bandwidth selection.

### 2.2.2 Bandwidth selection

There are different methods for us to specify the bandwidth h, such like Rule-of-thumb, Biased cross-validation, Unbiased cross-validation and so on. Basically, they are all

based on *AMISE* (asymptotic mean integrated squared error). So before we get to know those methods, we would like to introduce *AMISE* first.

There is a vector for us to compare the estimated density function with the real one and we call it *ISE* (integrated squared error). It's given by:

$$ISE = \int_{-\pi}^{\pi} \left\{ \hat{f}(x) - f(x) \right\}^2 dx. \tag{2.5}$$

By averaging *ISE* we could get another quantity, called *MISE* (mean integrated squared error), and the asymptotic approximation for *MISE* is just the *AMISE*. The *AMISE* is usually derived via the Taylor's series. Based on some certain assumptions, we have the *AMISE* as shown below (Tang, 2011):

$$AMISE(\hat{f}(x;h)) = \frac{1}{nh} R(K) + \frac{1}{4} h^4 \mu_2(k)^2 R(f^{''}). \tag{2.6}$$

Here $\mu_2(K) = \int x^2 K(x) dx$, $R(K) = \int K^2(x) dx$, and $f^{''}$ is the second derivative of the density $f$.

The optimal bandwidth can be obtained by minimizing MISE, and it's given by:

$$h_{AMISE} = (\frac{R(K)}{\mu_2(K)^2 R(f^{''}) n})^{1/5}. \tag{2.7}$$

However, neither the *AMISE* nor the $h_{AMISE}$ can be used directly since there is always the unknown density function $f$. So now we can introduce some popular method for bandwidth selection based on these above equations.

**Rule-of-thumb**

Deheuvels (1977) proposed the Rule-of-thumb firstly and it was popularized by Silverman (1986) later, and that's why it's also called by Silverman's Rule-of-thumb. Based on the equation (2.7), Deheuvels proposed the kernel function $K$ as the Gaussian distribution and the standard normal distribution as the reference distribution, and the estimator of $h$ is given by:

$$\hat{h}_{ROT} = 1.06 \hat{\sigma} n^{-1/5}, \tag{2.8}$$

where $\sigma^2$ is the sample variance.

This method is easy to compute, however, it could also yield some problems. When the density is not close to being normal, the estimation may become inaccurate. To fix that, Silverman (1986) proposed a modified estimator which can decrease this inaccuracy:

$$\hat{h}_{M.ROT} = 1.06 * min(\hat{\sigma}, \frac{Q}{1.34})n^{-1/5}, \tag{2.9}$$

where $Q$ is the interquartile range:

$$Q = X_{[0.75n]} - X_{[0.25n]}. \tag{2.10}$$

When the density is very close to the normal distribution, both estimators $\hat{h}$ are good and helpful. However, if the true density is not normal distribution or we can't determinate it, the second estimator gives a better result.

**Unbiased cross-validation**

The Unbiased cross-validation method is based on the formula of MISE:

$$\begin{aligned} MISE(\hat{f}(x;h)) &= E\int(\hat{f}(x;h) - f(x))^2 dx \\ &= E\int \hat{f}^2(x;h)dx - 2E\int \hat{f}(x;h)f(x)dx + \int f^2(x)dx. \end{aligned} \tag{2.11}$$

Chaubey et al. (2012) adapted this method Wand and Jones (1995) that we describe below.

Since the last term in the above function is constant, we can drive an unbiased estimator of MISE with replacing the first two terms by its estimator.

$$UCV(h) = \int \hat{f}^2(x;h)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_i(X_i;h). \tag{2.12}$$

Noted here the second term is the *leave-one-out estimator* given by

$$\hat{f}_i(X_i;h) = \frac{1}{n-1}\sum_{j\neq i}^{n} K_h(x - X_j), \tag{2.13}$$

where

$$K_h(u) = \frac{1}{h}K(\frac{u}{h}).$$

Then we can obtain an optimal $\hat{h}_{UCV}$ by finding the $h$ that minimized the function (2.13).

**Biased cross-validation**

The biased cross-validation method was proposed by Scott and Terrell (1987). Recall the asymptotic mean squared error:

$$AMISE(\hat{f}(x;h)) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(k)^2R(f^{''}). \tag{2.14}$$

To get the estimator of $h$, they used the estimated function of $R(f^{''})$ instead of a reference distribution,

$$\hat{R}(f^{''}) = \frac{1}{n^2}\sum\sum_{i\neq j}(K_h^{''} * K_h^{''})(X_i - X_j). \tag{2.15}$$

Then it leads to the score function:

$$BCV(h) = \frac{R(K)}{nh} + h^4\frac{\mu_2^2(K)}{4n^2}\sum\sum_{i\neq j}(K_h^{''} * K_h^{''})(X_i - X_j). \tag{2.16}$$

Scott and Terrell (1987) proposed to use the optimal bandwidth $\hat{h}_{BCV}$ which could minimize the function above.

**Conclusion**

After we introduced some widely used bandwidth selection methods, an important question must be asked by ourselves: which method is the most efficient one? It's really difficult to define which one is "the best", since it all depends on the situation. When we analyze a real data set, the best way for us to select the "best" bandwidth is to apply different bandwidth selectors to our data and try to compare all the possible bandwidths, and then we could find out the most suitable one for our data.

## 2.3 Asymmetric density estimation

Since the standard symmetric fixed kernels are not appropriate for some density functions with bounded support, asymmetric kernels presented. Kernel density estimator with asymmetric kernels such as gamma kernels have been proposed to solve the boundary consistency problem. For example, Chen (2000) used Gamma kernels and Scaillet (2004) used inverse Gaussian (IG) and reciprocal inverse Gaussian (RIG). To focus on the case of non-negative data, Chaubey et al. (2012) proposed an estimator which is based on the generalization of Hille's smoothing lemma (Feller, 1965).

**Lemma 2.3.1** *Let $u$ be any bounded and continuous function and $G_{x,n}$, n=1,2,... be a family of distributions with mean $\mu_n(x)$ and variance $h_n^2(x)$ such that $\mu_n(x) \to x$ and $h_n(x) \to 0$. Then*

$$\hat{u}(x) = \int_{-\infty}^{\infty} u(t) dG_{x,n}(t) \to u(x). \tag{2.17}$$

*The convergence is uniform in every subinterval in which $h_n(x) \to 0$ uniformly and $u$ is uniformly continuous.*

According to this lemma, the smooth estimator of an empirical distribution function $F(x)$ could be obtained easily by replace $u(x)$ with $F_n(x)$

$$\hat{F}_n(x) = \int_{-\infty}^{\infty} F_n(t) dG_{x,n}(t). \tag{2.18}$$

To obtain more details about $\hat{F}_n(x)$, Chaubey et al. (2012) considered the following theorem.

**Theorem 2.3.1** *Let $h_n(x)$ to be the variance of $G_(x, n)$ as in Lemma , and suppose that $h_n(x) \to 0$ as $n \to 0$ for every fixed $x$. Then we have*

$$\sup_x |\widetilde{F}_n(x) - F(x)| \xrightarrow{a.s.} 0 \tag{2.19}$$

*as $n \to 0$.*

A restricted condition on $G_{x,n}(X_i)$ can be seen easily if we change the form of $\widetilde{F}_n(x)$ to:

$$\widetilde{F}_n(x) = 1 - \frac{1}{n} G_{x,n}(X_i). \tag{2.20}$$

For $\widetilde{F}_n(x)$ to be a proper distribution function, $G_{x,n}$ must be a decreasing function of $x$.

Based on that, we can get the smooth estimator of the density function given by:

$$\begin{aligned} \hat{f}_n(x) &= \frac{d\widetilde{F}_n(x)}{dx} \\ &= -\frac{1}{n} \sum_{i=1}^{n} \frac{d}{dx} G_{x,n}(X_i). \end{aligned} \tag{2.21}$$

## 2.4 Bernstein polynomial estimation

The empirical distribution function is known to have good properties when we use it to estimate distribution function. However, since it's not a continuous function, it may not be appropriate to use it to estimate a continuous distribution function. Babu, Canty and Chaubey (2002) considered the application of Bernstein polynomials to approximate a bounded and continuous function, and they proved that with a continuous approximation of the empirical distribution function, the Bernstein polynomials could be naturally adapted to smooth an estimated distribution function that concentrated on the interval [0,1].

We firstly have a look at the definition of the empirical distribution function:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\left\{X_i \leq x\right\}. \tag{2.22}$$

The above function is appropriate when the support of distribution is $\mathbf{R}^+$. If we have the support of $F$ in the interval [a,b], where $a \leq b$, it would be better to transform the variable $X$ to $Y$ with support [0,1]. This transformation could be done by $Y = \frac{X-a}{b-a}$. Based on that, Babu, Canty and Chaubey (2002) introduced the following theorem to adapt Bernstein polynomials to estimation.

**Theorem 2.4.1** *(Feller, 1965) If $u(x)$ is a bounded and continuous function on the interval [0,1], then as $m \to \infty$*

$$u_m^*(x) = \sum_{k=0}^{m} u(k/m) b_k(m, x) \to u(x) \tag{2.23}$$

*uniformly for $x \in [0, 1]$, where*

$$b_k(m, x) = \binom{m}{k} x^k (1 - x)^{m-k}, \quad k = 0, \ldots, m. \tag{2.24}$$

With the theorem given below, Babu, Canty and Chaubey (2002) considered $F$ with support [0,1], and motivated the smooth estimator $\widetilde{F}$ based on Bernstein polynomial, which is given by:

$$\hat{F}_{n,m}(x) = \sum_{k=0}^{m} F_n(\frac{k}{m}) b_k(m, x), \quad x \in [0, 1] \tag{2.25}$$

Noted here this estimator is based on the empirical distribution function $F_n$:

$$F_n(x) = n^{-1} \sum_{i=1}^{n} I\left\{X_i \leq x\right\}. \tag{2.26}$$

To prove that $\hat{F}_{n,m}$ is a proper distribution function and it has derivative, consider an alternative form of the function in (2.25).

$$\hat{F}_{n,m}(x) = \sum_{k=0}^{m} f_n(\frac{k}{m}) B_k(m, x), \tag{2.27}$$

where

$$\begin{aligned} &f_n(0) = 0, \\ &f_n(\frac{k}{m}) = F_n(\frac{k}{m}) - F_n(\frac{k-1}{m}), \quad k = 1, \ldots, m \end{aligned} \tag{2.28}$$

and

$$
\begin{aligned}
B_k(m,x) &= \sum_{j=k}^{m} b_k(m,x) \\
&= m\binom{m-1}{k-1}\int_0^x t^{k-1}(1-t)^{m-k}dt. \qquad (2.29)
\end{aligned}
$$

The above equation is given by the definition of the cumulative distribution function of a binomial distribution. Since $b_k(m,x)$ is just the form of the probability mass function of a binomial distribution with parameters $m, k, x$, $B_k(m,x) = \sum_{j=k}^{m} b_k(m,x)$ will be the probability mass function of this given binomial distribution. From the definition we can get the equation of $B_k(m,x)$.

Noted here $\hat{F}_{n,m}$ is a polynomial in terms of $x$ so it's continuous and has all the derivatives, and $0 \leq \hat{F}_{n,m} \leq 1$ for $x$ in interval $[0,1]$. Based on the equations in (2.28) and (2.29), we can see that $f_n(\frac{k}{m})$ is a non-negative polynomial and $B_k(m,x)$ is non-decreasing in $x$. Thus, we know $\hat{F}_{n,m}$ is non-decreasing also.

Upon taking the derivative of $\hat{F}_{n,m}$, we can get the density estimator of $f$ as proposed in Babu, Canty and Chaubey (2002):

$$
\begin{aligned}
\hat{f}_{n,m}(x) &= \sum_{k=1}^{m} f_n(\frac{k}{m})\frac{d}{dx}B_k(m,x) \\
&= m\sum_{k=0}^{m-1} f_n(\frac{k+1}{m})b_k(m-1,x) \\
&= m\sum_{k=0}^{m-1}(F_n(\frac{k+1}{m}) - F_n(\frac{k}{m})b_k(m-1,x)). \qquad (2.30)
\end{aligned}
$$

We will adapt this estimator for estimation of a circular density in the next chapter.

# Chapter 3

# Transformation Based Nonparametric Density Estimators

## 3.1 General method

Recall the kernel density estimator and if we consider it for circular data, we can assume a continuous circular density function $f(\theta)$, where $\theta \in [-\pi, \pi)$, $f(\theta) \geq 0$ for $\theta \in R$ and $\int_{-\pi}^{\pi} f(\theta)d\theta = 1$.

For a random sample that satisfies the above density function, i.e. $(\theta_1, ..., \theta_n)$, we can write the kernel density estimator as follows:

$$\hat{f}(\theta; h) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{\theta - \theta_i}{h}). \tag{3.1}$$

To obtain this density estimator, Fisher (1989) used the quartic kernel function to adapt the linear kernel density estimator and defined the kernel function:

$$K(\theta) = \begin{cases} 0.9375(1 - \theta^2), & \text{if } \theta \in [-1, 1] \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

Noted here the data $\theta$ must be transformed to the interval [-1,1] and the factor 0.9375 is

to ensure that $\int K(\theta)d\theta = 1$ and $K(\theta)$ is a density function. It's indeed one method to do the transformation but here is one obvious problem-the estimator is not periodic. To fix this, Fisher (1995) suggested to perform the smoothing by replicating the data to 3 to 4 circles and consider only one interval, $[-\pi, \pi)$.

## 3.2 Transformation based kernel density estimator

Since the major different between circular data and linear data is the type of data, we may adapt the linear density estimator to the circular density estimator by transforming the data interval. In this section we will focus on kernel density estimator and try to transform the data on $[-\pi, \pi)$ to $(-\infty, \infty)$. (Chaubey, 2017)

Let $f(\theta)$ be the density function of the circular data and $p(x)$ be the density function of the linear data. Suppose $x = t(\theta)$ is the one-to-one transform function $t : (-\pi, \pi) \to (-\infty, \infty)$, e.g. $t(\theta) = \tan(\frac{\theta}{2})$.

Using the transformation and the fact that

$$f(\theta) = p(t(\theta))|\frac{dt(\theta)}{d\theta}|,$$

and recall the kernel density estimator of $x$:

$$\begin{aligned} \hat{p}_K(x; h) &= \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \\ &= \frac{1}{nh}K(\frac{x - \tan(\theta_i/2)}{h}), \end{aligned} \tag{3.3}$$

then this transformation could be done by the following steps:

$$\hat{f}_K(\theta; h) = \widetilde{p}(t(\theta); h)|\frac{d(t(\theta))}{d\theta}|, \tag{3.4}$$

where

$$\begin{aligned} t(\theta) &= \tan(\frac{\theta}{2}) \\ &= \frac{\sin\theta}{1+\cos\theta}, \end{aligned}$$

and

$$\begin{aligned} \frac{d(t(\theta))}{d(\theta)} &= (\frac{\sin\theta}{1+\cos\theta})' \\ &= \frac{\cos\theta(1+\cos\theta)+\sin^2\theta}{(1+\cos\theta)^2} \\ &= \frac{1}{1+\cos\theta}, \end{aligned}$$

then

$$\hat{f}_K(\theta;h) = \frac{1}{1+\cos(\theta)}\hat{p}(\frac{\sin\theta}{1+\cos\theta};h) \tag{3.5}$$

is the estimator for the circular density function.

## 3.3 Transformation based Bernstein polynomial density estimator

Another transformation between circular density estimator and linear density estimator can be done based on Bernstein polynomials (Chaubey, 2017). We introduced the Bernstein polynomials estimator (Babu, Canty and Chaubey, 2002) in chapter 2, and to differentiate the circular density function from the linear one, we denote the linear Bernstein polynomial density estimator by $\hat{p}(x;m)$, and it's given by

$$\hat{p}_B(x;m) = m\sum_{k=1}^{m} F_n(\frac{k}{m})b_k(x;m-k+1), \quad x\in[0,1], \tag{3.6}$$

where $b_k(x;m-k+1)$ is the Bernstein polynomial.

Chaubey (2017) considered the following transformation:

$$t(\theta) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(c\tan(\frac{\theta}{2})), \tag{3.7}$$

where $c$ is a positive number.

With this function, we could convert the interval of $\theta$ - which is $[-\pi, \pi)$ to [0,1), and it's also a one-to-one transformation for all $c > 0$. An important thing is that the transform function $t(\theta)$ is periodic.

Now the transformed estimator of $f(x)$ can be obtained and it's given by

$$\hat{f}_B(\theta; m) = \frac{1}{2\pi}\hat{p}_B(t(\theta); m)\frac{c(1 + \tan^2(\frac{\theta}{2}))}{1 + c^2\tan^2(\frac{\theta}{2})}. \tag{3.8}$$

# Chapter 4

# A Simulation Study

## 4.1 Introduction

In order to compare the two estimators we discussed in Chapter 3, I would like to use simulation method in this chapter. Concerning the data we would like to use is circular data, I'd like to introduce ISE and MSE as the reference for comparing them.

For ISE, integrated squared error, the function is given by:

$$ISE = \int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 dx. \tag{4.1}$$

From the function above, we can see that the value of ISE is the square of the distance between the estimated density and the simulated density. We can simplify the function as:(Chaubey et al., 2012)

$$ISE = \int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 dx \tag{4.2}$$

$$= \int_{-\infty}^{\infty} \hat{f}^2(x) dx - 2\int_{-\infty}^{\infty} \hat{f}(x) f(x) dx + \int_{-\infty}^{\infty} f^2(x) dx. \tag{4.3}$$

The above function motivated a method to smooth the parameter, which is called unbiased cross validation. Since the third term in the function $ISE$ is constant, and we can

estimate the second term by the $leave - one - out$ estimator $\frac{2}{n}\sum \hat{f}_n(X_i;\omega_i)$,
where $\omega_i = \{X_1, ..., X_{i-1}, X_{i+1}, ...X_n\}$. The estimated ISE with cross-validation in our case is given by:

$$CV_{ISE} = \int_{-\pi}^{\pi} \hat{f}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_n(X_i;\omega_i). \tag{4.4}$$

With this unbiased estimator of *ISE*, we could obtain the optimal $h$ in the kernel density estimator and $m$ in the Bernstein polynomial density estimator by minimizing the $ISE.CV$ function. To makes it more clearly, I will check the mean, median and standard deviation of $ISE$ and $MSE$ for these two estimators based on different distributions and different sample sizes.

Before we start the simulation, it's better to look at the range of bandwidth in the kernel estimator firstly. Here we use the Gaussian kernel function as the kernel function $K$ and it's given by:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}. \tag{4.5}$$

When we plug the kernel function into the kernel density estimator for circular data,

$$\hat{f}_K(\theta;h) = \frac{1}{1+\cos(\theta)}\hat{p}(\frac{\sin\theta}{1+\cos\theta};h), \tag{4.6}$$

and here the estimated density function for linear data is given by:

$$\hat{p}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h}), \tag{4.7}$$

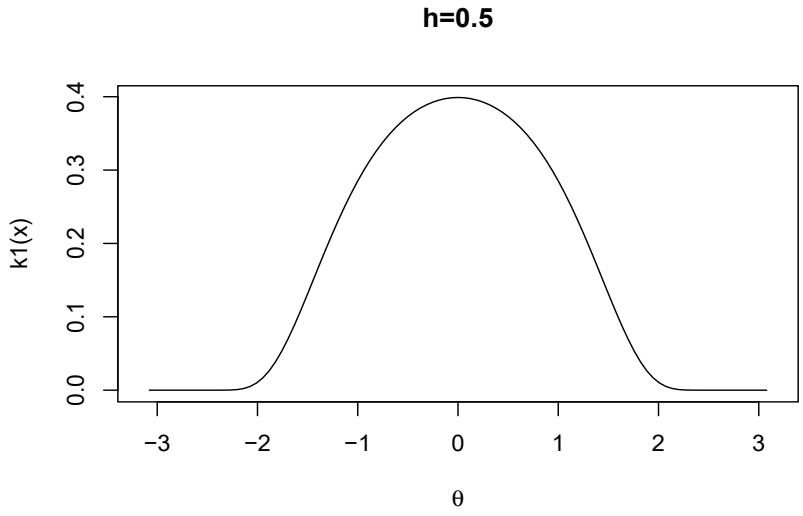
we can see that the density function is nothing but the mean of the following function:

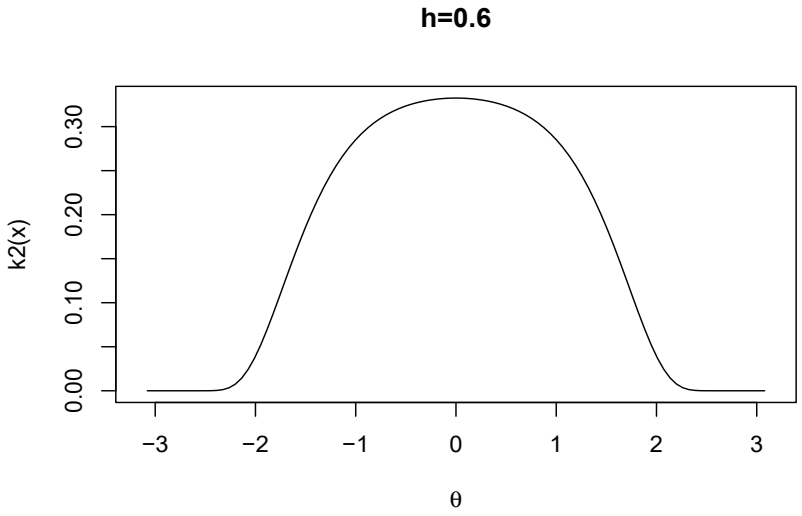
$$k(\theta,h) = \frac{1}{h(1+\cos(\theta))} K(\frac{\tan(\theta/2)-\tan(\theta_i/2)}{h}). \tag{4.8}$$

Noted here we want the kernel estimated density function to be symmetric around zero and it also has the maximum value in zero, so we have to ensure that when $\frac{x-x_i}{h}$ equals to $0$, the derivative of function $\hat{k}$ is also $0$. To make it more convenient for us to study, we use $\theta$ to replace $\tan(\theta/2) - \tan(\theta_i/2)$ in equation 4.8 and took a look at the plot of the

following function:

$$\hat{k}(\theta, h) = \frac{1}{h\sqrt{2\pi}(1+\cos\theta)} exp\left[-\frac{1}{2h^2}(\frac{\sin\theta}{1+\cos\theta})^2\right]. \tag{4.9}$$



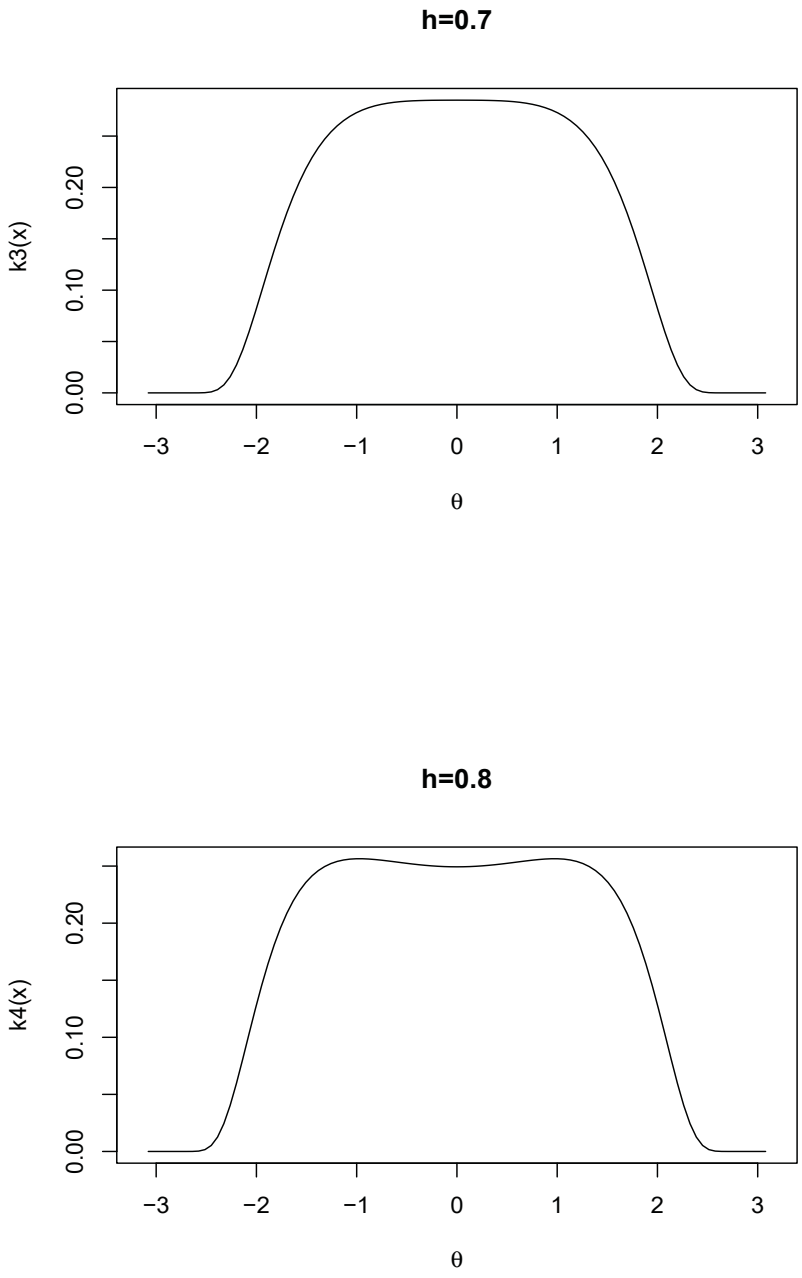



Figure 4.1: Kernel function with different h value

When $h = 0.5$, the given function is definitely symmetric around zero and its maximum value also appears at zero. With the value of h becomes larger, the upper part gets more flatter. For $h = 0.7$, the peak almost disappear and it's very flat. Even more, when $h = 0.8$, the upper part becomes hollow and two peak shows up. Since the special condition we want to have for function $\hat{k}(\theta, h)$, we restrict $h$ to be smaller than or equal to 0.6 in this thesis.

In this simulation we considered 3 models in total: wrapped Cauchy distribution with ($\sigma$=1, $\mu$=0), von Mises distribution with ($\kappa$=1, $\mu$=0), and a mixture distribution of 40% von Mises distribution with ($\kappa$=1, $\mu$=0) and 60% Wrapped Cauchy distribution with ($\sigma$=1, $\mu$=0). The domain for each model is in $[-\pi, \pi)$.

## 4.2 Global comparison

Here are the results of our global comparison for these two estimators. Noted here the function we used to calculate ISE is given by:

$$ISE = \int_{-\pi}^{\pi} (\hat{f}(x) - f(x))^2 dx, \tag{4.10}$$

and the parameter $h$ is obtained by unbiased cross validation method we mentioned before.

| ISE | | n=100 | n=200 | n=400 |
|---|---|---|---|---|
| mean | kernel | 0.007839894 | 0.004649004 | 0.002747486 |
| | Bernstein(c=1) | 0.00912195 | 0.005277216 | 0.003033216 |
| | Bernstein(c=0.5) | 0.03048433 | 0.02982057 | 0.01277822 |
| | Bernstein(c=2) | 0.04376135 | 0.04254877 | 0.04193504 |
| median | kernel | 0.006207218 | 0.0037773064 | 0.002226404 |
| | Bernstein(c=1) | 0.007930931 | 0.004729214 | 0.00269449 |
| | Bernstein(c=0.5) | 0.02962856 | 0.0294122 | 0.01277754 |
| | Bernstein(c=2) | 0.04343982 | 0.04271339 | 0.04222359 |
| sd | kernel | 0.006462172 | 0.003479234 | 0.00214282 |
| | Bernstein(c=1) | 0.005054517 | 0.002774371 | 0.001544781 |
| | Bernstein(c=0.5) | 0.003076889 | 0.001725131 | 0.009291902 |
| | Bernstein(c=2) | 0.01029532 | 0.007748815 | 0.005816618 |

Table 4.1: Transformation based on different estimators with wrapped Cauchy distribution ($\sigma$=1, $\mu$=0)

If we focus on Bernstein polynomial density estimator only, we notice that as the value of $c$ goes greater, the average value of *ISE* is also getting greater and the accuracy of estimation is declining. At the same time, when we decrease the value of $c$, the aver-

age value of *ISE* also gets greater, which gives us the result that $c = 1$ gives us a better estimation than other $c$ values. So we will only compare kernel density estimator with Bernstein polynomial density estimator with $c = 1$

When we compare the kernel density estimator with the Bernstein polynomial density estimator, we found that the kernel density estimator performs a little better than Bernstein polynomial density estimator when sample size $n$=100. As the sample size goes greater, the difference between their *ISE* values becomes smaller and smaller. When sample size $n$ is 400, these two estimators' *ISE* values are very close.

| ISE | | n=100 | n=200 | n=400 |
|---|---|---|---|---|
| mean | kernel | 0.007430471 | 0.004372632 | 0.002349617 |
| | Bernstein(c=1) | 0.007600224 | 0.004474659 | 0.002517603 |
| median | kernel | 0.005641455 | 0.003463993 | 0.001899657 |
| | Bernstein(c=1) | 0.006388351 | 0.003918022 | 0.002181035 |
| sd | kernel | 0.006650624 | 0.002686281 | 0.001797038 |
| | Bernstein(c=1) | 0.004872336 | 0.002686281 | 0.001405067 |

Table 4.2: Transformation based on different estimators with von Mises distribution ($\kappa$=1, $\mu$=0)

In the second model, the performance of kernel density estimator still looks like very similar with Bernstein polynomial density estimator since their *ISE* values are very close for all three sample sizes.

| ISE | | n=100 | n=200 | n=400 |
|---|---|---|---|---|
| mean | kernel | 0.005584709 | 0.00441631 | 0.002503583 |
| | Bernstein(c=1) | 0.00570089 | 0.004810359 | 0.002700961 |
| median | kernel | 0.004631101 | 0.003433499 | 0.002131873 |
| | Bernstein(c=1) | 0.005434047 | 0.00423467 | 0.002340753 |
| sd | kernel | 0.004009296 | 0.00334669 | 0.00183587 |
| | Bernstein(c=1) | 0.002532875 | 0.002758459 | 0.001556152 |

Table 4.3: Transformation based on different estimators with Mixture distribution

In the third model, we considered a mixture distribution of 40% von Mises distribution and 60% wrapped Cauchy distribution. The result shows that kernel density estimator still performs much similar as Bernstein polynomial density estimator since their *ISE* values are very close.

We cannot decide which estimator is better only depend on their *ISE* values, and we would like to continue our local comparison based on *MSE*.

## 4.3 Local comparison

In this section we will focus on the factor *MSE*, mean square error, which defined as

$$\begin{aligned} MSE(\hat{f}(\theta)) &= \frac{1}{N} E\left\{(\hat{f}(\theta) - f(\theta))^2\right\} \\ &= \int (\hat{f}(\theta) - f(\theta))^2 f(\theta) d\theta. \end{aligned} \tag{4.11}$$

Instead of computing the *MSE* directly, we would like to choose the estimator of it:

$$\hat{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{f}_i(\theta) - f(\theta))^2, \tag{4.12}$$

where $N$ is the replication number. (Allen, 1971)

This $\hat{MSE}$ is easier for us to compute, and it could also tell us the estimate ability of these two estimators.

We'd like to express the performance of *MSE* in form of graphs, and consider different sample size we have in each distribution, there will be 9 cases in total: wrapped Cauchy distribution with parameters ($\sigma$=1, $\mu$=0) and sample size $n$=(100, 200, 400) separately, von Mises distribution with parameters ($\kappa$=1, $\mu$=0) and sample size $n$=(100, 200, 400) separately , mixture distribution of 40% von Mises distribution with parameter ($\kappa$=1, $\mu$=0) and 60% wrapped Cauchy distribution with parameter ($\sigma$=1, $\mu$=0) and sample size $n$=(100, 200, 400) separately.

In the first case we compared kernel density estimator and Bernstein polynomial density estimator with $c = (1, 2, 0.5)$, and the effect of increasing or decreasing the value of $c$ is exactly the same as what we found in global comparison. $c = 1$ results in a better estimation in our case. So we will only compare kernel density estimator with Bernstein polynomial density estimator with $c = 1$ in the left 5 cases.

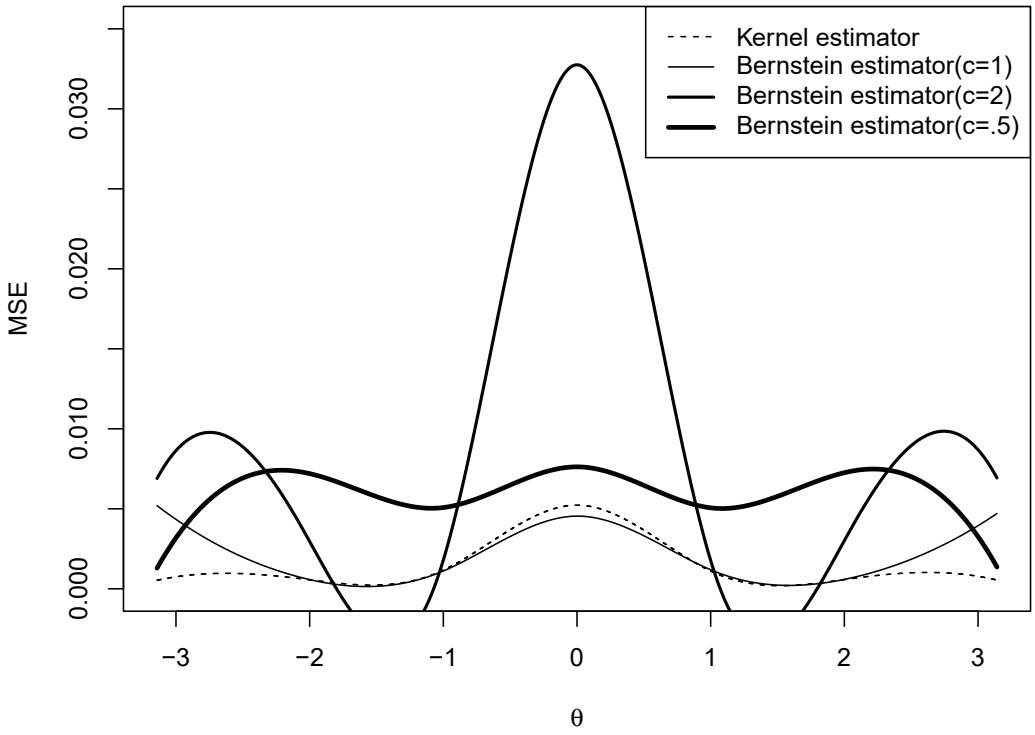



Figure 4.2: Mean squared error for wrapped Cauchy distribution ($\sigma$=1, $\mu$=0,n=100)

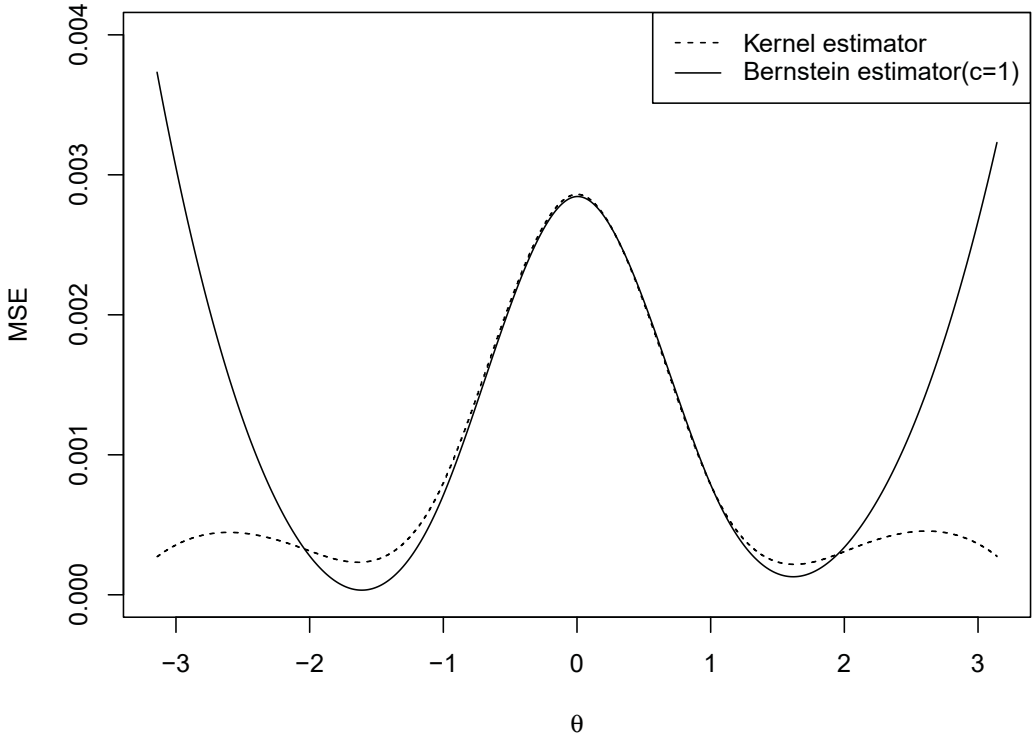



Figure 4.3: Mean squared error for wrapped Cauchy distribution ($\sigma$=1, $\mu$=0,n=200)

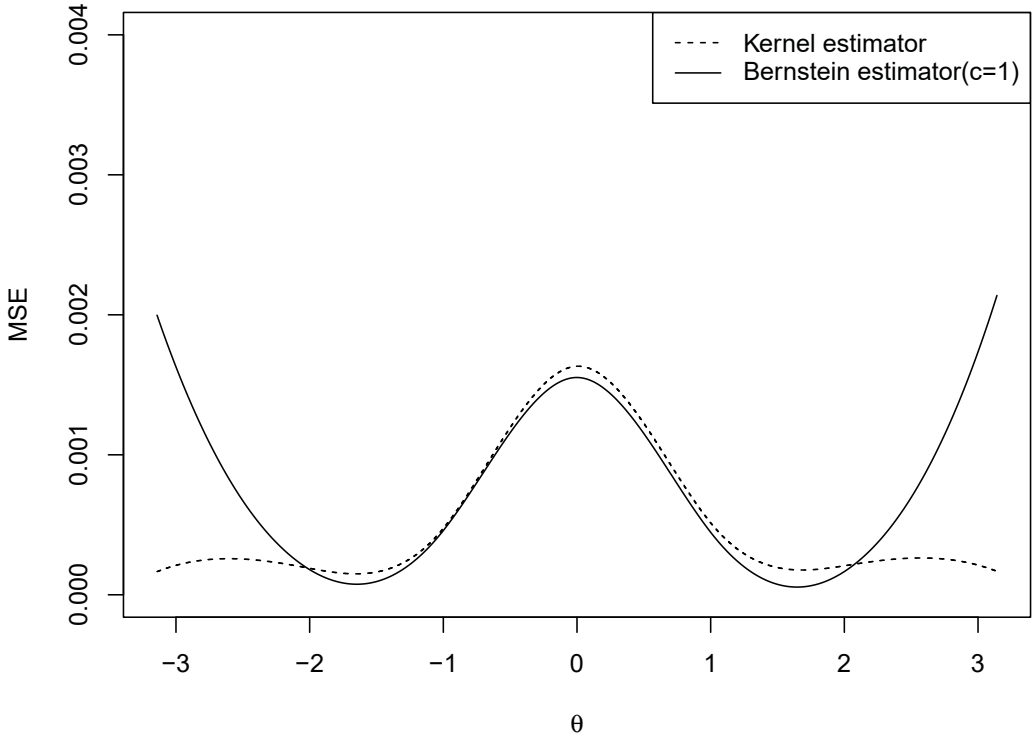



Figure 4.4: Mean squared error for wrapped Cauchy distribution ($\sigma$=1, $\mu$=0,n=400)

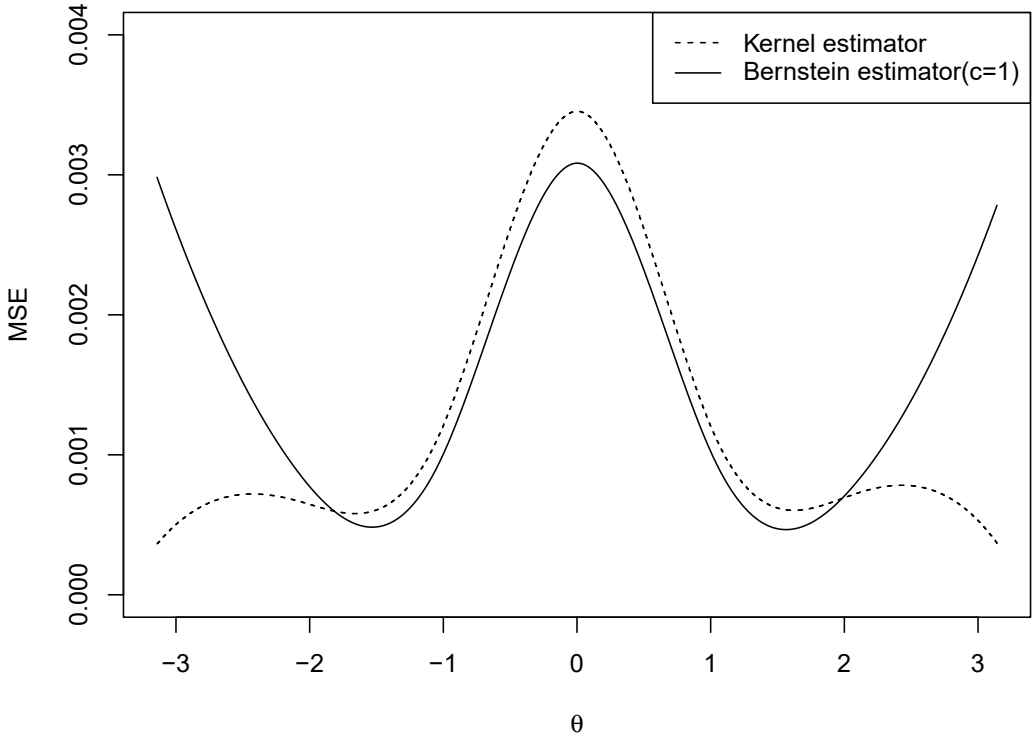



Figure 4.5: Mean squared error for von Mises distribution ($\kappa$=1, $\mu$=0,n=100)

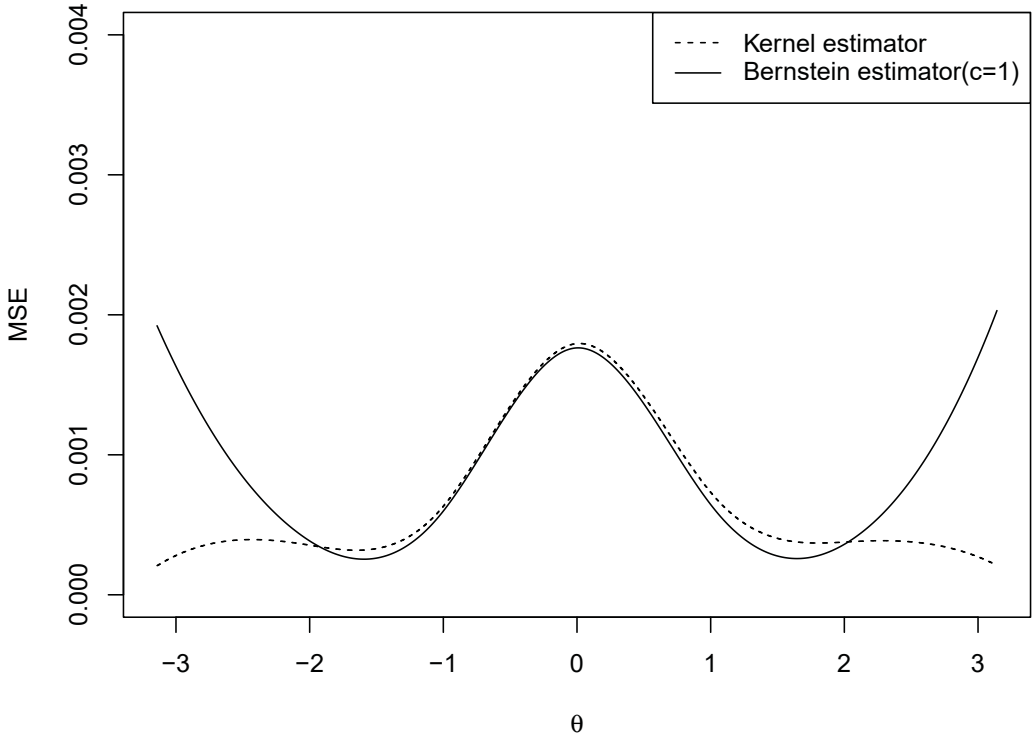



Figure 4.6: Mean squared error for von Mises distribution ($\kappa$=1, $\mu$=0,n=200)

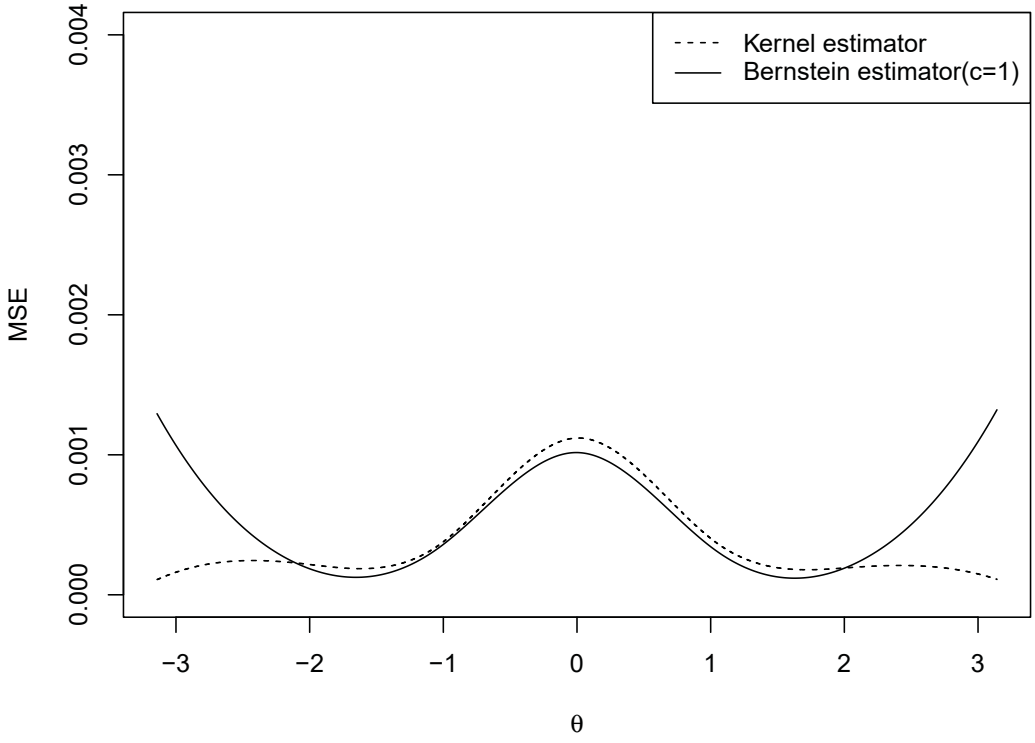



Figure 4.7: Mean squared error for von Mises distribution ($\kappa$=1, $\mu$=0,n=400)

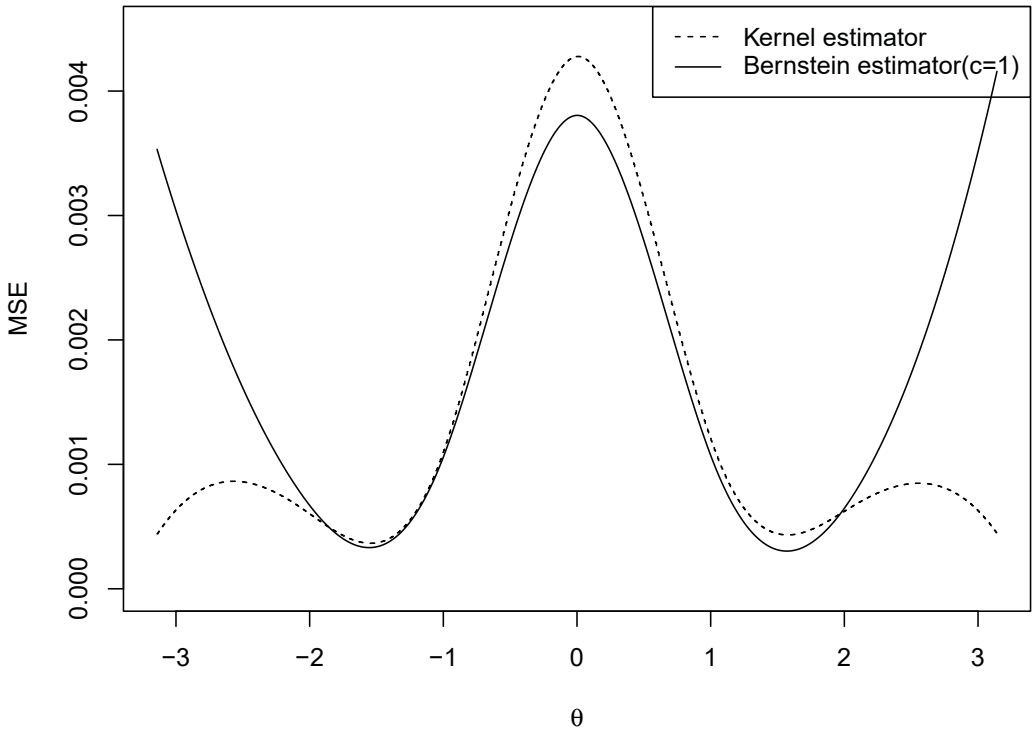



Figure 4.8: Mean squared error for Mixture distribution(n=100)

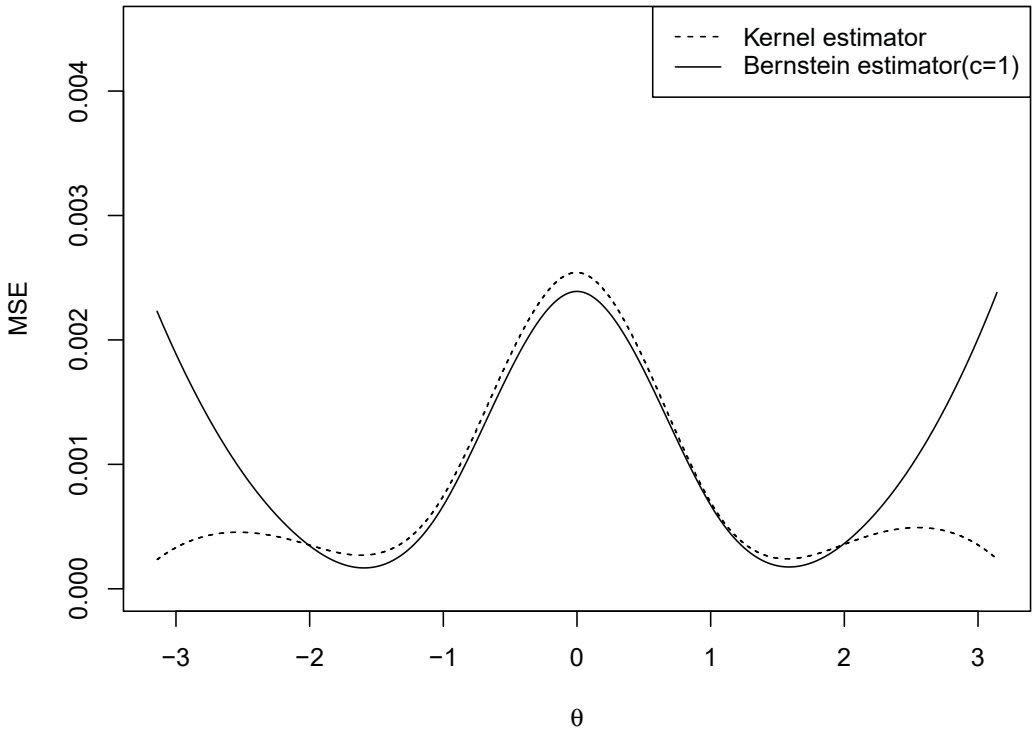



Figure 4.9: Mean squared error for Mixture distribution(n=200)

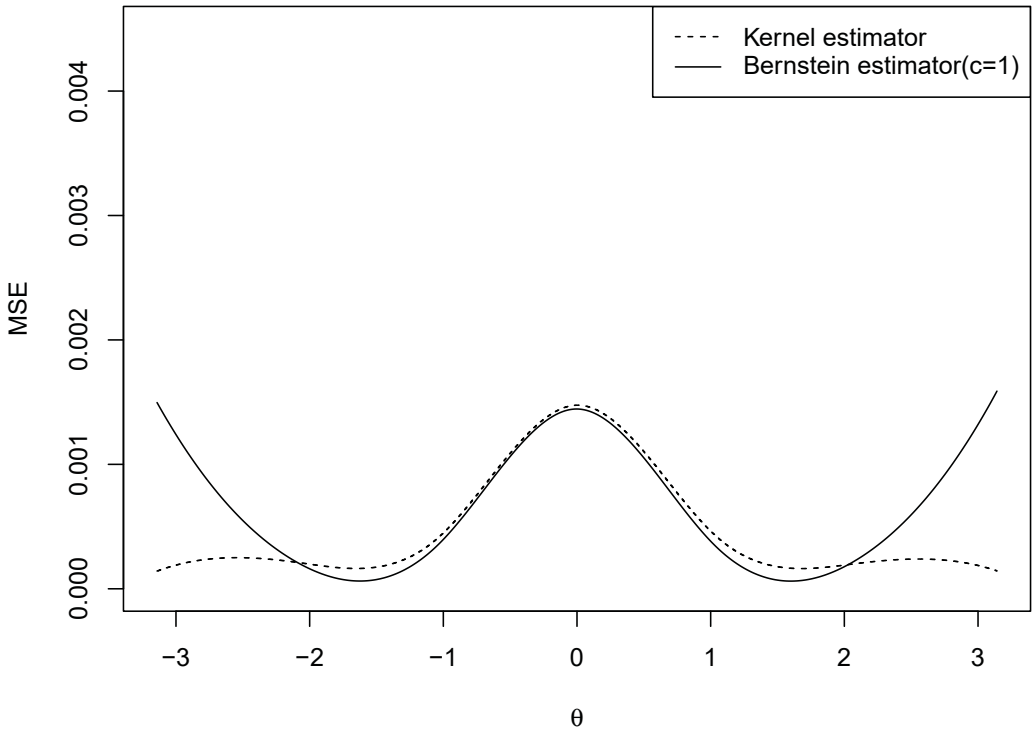



Figure 4.10: Mean squared error for Mixture distribution(n=400)

## 4.4 Conclusion and further research

### 4.4.1 Conclusion

In the first case ($n = 100$) of the first distribution (Figure 4.2), if we focus on the tails, we can see that kernel estimator performs better than Bernstein estimator, and comparing different values of $c$, $c = 0.5$ gives a better result than $c = 1$ and $c = 2$. However, in the central part, Bernstein polynomial estimator with $c = 1$ performs pretty much similar to the kernel density estimator, and as the value of $c$ becomes greater (c=2) smaller (c=.5), the value of MSE also increases conspicuously. Further looking at the result with different sample sizes, we see that these two estimators are quite similar when estimating the central part, and for the tails, kernel density estimator performs much better than Bernstein polynomial estimator. So in general we can say that kernel density estimator is more accurate than Bernstein polynomial estimator for the wrapped Cauchy distribution. However, investigation by considering other circular probability density functions for simulation may be desired.

Figure 4.5, 4.6 and 4.7 give the results based on von Mises distribution. In this case kernel estimator still performs better than Bernstein estimator when we estimate the tails. However, for estimating the central part, Bernstein estimator performs a little bit better than kernel estimator when the sample size is not big enough. As the sample size goes bigger, the performances of these two estimator becomes very similar.

For the third distribution (See Figure 4.8, 4.9 and 4.10), the result also shows that kernel estimator has a stronger estimation ability than Bernstein estimator. In general, these two transformations are very comparable when we estimate the central part for all distributions, however, the kernel transformation are much more efficient than Bernstein transformation when it comes to the tail.

To sum it up, for all distributions we mentioned in this chapter, the performances of these two estimators are very similar when estimating the central part and the kernel estimator is more efficient than the Bernstein estimator with the tail part. In general

we can say, the kernel density estimator is better than Bernstein polynomial density estimator.

### 4.4.2 Further research

Since we just compared two estimators based on three models, it's conspicuous that kernel density estimator has a stronger ability to alleviate the boundary problem than Bernstein polynomial density estimator in these models. However, we may need to do more about the boundary problems in the transformed Bernstein polynomial density estimator. On the other hand, we just used three different sample size to do the research and small sample size may cause many problems, so more tests with larger sample size may be needed in the further research.

Another important line of investigation would be to consider other circular probability distribution for simulation which are not necessarily symmetric. One way to generate such distribution would be to consider mixture distribution such as the one considered in this thesis, but with different mean directions.

Also for a definitive treatment for the choice of $c$ could be to analyze the asymptotic nature of the *MISE*; the leading term may cast some light on its optimal choice.

# APPENDIX: R codes for computing ISE and MSE for transformed kernel density estimator (Based on Wrapped Cauchy distribution ($\sigma$=1, $\mu$=0))

```
##Global Error: ISE
##Kernel estimator
##Wrapped Cauchy Distribution

##p.d.f of wrapped cauchy distribution with mu=o, rho=exp(-1)
WCauchyf=function(theta){
        (1/(2*pi))*((1-(exp(-1))^2)/(1+(exp(-1))^2-2*exp(-1)*cos(theta)))
}

##fhat function
fhat.kernel<-function(theta,thetas,h){
dcnorm<-function(theta,h){
        (1/((2*pi)^0.5*h*(1+cos(theta))))*exp(-0.5*(sin(theta)/
        (h*(1+cos(theta))))^2)
}
        xs<-tan(thetas/2)
        mean(dcnorm(theta-thetas,h=h))
```

```
}

#CV.ISE function
CVISE.kernel<-function(h,thetas){
    n<-length(thetas)
    FT.Int<-function(theta,thetas=thetas,h=h)
            {(fhat.kernel(theta,thetas,h))^2}
        FT<-Vectorize(FT.Int,"theta")
        first.termK<-integrate(FT,-pi,pi,thetas=thetas,h=h)$value
  #second term
        ST<-0
        for (i in 1:n){
            ST=ST+fhat.kernel(thetas[i],thetas=thetas[-i],h=h)
        }
        second.termK<-2*ST/n
        cv<-first.termK-second.termK
        return(cv)
        }

##Minimize CV.ISE to find h
h.cv=as.numeric(optimise(CVISE.kernel,interval=c(0,0.6),
        thetas=thetas)[1])


##Definition of ISE
func.integrand=function(theta,thetas,h){
        (fhat.kernel(theta,thetas,h)-WCauchyf(theta))^2
}

ISE.kernel=replicate(1000,{
        thetas=rwrpcauchy(100,location=pi,rho=exp(-1))-pi
```

```
        h.cv=as.numeric(optimise(CVISE.kernel,interval=c(0,0.6),
        thetas=thetas)[1])
        integrate(func.integrand,-pi+0.1,pi,thetas=thetas,h=h.cv)$value
})

mean(ISE.kernel)
median(ISE.kernel)
sd(ISE.kernel)

##Local Error: MSE
##Kernel estimator
##Wrapped Cauchy Distribution

##p.d.f of wrapped cauchy distribution with mu=o, rho=exp(-1)
WCauchyf=function(theta){
  (1/(2*pi))*((1-(exp(-1))^2)/(1+(exp(-1))^2-2*exp(-1)*cos(theta)))
}

#MSE definition for kernel transformation
mse.kernel.wc=function(theta,thetas,h){
  (fhat.kernel(theta,thetas,h)-WCauchyf(theta))^2
}

theta.mse=seq(-pi,pi,length.out = 7)

MSE.wc=replicate(1000,{
  thetas=rwrpcauchy(100,location = pi,rho=exp(-1))-pi
  h.cv=as.numeric(optimise(CVISE.kernel,interval=c(0,0.6),
       thetas=thetas)[1])
  MSEmatrixy=sapply(theta.mse,mse.kernel.wc,thetas=thetas,h=h.cv)
  MSEmatrixy
```

```
})

KernelMSE1=rowMeans(MSE.wc)

#MSE Plot for kernel estimator
sp1=spline(theta.mse,KernelMSE1,n=1000)
plot(sp1,ylim=c(0,0.035),xlab=expression(theta),ylab="MSE",
                type="s",lty=2)
```

# REFERENCES


Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**(3):469–475

Babu, G. J., Canty, A. J., and Chaubey, Y. P. (2002). Application of bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, **105**(2):377–392.

Chaubey, Y. P., Li, J., Sen, A., and Sen, P. K. (2012). A new smooth density estimator for non-negative random variables. *Journal of the Indian Statistical Association*, **50**:83–104

Chaubey, Y. P. (2017). On smooth density estimation for circular data. *Invited Paper,* Presented at 61st World Statistics Congress, Marrakech July 16–21, 2017.

Chaubey, Y. P. (2018). Smooth kernel estimation of a circular density function: A connection to orthogonal polynomials on the unit circle. *Journal of Probability and Statistics,* **2018**, Article ID 5372803.

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

Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. *Rev. Statist. Appl*, **25**(3):5–42.

Feller, W. (1965). *An Introduction to Probability Theory and its Applications*, Vol. 1, Wiley: New York.

Fisher, N. I. (1989). Smoothing a sample of circular data. *Journal of structural geology*, **11**(6):775–778.

Fisher, N. I. (1995). *Statistical Analysis of Circular Data*. Cambridge University Press, London.

Jammalamadaka, S. R. and Sengupta, A. (2001). *Topics in Circular Statistics*. World Scientific, Singapore.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Education.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**(3):1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, **27**(3):832–837.

Scaillet, O. (2004). Density estimation using inverse and reciprocal inverse gaussian kernels. *Nonparametric Statistics*, **16**(1-2):217–226.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**(400):1131–1146.

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

Tang, M. (2011). *A Comparison of Two Nonparametric Density Estimators in the Context of Actuarial Loss Model*. PhD Thesis, Concordia University.

von Mises, R. (1918). Uber die'ganzzahligkeit'der atomgewicht und verwandte fragen. *Phys. Z.*, **19**:490–500.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman and Hall Ltd., London.