People's Democratic Republic of Algeria
Ministry of Higher Education and Scientifc Research
Mohamed Khider University
Departement Of Mathematics

A Thesis Presented for the Degree of DOCTOR in Mathematics

BIAS CORRECTION AT END POINTS IN KERNEL DENSITY ESTIMATION

HIND Bouredji

Examination Committee Members :

Mr. YAHIA Djebrane	Pr.	Biskra University	Chairman
Mr. SAYAH Abdallah	M.C.A.	Biskra University	Supervisor
Mr. BRAHIMI Brahim	Pr.	Biskra University	Examinator
Mr. DJEFFAL El Amir	Pr.	Batna University	Examinator

2022

Republique Algerienne Democratique et Populaire
Ministere de l'Enseignement Superieur et de la Recherche Scientifique
Universite de Mohamed Khider
Departement de Mathematiques

Mémoire présenté en vue de l'obtention du Diplôme Doctorat in Mathématique

LA CORRECTION DU BIAIS AUX POINTS LIMITES DANS L'ESTIMATION A NOYAU DE LA DENSITE

HIND Bouredji

Membres du Comite d'Examen :

Mr. YAHIA Djebrane	Pr.	Universite de Biskra	President
Mr. SAYAH Abdallah	M.C.A.	Universite de Biskra	Encadreur
Mr. BRAHIMI Brahim	Pr.	Universite de Biskra	Examinateur
Mr. DJEFFAL El Amir	Pr.	Universite de Batna	Examinateur

2022

Dedicace

To my Dear Parents

And

my Sisters and Brothers

Remerciements

I glorify Allah the Almighty for giving me courage and patience that allowed meto do this work.

First and foremost I am deeply grateful to my supervisor Prof. Sayah Abdallah, for hisvaluable support. Without his constructive comments, patience and knowledge this thesiswould have not been possible. I honestly could have not wished for a better supervisor.Besides my advisor, I would like to thank the examination committee members: Prof. YAHIADjebrane, Prof. BRAHIMI Brahim, Prof. DJEFFAL El Amir for their interest in my work andreviewing this thesis, for their insightful comments and encouragement, but also for the hardquestions which incited me to widen my research from various perspectives.

A special thank goes for my best friend Almi Nassima for all your help

I would like toexpress my greatest gratitude to the people who have helped and supported me throughoutmy doctorate.

A special thank goes to my parents, my sisters and my brothers: for all theemotional support, understanding, caring they provided, for their never-ending confidence.This thesis is for you.

Contents

List of Figures

List of Tables

Introduction

An important aspect of statistics, often neglected nowadays, is the presentation of the data back to the client in order to provide explanation and illustration of conclusions that may possibly have been obtained by other means. Density estimates are ideal for this purpose, for the simple reason that they are fairly easily comprehensible to non-mathematicians. It is a research area in statistics and has been studied extensively in this field. Oftentimes it is not possible to make such strict assumptions about the form of the underlying density function. The existing literature on density function estimation can be categorized into two broad themes: parametric and nonparametric methods. Non-parametric approaches are more appropriate in these situations. These techniques make few assumptions about the density function and allow the data to drive the estimation process more directly. Nonparametric kernel smoothing belongs to a general category of techniques for nonparametric estimations including : density, distribution, regression, quantiles.These estimators are now popular and in wide use with great success in statistical applications. Nonparametric density estimation is one of the most researched and still active areas in statistical theory, and the techniques and the theory are highly sophisticated. A lot of development in statistics has taken place

around the themes, methods, and mathematics of density estimation. Density estimation has experienced a wide explosion of interest over the last 20 years. Early results on kernel density estimation are due to Rosenblatt (1956) and Parzen (1962). Good references in this area are Silverman (1986), and Wand and Jones (1995). Kernel estimates may suffer from boundary effects. This type of boundary effect for kernel estimators of curves with compact supports is well-known in density function estimation frameworks. In the density estimation context, a various boundary bias correction methods have been proposed. Schuster (1999) and Cline and Hart (1991) considered the reflection method, which is most suitable for densities with zero derivatives near the boundaries. Boundary kernel method and local polynomial method are more general without restrictions on the shape of densities. Local polynomial method can be seen as a special case of boundary kernel method and draws much attention due to its good theoretical properties. Though early versions of these methods might produce negative estimates or inflate variance near the boundaries, remedies and refinements have been proposed, see Muller (1991), Jones (1993), Jones and Foster (1996), Cheng (1997), Zhang and Karunamuni (1998; 2000); and Karunamuni and Alberts (2005). Cowling and Hall (1996) proposed a pseudo-data method that estimates density functions based on the original data plus pseudo-data generated by linear interpolation of order statistics. Zhang et al. (1999) combined the pseudo-data, transformation and reflection method. In the regression function estimation context, Gasser and Muller (1979) identified the unsatisfactory behavior of the Nadaraya Watson regression estimator for points in the boundary region. They proposed optimal boundary kernels but did not give any formulas. However, Gasser and Muller (1979) and Muller

(1988) suggested multiplying the truncated kernel at the boundary zone or region by a linear function. Rice (1984) proposed another approach using a generalized jackknife. Schuster (1985) introduced a reflection technique for density estimation. Eubank and Speckman (1991) presented a method for removing boundary effects using a bias reduction theorem. This thesis is organized as follows :

- Chapter 1. In this chapter, we provide an overview of the essential definitions of the probability density estimation, a common problem in statistics is that of estimating a density f, existing methods to estimate on unknown density function from data can be classified into two groups, namely : parametric and nonparametric methods. Parametric methods are dependent on assumption that the functional form of the density function is specified. Non-parametic density estimators in general make no assumptions on the type of the density that produced the samples

- Chapter 2. This chapter is designed for the methods of boundary correction in kernel density estimation, we focused on the boundary effect, some methods of boundary correction have been discussed.

- Chapter 3. In this chapter, We have focused also on the bias correction at end points in kernel density estimation. Our new approach is the preferable method, because it has a good performance in most of the investigated situations.

CHAPTER 1

Probability Density Estimation

This chapter discusses the method of density estimation, it can be broadly classified into two groups: parametric density estimation and non-parametric density estimation. The first approach requires specification of a family of densities and estimation of the unknown parameter using a suitable estimation method. This approach may be prone to bias that arises from either estimation of the parameter or from incorrect specification of the probability distribution. The second approach, does not assume a specific parametric family.

1.1 Parametric Density Estimation

Parametric methods assume a certain type of density for approximating the true underlying density, which can be any type of parametric density representation like Gaussian, Laplacian, uniform, etc., or mixtures of such densities. This assump-

tion does not mean that the true underlying density has to be of this type. The idea behind this approach is to find the parameters of the chosen density type such that the observed data has the highest possible probability. Parametric methods make strict a priori assumptions about the form of the underlying density function. For instance, a parametric approach may assume the random variables have a PDF is a polynomial of a particular degree. Such assumptions significantly simplify the problem, since only the parameters of the chosen family of functions need to be determined.

1.1.1 Method of Moments

In the method of moments approach, the parameters of a probability distribution model are estimated by matching the moments of the dataset with that of the candidate model. The number of moments required corresponds to the number of unknown model parameters. Application of this method is straightforward, as closed-form expressions for the moments can be readily derived for most common distributions. However, the raw moments may be biased due to the presence of outliers and/or the lack of perfect agreement between the data and the model. To show how the method of moments determines an estimator, we first consider the case of one parameter.

Definition 1.1.1 *Let be $X_1, X_2, ...$ an independent random variables chosen according to the probability density f_X(x,θ)$associatedtoanunknownparametervalue\theta$. The common mean of the X_i, μ_X, is a function $k(\theta)$ of θ. For example, if the X_i are continuous random*

variables, then

$$\mu_X = \int_{-\infty}^{\infty} x f_X(x, \theta) dx = k(\theta)$$

The law of large numbers states that

$$\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i \to \mu_X \qquad \text{as } n \to \infty$$

Thus, if the number of observations n is large, the distributional mean, $\mu = k(\theta)$, should be well approximated by the sample mean, i.e.,

$$\overline{X} \approx k(\theta)$$

This can be turned into an estimator by setting

$$\overline{X} = k(\theta)$$

and solving for $\widehat{\theta}$. More generally, for independent random variables $X_1, X_2, ...$ chosen according to the probability distribution derived from the parameter value θ and m a real valued function, if $k(\theta) = E_\theta m(X_1)$, then

$$\frac{1}{n} \sum_{i=1}^{n} m(X_i) \to k(\theta) \qquad \text{as } n \to \infty$$

The method of moments results from the choices $m(x) = x^m$. Write

$$\mu_m = EX^m = k_m(\theta) \tag{1.1.1}$$

for the m-th moment.

Our estimation procedure follows from these 4 steps to link the sample moments to parameter estimates.

Step1. If the model has d parameters, we compute the functions k_m in equation for the first d moments

$$\mu_1 = k_1(\theta_1, \theta_2, \ldots, \theta_d), \mu_2 = k_2(\theta_1, \theta_2, \ldots, \theta_d), \ldots, \mu_d = k_d(\theta_1, \theta_2, \ldots, \theta_d)$$

obtaining d equations in d unknowns.

Step2. We then solve for the d parameters as a function of the moments.

$$\theta_1 = g_1(\mu_1, \mu_2, \ldots, \mu_d), \theta_2 = g_2(\mu_1, \mu_2, \ldots, \mu_d), \ldots, \theta_d = g_d(\mu_1, \mu_2, \ldots, \mu_d) \tag{1.1.2}$$

Step3. Now, based on the data $x = (x_1, x_2, ..., x_n)$, we compute the first d sample moments,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \ldots, \overline{x^d} = \frac{1}{n}\sum_{i=1}^{n} x_i^d.$$

Using the law of large numbers, we have, for each moment, $m = 1, ..., d$, that $\mu_m \approx \overline{x^m}$.

Step4. We replace the distributional moments μ_m by the sample moments x^m, then the solutions in (1.1.2) give us formulas for the method of moment estimators $(\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_d)$. For the data x, these estimates are

$$\widehat{\theta}_1(x) = g_1(\overline{x}, \overline{x^2}, \ldots, \overline{x^d}), \widehat{\theta}_2(x) = g_2(\overline{x}, \overline{x^2}, \ldots, \overline{x^d}), \ldots, \widehat{\theta}_d(x) = g_d(\overline{x}, \overline{x^2}, \ldots, \overline{x^d}),$$

Example 1.1.2 *Let $X_1, X_2, ..., X_n$ be a random sample with density of beta of parametre a and b*

$$f_X(x, a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \qquad x \in [0; 1].$$

where $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ *and* Γ *is the gamma function. The mean and the variance are, respectively,*

$$\mu = \frac{a}{a+b}, \qquad \sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

In this situation, we have two parameter, namely a, b*. Thus, in* $step\ 1$*, we will need to determine the first moment*

$$\mu = \bar{X} = \frac{a}{a+b}$$

$step\ 2$*, we will need to determine the second moment*

$$S^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

to find the method of moments estimator $(\hat{a}, \hat{b})$ *for* (a,b) *we need to solve the system*

$$\begin{cases} \bar{X} = \dfrac{\hat{a}}{\hat{a}+\hat{b}} \\ \\ S^2 = \dfrac{\hat{a}\hat{b}}{(\hat{a}+\hat{b})^2(\hat{a}+\hat{b}+1)} \end{cases}$$

In the following figure we use beta density of parametre $(2,4)$*, our paramters estimated by this method are* $\hat{a} = 2.0774$ *and* $\hat{b} = 4.2613$

1.1.2 Maximum Likelihood Estimation

Maximum likelihood estimation is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the

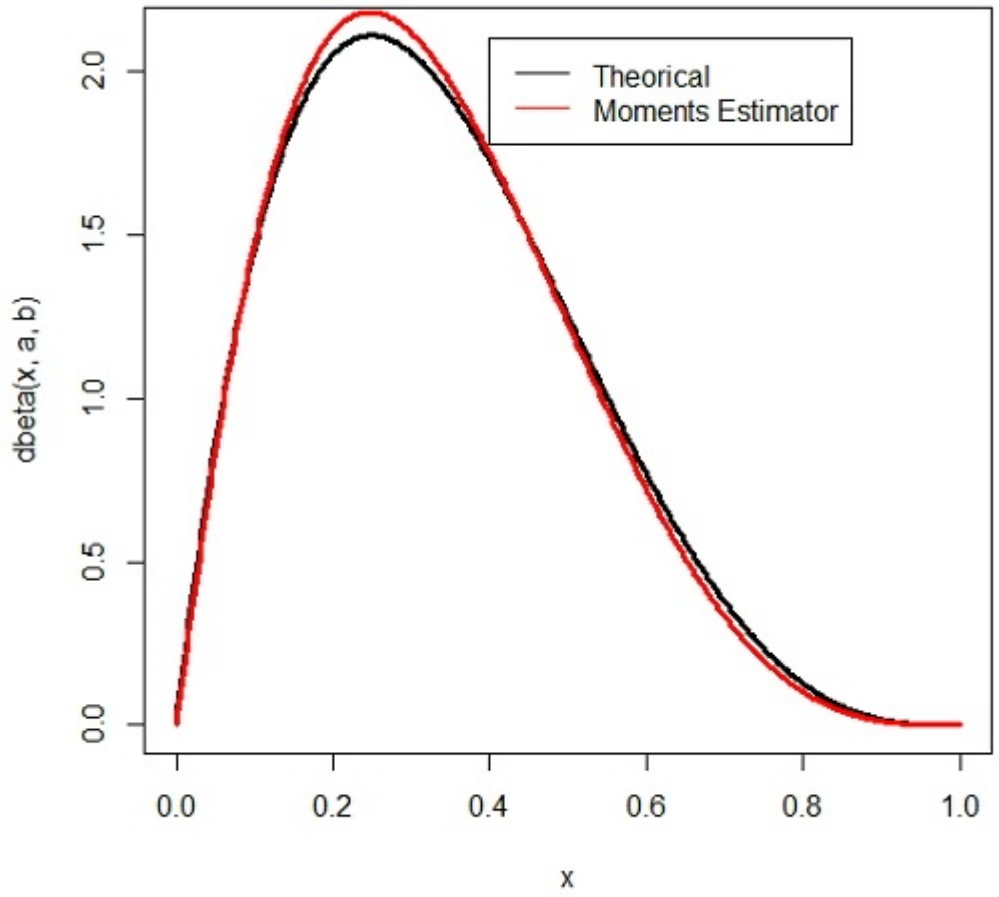

Figure 1.1: Moment estimaator of the beta density of parametre $(2, 4)$

observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.

Definition 1.1.3 *The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.*

$$L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i, \theta)$$

maximizing $L(\theta)$ with respect to θ will give us the

$$\widehat{\theta} = \arg\max \sum_{i=1}^{n} f_{X_i}(x_i, \theta)$$

Example 1.1.4 *The Pareto distribution has been used in economics as a model for a density function with a slowly decaying tail*

$$f_X(x, \theta) = \frac{\theta}{x^{\theta+1}}, \qquad x > 1$$

The log-likelihood function is

$$\begin{aligned} L(\theta) &= \textstyle\sum_{i=1}^{n} \log f_{X_i}(x_i, \theta) \\ &= \textstyle\sum_{i=1}^{n} (\log\theta - (\theta+1)\log X_i) \\ &= n\log\theta - (\theta+1)\textstyle\sum_{i=1}^{n} \log X_i \end{aligned}$$

Let the derivative with respect to θ be zero:

$$\frac{dL(\theta)}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \log X_i = 0$$

Solving the equation yields the MLE of θ :

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n} \log X_i}$$

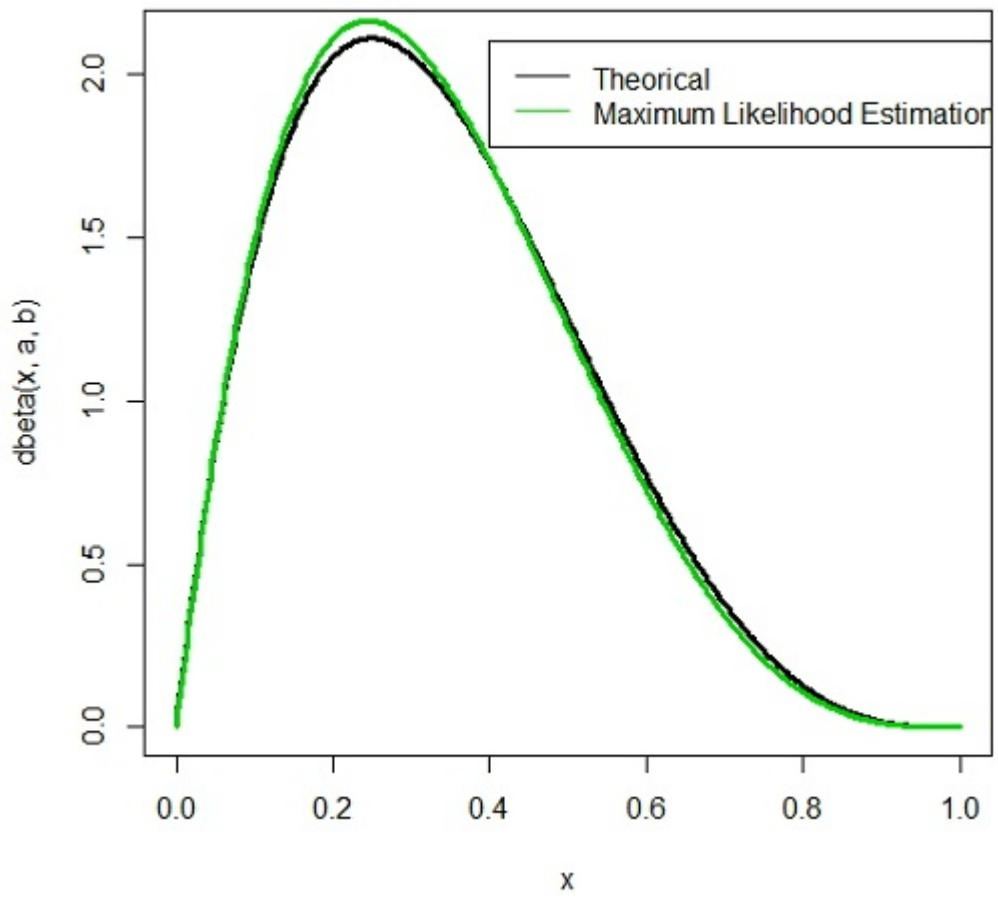

Figure 1.2: Maximum Likelihood estimator of the Beta density

We applied this method to the precedant example beta$(2, 4)$*, then we have* $\hat{a} = 2.0400$ *and* $\hat{b} = 4.1841$ *that give the following figure*

By comparing the two methods, we notice that The maximum probability method is more precise than the method of moments, as shown in the following figure

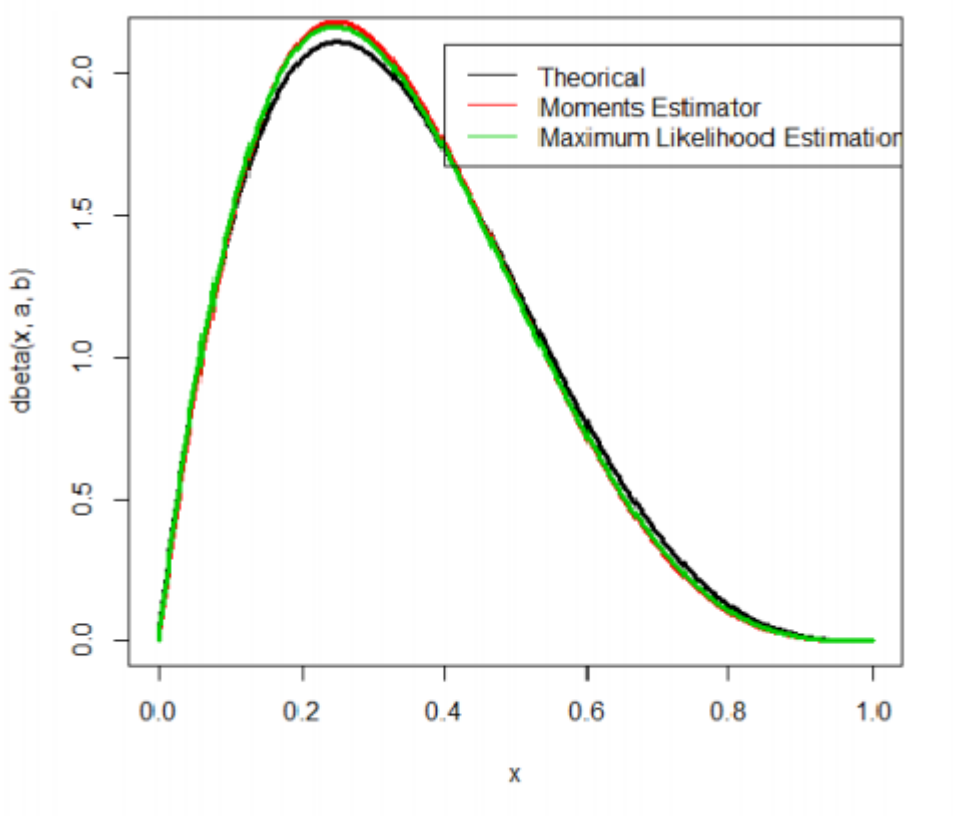

Figure 1.3: Comparing of the two methods Moment and Maximum Likelihood

1.2 Nonparametric Density Estimation

Estimation of a probability density is similar to its construction from observed data. The first approach is that of parametric estimation. The idea principle of this approach is to assume that the density f to be estimated belongs to a family of distributions having a finite number of parameters (we mainly cite the maximum likelihood method and the method of moments). Son advantage lies in the algorithmic simplicity of its implementation. By opposition, the non-parametric approach makes no a priori assumption on the membership of f to a known family of laws. The estimate therefore no longer concerns a parameter in this family of law, but directly the function itself (hence the term nonparametric). Nonparametric density estimation is an important data analytic tool which provides a very effective way of showing structure in a set of data at the beginning of its analysis. The grouping of data in the form of a frequency histogram is a classical methodology that is used in nonparametric density estimatoion, this method was introduced by John Graunt (1962). Since histograms are not smooth, Parzen (1962) devoloped the kernel density estimator proposed on (1956) by Rosenblatt, which are smoother and which converge to the true density faster.

1.2.1 Histogram

The simplest non-parametric technique for density estimation is the histogram. A histogram is appropriate for continuous data. It consists of a series of contiguous bars, the areas of which are proportional to the counts of observations that fall

within the intervals covered by the bars. Anyone who has drawn a histogram will realize that the appearance is strongly affected by the origin or anchor point at which one chooses to start the histogram, and the width of the intervals used bin width. Whallon (1987) illustrates this well. The grouping of data in the form of a frequency histogram is a classical methodology that is intrinsic to the foundations of a variety of estimation procedures. Providing useful visual information, it has served as a data presentation device, however, as a density estimation method, it has played a fundamental role in nonparametric statistics. This method is usually formed by dividing the real line into equally sized intervals often called bins B.

Definition 1.2.1 *For positive and negative integers m, the bins are of the form $[x_0 + mh, x_0 + (m+1)\,h)$, $x_0 \in \mathbb{R}$. The intervals have been chosen closed on the left and open on the right for definiteness. Then the histogram estimate at a point x is given by,*

$$f_H(x) = \frac{\mathbf{1}\{X_i \in B_i\}}{nh},\ i = 1, \ldots, n.$$

Note that, to construct the histogram, we have to choose both an origin and a bin width, it is the choise of bin width which, primarily, controls the amount of smoothing inherent in the procedure. The smoothness of the histogram estimate is controlled by the smoothing parameter h, a characteristic shared by all nonparametric cure estimators. Choosing a small bandwidth leads to a jagged estimate, while larger bandwidths tend to produce over smoothed histogram estimates. Histograms for the graphical presentation of bivariate or trivariate data present several difficulties; for example, one cannot easily draw contour diagrams to present the data, and the problems raised in the univariate case are exacerbated

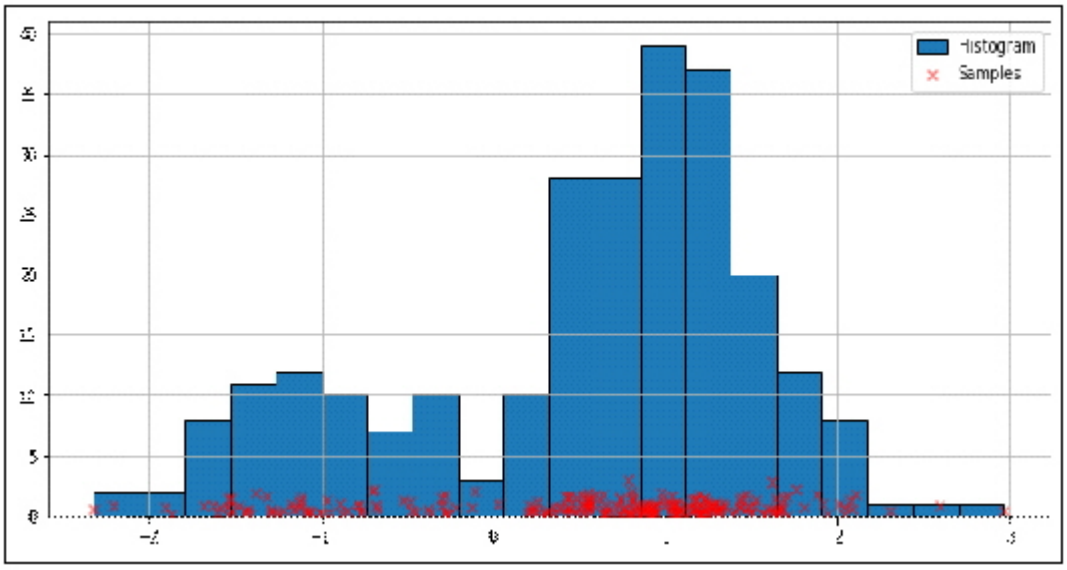

Figure 1.4: The histogram estimation for density function

by the dependance of the estimates on the choice not only of an origin but also of the coordinate direction(s) of the grid of cells. Finally, it should be stressed that, in all cases, the histogram still requires a choice of the amount of smoothing.

Lemma 1.2.2 *The basic properties of* $f_H(x)$ *are,*

- *The bias of $f_H(x)$ is*

$$\begin{aligned} E(f_H(x)) &= \frac{1}{h}\int_x^{x+h} f(t)dt \\ &= \frac{F(x+h)-F(x)}{h} \\ &= f(x)+\frac{h}{2}f^{(1)}(x)+o(h) \\ &= \frac{h}{2}f^{(1)}(x)+o(h) \underset{n\to\infty}{\to} f(x). \\ &= Bias(f_H(x)) \end{aligned}$$

- *The variance of $f_H(x)$ is*

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

- *The mean squared error* (MSE) *can be written as*

$$\begin{aligned} MSE(f_H(x)) &= E(f_H(x) - f(x))^2 \\ &= E((f_H(x)) - f(x))^2 + Var(f_H(x)) \\ &= Var(f_H(x)) + Bias^2(f_H(x)). \\ &= \left(\frac{h}{2}f^{(1)}(x)\right)^2 + o(h^2) + \frac{f(x)}{nh} + O\left(\frac{1}{n}\right) \end{aligned}$$

The histogram estimation method is the most natural and widely used because it is easily implemented. However, the density estimator provided by a histogram cannot be adapted to the fairly common situation where there is a priori information on the regularity of the density to be estimated. More precisely, if we know in advance that the density of the observed sample is, for example, twice continuously differentiable, we would naturally want to estimate this density by a function which, too, is twice continuously differentiable, for this fact, we present an other estimator.

1.2.2 Naive estimator

Definition 1.2.3 *From the definition of a probability density, if the random variable* X *has density* f*, then*

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

For any given h*, we can of course estimate* $P(x - h < X < x + h)$ *by the proportion of the sample falling in the interval* $(x - h, x + h)$*. Thus a natural estimator* f_n *of the density is gien by choosing a small number* h *and setting*

$$f_n(x) = \frac{1}{2nh}\left[\textit{no. of } X_1, \ldots, X_n \textit{ falling in } (x-h, x+h)\right]$$

we shall call this the naive estimator or the Rosenblatt estimator. To express the estimator more transparently, define the weight function w by

$$w(x) = \begin{cases} \frac{1}{2} & \textit{if } |x| < 1 \\ \\ 0 & \textit{otherwise} \end{cases} \tag{a}$$

Then it is easy to see that the naive estimator can be written

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{x - X_i}{h}\right)$$

It follows from (a) that the estimate is constructed by placing a 'box' of width $2h$ and height $(2nh)^{-1}$ on each obseration and then summing to obtain the estimate. We shall return to this interpretation below, but it is instructive first to consider a connection with histograms.

Lemma 1.2.4 *The statistical properties of f_n are,*

$$\begin{aligned} Var(f_n(x)) = \frac{1}{4nh_n^2}\big[&-F^2(x+h_n) + F(x+h_n) + F(x+h_n)F(x-h_n) \\ &- F(x-h_n) + F(x+h_n)F(x-h_n) - F(x-h_n) \\ &- F^2(x-h_n) + F(x-h_n)\big] \\ = \frac{1}{4nh_n^2}&\left[F(x+h_n) - F(x-h_n) - \left(F(x+h_n) - F(x-h_n)\right)^2\right] \end{aligned}$$

Now

$$
\begin{aligned}
MSE\left(f_n\left(x\right)\right) &= E\left[\left(\hat{f}_n\left(x\right)-f\left(x\right)\right)^2\right] \\
&= Var\left(\hat{f}_n\left(x\right)\right)+Biais^2\left(\hat{f}_n\left(x\right)\right) \\
&= \frac{1}{4nh_n^2}\left[F\left(x+h_n\right)-F\left(x-h_n\right)-\left(F\left(x+h_n\right)-F\left(x-h_n\right)\right)^2\right] \\
&+ \left[\frac{F\left(x+h_n\right)-F\left(x-h_n\right)}{2h_n}-f\left(x\right)\right]^2
\end{aligned}
$$

Assuming that f is thrice differentiable at x therefore we have

$$
\begin{aligned}
F\left(x+h_n\right)-F\left(x-h_n\right) &= 2h_nf\left(x\right)+\frac{h_n^3}{3}f''\left(x\right)+O(h_n^4) \\
MSE\left(f_n\left(x\right)\right) &= \frac{1}{4nh_n^2}\left[2h_nf\left(x\right)+\frac{h_n^3}{3}f''\left(x\right)+O(h_n^4)\right. \\
&\quad \left.-\left(2h_nf\left(x\right)+\frac{h_n^3}{3}f''\left(x\right)+O(h_n^4)\right)^2\right] \\
&\quad + \frac{1}{4h_n^2}\left[2h_nf\left(x\right)+\frac{h_n^3}{3}f''\left(x\right)+O(h_n^4)-2h_nf\left(x\right)\right]^2 \\
&= \frac{f\left(x\right)}{2nh_n}+\frac{h_n^4}{36}(f''\left(x\right))^2+o\left(\frac{1}{nh_n}+h_n^4\right).
\end{aligned}
$$

The naive estimator is not wholly satisfactory from the point of view of using density estimates for presentation. It follows from the definition that hat f is not a continuous function, but has jumps at the points $X_i \pm h$ and has zero derivative everywhere else. This gives the estimates a somewhat ragged character which is not only aesthetically undesirable, but, more seriously, could provide the untrained observer with a misleading impression. Partly to overcome this difficulty,

and partly for other technical reasons given later, it is of interest to consider the generalization of the naive estimator given in the following section.

1.2.3 Kernel estimator

The kernel method originated from the idea of Rosenblatt and Parzen dedicated to density estimation. The distribution function $F(x)$ is naturally estimated by the EDF. It might seem natural to estimate the density $f(x)$ as the derivative of $F_n(x)$, $\frac{d}{dx}F_n(x)$. but this estimator would be a set of mass point, not a density, and as such is not a useful estimate of $f(x)$.

Instead, consider a discrete derivative. For some small $h > 0$, let

$$f_R(x) = \frac{F_n(x+h) - F_n(x-h)}{2b}.$$

We can write this as

$$\begin{aligned} f_R(x) &= \frac{1}{2nh}\sum_{i=1}^{n} I(x-h \le X_i \le x+h) \\ &= \frac{1}{2nh}\sum_{i=1}^{n} I\left(\frac{|X_i - x|}{h} \le 1\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right), \end{aligned}$$

where

$$k(t) = \begin{cases} \frac{1}{2}, & |t| \le 1 \\ \\ 0, & |t| > 1. \end{cases}$$

is the uniform density function on $[-1, 1]$. $f_R(x)$ is a special case of what is called a Rosenblatt-Parzen kernel density estimator is as follows (Wand and Jones 1995;

Silverman 1996):

$$f_R(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right),$$

where $X_1, ..., X_n$ be independent random variables identically distributed which are drawn from a continuous distribution $F(x)$ with density function $f(x)$. n is the sample size, $h := h_n$ ($h \to 0$ and $nh \to \infty$) is the smoothing parameter, called the bandwidth, which controls the smoothness of the estimator, $k(.)$ is the weighting function called the kernel function. When $k(.)$ is symmetric and unimodal function and the following conditions are fulfilled:

1. $k(t) \geq 0, \forall t \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} k(t)dt = 1$, hence k is a density function.
3. $k(-t) = k(t)$, hence k is a symmetric function.
4. $\int_{-\infty}^{\infty} tk(t)dt = 0$.
5. $\int_{-\infty}^{\infty} t^2k(t)dt < \infty$.

The basic idea of a kernel estimation of densities that each data point, x_i, is associated with a kernel k, and the bandwidth h are summed to get the KDE, the population choice of kernel and bandwidth are discuted in the following subsuction . The function f_R is called the kernel density estimator or the Parzen-Rosenblatt estimator.

Theorem 1.2.5 *We discuss some of the numerical properties of the kernel estimator f_R, first, if $k(x)$ is non-negative then it is easy to see that $f_R \geq 0$: However, this is not*

guarenteed if k is a higher-order kernel. That is, in this case it is possible that $f_R < 0$ for some values of x. Second, f_R integrates to one, to see this, first note that if k is a kernel, then f_R is a density,we have

$$\begin{aligned}\int_{\mathbb{R}} f_R(x)dx &= \int_{\mathbb{R}} \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) dx \\ &= \frac{1}{nh}\sum_{i=1}^{n} \int_{\mathbb{R}} k\left(\frac{X_i - x}{h}\right) dx \\ &= \int_{\mathbb{R}} k\left(t\right) dt \\ &= 1\end{aligned}$$

Third, we can also calculate the numerical moments of the density f_R: Again using the change of-variables $t = \frac{X_i - x}{h}$; the mean of the estimated density is

$$\begin{aligned}\int_{\mathbb{R}} x f_R\left(x\right) dx &= \frac{1}{n}\int x\frac{1}{h}k\left(\frac{X_i - x}{h}\right) dx \\ &= \frac{1}{n}\int \left(X_i + th\right) k\left(t\right) dt \\ &= \frac{1}{n}X_i \int k\left(t\right) dt \\ &+ \frac{1}{n}h\int tk\left(t\right) dt \\ &= \frac{1}{n}X_i\end{aligned}$$

the sample mean of the X_i. The second moment of the estimated density is

$$\begin{aligned}\int_{\mathbb{R}} x^2 f_R(x)\,dx &= \frac{1}{n}\int x^2 \frac{1}{h}k\left(\frac{X_i - x}{h}\right)dx \\ &= \frac{1}{n}\int (X_i + th)^2 k(t)\,dt \\ &= \frac{1}{n}X_i^2 + \frac{2}{n}X_i h\int k(t)\,dt \\ &+ \frac{1}{n}h^2\int t^2 k(t)\,dt \\ &= \frac{1}{n}X_i^2 + h^2\int t^2 k(t)\,dt\end{aligned}$$

The bias of f_R is then

$$Bias\left(f_R(x)\right) = \frac{f^{(2)}(x)}{2}h^2\int t^2 k(t) + O(h^(3)) \tag{1.2.1}$$

and

$$Var\left(f_R(x)\right) = \frac{f(x)}{nh}h^2\int k^2(t) + o(nh)^{(-1)} \tag{1.2.2}$$

Then from (1.2.1) and (1.2.2) we have

$$MSE\left(f_R(x)\right) \sim \frac{1}{4}f^{(2)}(x)^2 h^4(\int t^2 k(t))^2 + \frac{f(x)}{nh}f(x)\int k^2(t) \tag{1.2.3}$$

A good introduction to kernel density estimation with an interesting collection of its use in data analysis is given by the monograph of Silverman (1986). It turns out that the choise of h is much more important for the bahaviour of $f_R(x)$ than the choise of K.Small values of h make the estimate look "wiggly" and show spurious features, whereas to big of h will lead to an estimate which is too smooth in the sense that it is too biased and may not reveal structural features .

Kernel Functions and Bandwidth

- Kernel Functions. Much of the first decade of theoretical work focused upon various aspects of estimation properties relating to the characteristics of a kernel. The quality of a density estimate is now widely recognized to be primarily determined by the choice of smoothing parameter, and only in a minor way by the choice of kernel. Thus the topic could be de-emphasized. However, there has been a recent spurt of useful research on kernel design in special situations. While many potential hazards face the user of density estimation (for example, underestimating the smoothness of the unknown density), the specification of desired properties for the kernel is entirely at the disposal of the worker, who should have a good understanding of the following results.

Kernel	$k(t)$	Eff
Epanechnikov	$\frac{3}{4}(1-t^2)$ for $\|t\|<1$	1
Biweight	$\frac{15}{16}(1-t^2)^2$ for $\|t\|<1$	0.994
Triangular	$1-\|t\|$ for $\|t\|<1$	0.986
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$	0.951
Tricube	$\frac{70}{81}\left(1-\|t\|^3\right)^3$ for $\|t\|<1$	0.998
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}t\right)$ for $\|t\|<1$	0.999
Logistic	$\frac{1}{\exp(t)+2+\exp(-t)}$	0.887
Triweight	$\frac{35}{32}(1-t^2)^3$ for $\|t\|<1$	0.987
Uniform	$\frac{1}{2}$ for $\|t\|<1$	0.930

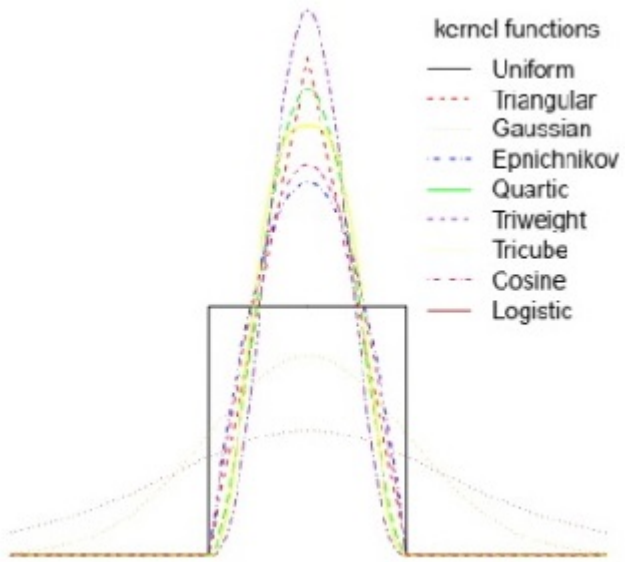

Figure 1.5: Some commoly used kernel functions

- Bandwidth parameter. Selecting an appropriate bandwidth for a kernel density estimator is of crucial importance, and the purpose of the estimation may be an influential factor in the selection method. In many situations, it is sufficient to subjectively choose the smoothing parameter by looking at the density estimates produced by a range of bandwidths. One can start with a large bandwidth, and decrease the amount of smoothing until reaching a "reasonable" density estimate. However, there are situations where several estimations are needed, and such an approach is impractical. An automatic procedure is essential when a large number of estimations are required as part of a more global analysis, the following figure show this phonemena for rundam value of h

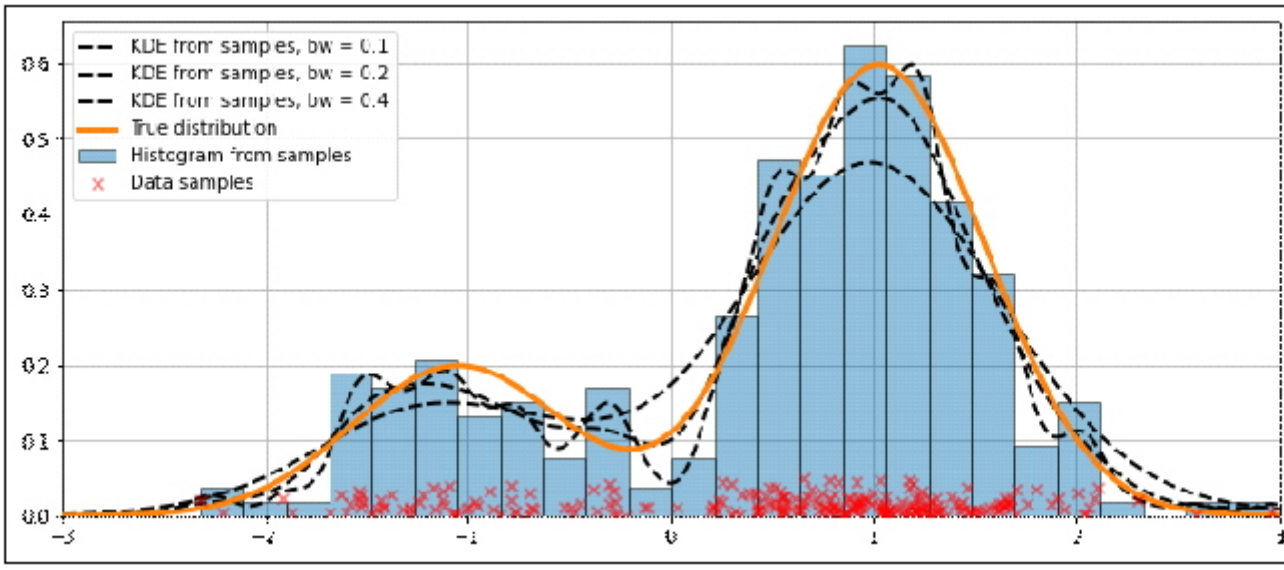

Figure 1.6: Comparing of appropriate bandwidth for a density estimator

The problem of selecting the smoothing parameter for kernel estimation has been explored by many authors, and no procedure has yet been considered the best in every situation. Automatic bandwidth selection methods can basically be divided in two categories: plug-in and classical. Plug-in methods refer to those that find a pilot estimate of f, sometimes using a pilot estimate of h, and "plug it in" the estimation of $MISE$. Classical methods, such as cross-validation, Bootstrap and another methods are basically extensions of methods used in parametric modeling. We present in more detail the reference method and the most used automatic bandwidth selection procedures.

Plug-in methods

It is well known that plug-in bandwidth estimators tend to select larger bandwidths when compared to the classical estimators. They are usually tuned by arbitrary specification of pilot estimates and most often produce over smoothed results when the smoothing problem is difficult. On the other hand, smaller bandwidths tend to be selected by classical methods, producing under smoothed results. The goal of a selector of the smoothing parameter is to make that decision purely from the data, finding automatically which features are important and which should be smoothed away The slow rate of convergence encouraged much research on faster converging methods.

– Optimal bandwidth. We Consider the following $AMISE$ version of the function of a probability density $f(x)$.

$$h_{MISE} = \left(\frac{R(K)}{\mu_2^2(K) R(f^{(2)})} \right)^{1/5} n^{-1/5}$$

– Rule Of Thumb. If we choose f as being the normal distribution of mean 0 and variance σ^2 we will have:

$$R\left(f^{(2)}\right) = \int \left(f^{(2)}(x)\right)^2 dx = \frac{3}{8\sqrt{\pi}} \sigma^{-1/5} \qquad \text{(rot)}$$

Moreover, if k is a Gaussian kernel, then the value for the h_{rot} is obtained by substituting this kernel and the value $R\left(f^{(2)}\right)$ obtained in the formula (2.29)

Classical methods

Cross-validation is a popular and readily implemented heuristic for selecting the smoothing parameter in kernel estimation.

– Unbiased cross-validation. Rudemo (1982) and Bowman (1984) proposed a so-called unbiased cross-validation (UCV) in kernel density estimator, is probably the most popular and best studied one. An adaptation of unbiased cross-validation is proposed by Wolfgang et all (1990) for bandwidth choice in the r^{th} derivative of kernel density estimator. The essential idea of this methods, it aims to estimate h the minimizer of $UCV\,(h)$ given by

$$UCV\,(h) = \frac{R\,(k)}{nh} + \frac{1}{n\,(n-1)\,h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} k^2\left(\frac{X_j - X_i}{h}\right)$$

The minimization criterion is defined by:

$$h_{ucv} = \arg\min_{h>0} UCV\,(h)$$

– Biased cross-validation. Biased cross-validation was proposed by Scott and George (1987), which has as its immediate target the $AMISE$.

$$AMISE\,(f_n) = \frac{R\,(k)}{nh} + \frac{h^4}{4}\mu_2^2\,(k)\,R\left(f^{(2)}\right)$$

– Complete cross-validation.

$$CCV\,(h) = R\left(\hat{f}_h\right) - \overline{\theta}_0\,(h) + \frac{1}{2}\mu_2\,(k)\,h^2\overline{\theta}_1\,(h) + \frac{1}{24}\left(6\mu_2^2\,(k) - \delta\,(k)\right)h^4\overline{\theta}_2\,(h)$$

where

$$R\left(\hat{f}_h\right) = \frac{R\left(k\right)}{nh} + \frac{1}{n\left(n-1\right)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} k\left(\frac{X_j - X_i}{h}\right)$$

and

$$\overline{\theta}_0\left(h\right) = \frac{1}{n\left(n-1\right)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} k\left(\frac{X_j - X_i}{h}\right)$$

with

$$\delta\left(k\right) = \int_{\mathbb{R}} x^4 k\left(x\right) dx$$

– Modified cross-validation

$$MCV\left(h\right) = \frac{R\left(k\right)}{nh} + \frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\psi\left(\frac{X_j - X_i}{h}\right)$$

where

$$\psi\left(t\right) = \left(k * k - k - \frac{\mu_2\left(k\right)}{2}k^{(2)}\right)\left(t\right)$$

– Maximum likelihood cross-validation

$$MLCV\left(h\right) = \left(n^{-1}\sum_{i=1}^{n}\log\left[\sum_{j\neq i} k\left(\frac{X_j - X_i}{h}\right)\right] - \log\left[\left(n-1\right)h\right]\right)$$

$$h_{mlcv} = \arg\max_{h>0} MLCV\left(h\right)$$

Bootstrap

A methodology that has been recently explored is that of selecting the bandwidth using bootstrap. It focuses on replacing the MSE by MSE^*, a bootstrapped version of MSE, which can be minimized directly. Some authors resample from a

subsample of the data $X_1, ..., X_n$, others replace from a pilot density based on the data, more precisely, from

$$\widetilde{f}_n^b = \frac{1}{nb_n} \sum_{i=1}^{n} L\left(\frac{x - X_i}{b_n}\right)$$

where L is another kernel and b_n is a pilot bandwidth. Since the bandwidth choice reduces to estimating s in $h = n^{-1/5}s$, Ziegler introduces

$$f_{n,s}^*(x) = \frac{1}{n^{4/5}s} \sum_{i=1}^{n} k\left(\frac{x - X_i^*}{n^{-\frac{1}{5}}s}\right)$$

and obtain $MSE_{n,s}^*(x) = E^*\left(\left(f_{n,s}^*(x) - \widetilde{f}_n^b(x)\right)^2\right)$. The proposed bandwidth is

$$h_n = n^{-1/5} \arg\min_s MSE_{n,s}^*$$

1.2.4 Variable kernel method

The estimate is constructed similary to the classical kernel estimate, but the scale parameter of the bumps placed on the data points is allowed to vary from one data point to another. Let k be a kernel function and τ a positive integer. Define $d_{j,\tau}$ to be the distance from X_j to the τ th nearst point in the set comprising the other $n - 1$ data points. Then the variable kernel estimate with smoothing parameter h is defined by

$$f_V(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{hd_{j,\tau}} k\left(\frac{x - X_j}{hd_{j,\tau}}\right)$$

The window width of the kernel placed on the point X_j is proportional to $d_{j,\tau}$ so that data points in regions where the data are sparse will have flatter kernels

associated with them. For any fixed τ, the overall degree of smoothing will depend on the parameter h. The choise of k determines how responsive the window width choice will be to very local detail.

1.2.5 General weight function estimators

It is possible to define a general class of density estimators which includes several of the estimators discussed above. Suppose that $\omega(x, y)$ is a function of two arguments, which in most cases will satisfy the conditions

$$\int_{-\infty}^{+\infty} \omega(x, y)\, dy = 1 \tag{1.2.4}$$

and

$$\omega(x, y) \geq 0 \text{ for all } x \text{ and } y. \tag{1.2.5}$$

We should think of ω as being defined in such a way that most of the weight of the probability density $\omega(x, .)$ falls near x. An estimate of the density underlying the data may be obtained by putting

$$f_{GW}(t) = \frac{1}{n} \sum_{i=1}^{n} \omega(X_i, t) \tag{1.2.6}$$

We shall refer to estimates of the form (1.2.6) as general weight function estimates. It is clear from (1.2.6) that the conditions (3.2.13) and (1.2.5) will be sufficient to ensure that f_{GW} is a probability density function, and that the smoothness properties of f_{GW} will be inherited from those of the functions $\omega(x, .)$.

To obtain the histogram as a special case of (1.2.6), set

$$\omega(x,y) = \begin{cases} \dfrac{1}{h(x)} & \text{if } x \text{ and } y \text{ fall in the same bin} \\ 0 & \text{otherwise} \end{cases}$$

where $h(x)$ is the width of the bin containing x. The kernel estimate can be obtained by putting

$$\omega(x,y) = \frac{1}{h} k\left(\frac{y-x}{h}\right).$$

1.2.6 Multivariate kernel density estimation

Kernel density estimation for multivariate data is an important technique that has a wide range of applications. However, its widespread usefulness has been limited by the difficulty in computing an optimal data-driven bandwidth.

Definition 1.2.6 *Let $X = (X_1, \ldots, X_d)^t$ denote a d-dimensional random vector with density $f(x)$ defined on $\mathbb{R}^d$, and let $\{x_1, \ldots, x_n\}$ be an independent random sample drawn from $f(x)$. The general form of the kernel estimator of $f(x)$ is,*

One estimator we consider is the well known kernel density estimator (KDE) for f defined as

$$f_H(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K\left(H^{-1}(x - \mathbf{X}_i)\right)$$

where $K(.)$ is a multivariate kernel function, and H is a symmetric positive definite $d \times d$ matrix known as the bandwidth matrix verify $R^d \to \mathbb{R}^+$ qui vérifie les conditions suivants:

(k_1) $\int\limits_{\mathbb{R}^d} K(t)\,dt = 1$

(k_2) $\int\limits_{\mathbb{R}^d} tK(t)\,dt = \overrightarrow{0} \Leftrightarrow \left(\int\limits_{\mathbb{R}^d} t_i K(t)\,dt\right)_{1\leq i\leq d} = 0_{\mathbb{R}^d}$

(k_3) $\int\limits_{\mathbb{R}^d} tt^T K(t)\,dt = \mu_2(K)\,\mathbf{I}_d > 0$ *avec* $\mathbf{I}_d$ $estl'identitdans$ R^d.

(k_4) $\int\limits_{\mathbb{R}^d} K(t)^2\,dt = \|K\|_2^2 < \infty$

In practice, we use the following kernels

The produit kernel: $K(t) = \prod\limits_{j=1}^{d} k_j(t_j)$.

The spheric kernel: $K(t) = C_d \times k\left(\left(t^T t\right)^{1/2}\right)$ where C_d : the volum of the unit dimension d.

Nonparametric estimation of unknown densities on partially or totally bounded supports, with or without correlation in its multivariate components, is a recurrent practical problem. Because of symmetry, the multivariate classical or symmetric kernels, not depending on any parameter, are not appropriate for these densities. In fact, these estimators give weights outside the support causing a bias in boundary regions.

Lemma 1.2.7 *The bias of* f_H *is*

$$\begin{aligned} bais\left[f_H(x)\right] = \mathbf{E}\left[f_H(x) - f(x)\right] &= \frac{1}{2} trace\left(\int\limits_{\mathbb{R}^d} t^T tK(t)\,dt\left(H^T \mathcal{H}(x) H\right)\right) + o\left(\|Ht\|_2^2\right) \\ &= \frac{1}{2}\mu_2(K)\, trace\left(H^T \mathcal{H}(x) H\right) + o\left(\|Ht\|_2^2\right) \\ &= \frac{1}{2}\mu_2(K)\, trace\left(H^T \mathcal{H}(x) H\right) + \int\limits_{\mathbb{R}^d} o\left(\|Ht\|_2^2\right) K(t)\,dt \end{aligned}$$

CHAPTER 2

Boundary correction methods in kernel density estimation

The performance of the kernel density estimator at the boundary points, i.e. for $x \in [0; h) \cup (a - h; a]$, diffiers from the interior points due to so-called "boundary effiects" that occur in nonparametric curve estimation problems. Assume that the density has support $[0, \infty)$ and its second derivative exists and is continuous at x, then it can be shown that the bias of $f_n(x)$ is of order $O(h)$ instead of $O(h^2)$ at boundary points, for $x = ch$, the bias of the kernel estimator has the form:

$$Bias\left(f_R\left(x\right)\right) = -f(x)\int_c^1 k\left(t\right)dt - hf^{(1)}(x)\int_{-1}^c tk\left(t\right)dt + \frac{h^2}{2}f^{(2)}(x)\int_{-1}^c t^2k\left(t\right)dt + o\left(h^2\right),$$

In the past, Many boundary or endpoint estimation problems in statistics are closely related to problems involving nonparametric curve estimation. The methods used are generally biased, and in fact the sign or direction (in the case of spatial problems) of bias is generally known. However, the relative error of

bias estimators is typically of larger order than would be found in the related curve estimation setting, flowing to marked asymmetries inherent to boundary estimation. Consequently, an idea on how to include boundary corrections in these estimators is presented. The first statement implies that the density has a support which is bounded on the left hand side. Without loss of generality the support is set to be $[0, \infty)$. Nonparametric kernel density estimation is now popular and in wide use with great success in statistical applications. The reflection method is specifically designed for the case $f^{(1)}(0) = 0$ where $f^{(1)}$ denotes the first derivative of f. The boundary kernel method is more general than the reflection method in the sense that it can adapt to any shape of density. These included a boundary kernel and its close counterpart the local linear fitting method, the transformation and reflection based method given by Zhang et al. (1999), Jones and Foster's (1993) nonnegative adaptation estimator, Cowling and Hall's (1996) pseudo-data method, and a recent estimator due to Hall and Park (2002) based on a transformation of the data "inside" the kernel.

2.1 The reflection method

The reflection method is one of the fairly frequently methods applied in practice of bias reduction in the kernel density estimation. The modification of the classical kernel density estimator consists in isolating that part of the kernel function which is outside the interval of the support of the random variable and then on its symmetrical reflection. This reflection is done in relation to the boundary of the support. It can be shown that the estimator taking into account the reflection

method of the kernel function has a support which is the same as the support of the random variable. The kernel density estimator with the reflection kernel function for a random variable with the support $[0, \infty)$ is of the form:

$$f_{Ref}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{k\left(\frac{x - X_i}{h}\right) + k\left(\frac{x + X_i}{h}\right)\right\}$$

Theorem 2.1.1 *The bias and the variance of the estimator are,*

- *The bias:*

$$Bias\left(f_{Ref}(x)\right) = \frac{h^2}{2} f^{(2)}(x) \int t^2 k(t)\, dt + o\left(h^2\right)$$

- *The variance:*

$$Var\left(f_{Ref}(x)\right) = \frac{1}{nh} f(x) \int k^2(t)\, dt + o(nh)^{-1}$$

Remark 2.1.2 *f_{Ref} is consistent, but the bias is of order $O(h)$ near the boundary.*

2.2 The transformation method

The transformation idea is based on transforming the original data $X_1, ..., X_n$ to $g(X_1), ..., g(X_n)$, where g is a non-negative, continuous and monotonically increasing function from $[0, 1)$ to $[0, 1)$. Consists of a three-step process. First, a transformation g is selected from a parametric family so that the density of $Y = g(X)$ has a first derivative that is approximately equal to 0 at the boundaries of its support. Next, a kernel estimator with reflection is applied to the Y_i's. Finally, this estimator is converted by the change of variables formula to obtain an estimate of f definied by

$$f_{Tr}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - g(X_i)}{h}\right),$$

Note this isn't really estimating the density function of X, but instead of $g(X)$.

2.2.1 Locally Adaptive Transformation Estimator

For convenience, we shall assume that the unknown probability density function f has support $[0;1)$, and consider estimation of f based on a random sample $X_1, ..., X_n$ from f. Our transformation idea is based on transforming the original data $X_1, ..., X_n$ to $g(X_1), ..., g(X_n)$, where g is a non-negative, continuous and monotonically increasing function from $[0;1)$ to $[0;1)$. Based on the transformed data, we now deflne,

$$f_{LATr}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - g(X_i)}{h}\right) / \int_{-1}^{c} k(t)dt. \tag{2.2.1}$$

Lemma 2.2.1 *Let f_{LATr} be deflned by (2.2.1). Assume that $f^{(2)}$ and $g^{(2)}$ exist and are continuous on $[0;1)$, where $f^{(i)}$ and $g^{(i)}$ denote the i-th derivative of f and g, respectively, with $f^{(0)} = f$; $g^{(0)} = g$. Further assume that $g(0) = 0$ and $g^{(1)}(0) = 1$. Then for $x = ch$, we have*

$$
\begin{aligned}
Bias\left(f_{LATr}\left(x\right)\right) = & \frac{-h}{\int_{-1}^{c} k(t)dt}\left\{f(0)g^{(2)}(0)\int_{-1}^{c}(c-t)k\left(t\right)dt + f^{(0)}\int_{-1}^{c} tk\left(t\right)dt\right\} \\
& + \frac{h^2}{2\int_{-1}^{c} k\left(t\right)dt}\left\{-f^{(2)}(0)c^{(2)}\int_{-1}^{c} k\left(t\right)dt + \int_{-1}^{c}(t-c)^{(2)}k\left(t\right)dt\right\} \\
& \times \left[f^{(2)}\left(0\right) - f(0)g^{(3)}(0) - 3g^{(2)}(0)\left[f^{(1)}\left(0\right) - f\left(0\right)g^{(2)}(0)\right]\right] \\
& + o\left(h^2\right),
\end{aligned}
\tag{2.2.2}
$$

and

$$
\begin{aligned}
Var\left(f_{LATr}\left(x\right)\right) = & \frac{f(x)}{(\int_{-1}^{c} k\left(t\right)dt)^2 nh^{-1}}\int_{-1}^{c} k^{(2)}\left(t\right)dt \\
& + o\left(\frac{1}{nh}\right).
\end{aligned}
\tag{2.2.3}
$$

Note that the leading term of the variance of f_{LATr} is not afiected by the transformation g. When $c = 1$, $Var\left(f_{LATr}\left(x\right)\right) = f(x)/nh\int_{-1}^{1} k^{(2)}\left(t\right)dt + o(1/nh)$, which is exactly the expansion of the interior variance of the traditional estimator.

2.2.2 Double transformation Estimator

The reflection estimator computes the estimate density based on the original and the reflected data points. Unfortunately, this does not always yield a satisfying result since this estimator enforces the shoulder condition and still contains a bias

of order h if the density does not fulfill this condition. The generalized reflection and transformation density estimators is given by

$$f_{DT}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{x - g_1(X_i)}{h}\right) + K\left(\frac{x + g_2(X_i)}{h}\right)\right\} \qquad x \geq 0 \qquad (2.2.4)$$

where g_1 and g_2 are transformations that need to be determined.

Lemma 2.2.2 *Assume that g_i^{-1} exists, $g_i(0) = 0$; $g_i^{(1)}(0) = 1$, and that $g_i^{(2)}$ and $g_i^{(3)}$ exist and are continuous on $[0; 1)$, where $g_i^{(j)}$ denotes the jth-derivative of g_i, with $g_i^{(0)} = g_i$ and g_i^{-1} denoting the inverse function of $g_i; i = 1, 2$. Suppose that $f^{(j)}$, the j-th derivative of f, exists and is continuous on $[0; 1), j = 0, 1, 2$, with $f^{(0)} = f$. Then the bias and variance of (2.2.4) are given by, for $x = ch$,*

$$
\begin{aligned}
Bias f_{DT}(x) = h \Bigg[& 2f^{(1)}(0) \int_c^1 (t-c)k(t)\,dt - g_1^{(2)}(0) f(0) \left[\int_c^1 (t-c)k(t)\,dt \right] \\
& - g_2^{(2)}(0) f(0) \int_c^1 (t-c)k(t)\,dt \Bigg] \\
& + \frac{h^2}{2} \Bigg[f^{(2)}(0) \int_{-1}^1 t^2 k(t)\,dt \\
& - \left\{ g_1^{(3)}(0) f(0) - 3 g_1^{(2)}(0) \left[f^{(1)}(0) - f(0) g_1^{(2)}(0) \right] \right\} \int_c^1 (t-c)^2 k(t)\,dt \\
& - \left\{ g_2^{(3)}(0) f(0) - 3 g_2^{(2)}(0) \left[f^{(1)}(0) - f(0) g_2^{(2)}(0) \right] \right\} \int_{-1}^c (t-c)^2 k(t)\,dt \Bigg] \\
& + o(h^2)
\end{aligned}
\tag{2.2.5}
$$

and

$$
Var f_{DT}(x) = \frac{f(0)}{nh} \left[2 \int_c^1 k(t) k(2c-t)\,dt + \int_{-1}^1 k^2(t)\,dt \right] + o\left(\frac{1}{nh}\right) \tag{2.2.6}
$$

Note that the contribution of g_1 on the bias vanishes as $c \longrightarrow 1$.

2.3 The pseudo-data method

The pseudo-data method estimator is defined (see Cowling and Hall (1996)), this generates data beyond the left endpoint of the support of the density. Generates

some extra data $X_{(i)}$' using what they call the "three-point-rule", which are then combined with the original data X_i 's to form a kernel type estimator.

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

where,

$$X_{(i)} = 5X_{(i/3)} - 4X_{(2i/3)}) + \frac{10}{3}X_{(i)}, \quad i = 1, 2, ..., n$$

and $X_{(i)}$ is the ith-order statistic of sample $X_1, ..., X_n$ and m is an integer such that $nb < m < n$

2.4 The boundary kernel method

The boundary kernel method is more general than the reflection method in the sense that it can adapt to any shape of density. However, a drawback of this method is that the estimates might be negative near the endpoints; especially when $f(0) \approx 0$.The boundary kernel and related methods usually have low bias but the price for that is an increase in variance. The boundary kernel estimator with bandwidth variation is defined (see Zhang and Karunamuni (1998)) as

$$f_{Bou}(x) = \frac{1}{nh}\sum_{i=1}^{n} K_{(c)}\left(\frac{x - X_i}{h}\right)$$

2.5 The local linear method

Special case of the boundary kernel method that is thought of by some as a simple, hard to beat, defaut approach parthy because of "optimal" theoretical properties in

the boundary kernel implicit in local linear fitting.

$$f_{LL}(x) = \frac{S_{n,2}(x) T_{n,0}(x) - S_{n,1}(x) T_{n,1}(x)}{S_{n,2}(x) S_{n,0}(x) - S_{n,1}(x)^2}$$

where

$$S_{n,j}(x) = \sum_{i=1}^{n} k\left(\frac{x - X_i}{n}\right)(X_i - x)^j \quad j = 0, 1, 2$$

and

$$T_{n,j}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{n}\right)(X_i - x)^j c_i \quad j = 0, 1$$

2.6 The cut-and-normalized method

Due to Gasser and Müller (1979), a very naive correction could then be to divide the original estimator by this factor $\int_{-1}^{c} k(t)\, dt$. The order of the bias is then h, which still is not very satisfying since in $[h, \infty)$ it becomes of order h^2. The goal is to achieve such an order in the boundary interval. This is a local correction since the integral depends on the relative position of x with respect to the bandwidth h

$$f_{CN}(x) = \frac{1}{nh} \frac{1}{\int_{-1}^{c} k(t)\, dt} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \qquad x \geq 0$$

Theorem 2.6.1 *If f is continuously differentiable in a neighbourhood of x, we have*

- The bias:

$$Bias f_{CN}(x) = \frac{f^{(2)} h^2}{2\int_{-1}^{c} k(t)\, dt}(x) \int_{-1}^{1} t^2 k(t)\, dt + o\left(h^2\right),$$

and

- The variance:

$$Var f_{CN}(x) = \frac{f(x)}{nh \int_{-1}^{c} k(t)\, dt} \int_{-1}^{1} k^2(t)\, dt + o\left(\frac{1}{nh}\right).$$

CHAPTER 3

Bias correction at end points in kernel density estimation

In this paper, we propose a new approach of boundary correction for kernel density estimation with the support $[0, 1]$, in particular at the right endpoints and we derive the theoretical properties of this new estimator and show that it asymptotically reduce the order of bias at the boundary region, whereas the order of variance remains unchanged. Our Monte Carlo simulations demonstrate the good finite sample performance of our proposed estimator. Two examples with real data are provided.

3.1 Introduction

Suppose we observe n independent identically distributed aleatoire random variables, with unknown continuous density function f. The kernel density

estimator which presented by Rosenblatt [19] then developed by Parzen [8], is defined as,

$$f_R(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}, \tag{3.1.1}$$

where h is a positive smoothing parameter, called the bandwidth, in which $h \to 0$ and $nh \to \infty$ when $n \to \infty$, and k is the kernel function with compact support $[-1, 1]$, satisfying the following conditions,

$$k(t) \geq 0, k(t) = k(-t), \int_{-1}^{1} k(t)\, dt = 1, 0 \neq \int_{-1}^{1} t^2 k(t)\, dt < \infty. \tag{3.1.2}$$

Let introduce the notation,

$$\int_{-1}^{1} t^j k(t)\, dt = \mu_j, j = 1, 2, 3, \tag{3.1.3}$$

to be more precise $\mu_1 = \mu_3 = 0$ since k is symmetric. Best reference in this area is Silverman [4] and Wand and Jones [48]. With appropriate choice of h, we can divided the support of the density onto regions, the intervals $[0, h)$ and $(1 - h, 1]$ are called the left and the right boundary region respectively and the interior region formed by the interval $[h, 1 - h]$. The performance of the kernel density estimator at least in one side of the support $(x \in [0, h) \cup (1 - h, 1])$, differs from the interior points due to so-called boundary problems and the region formed by the points with boundary problems is called the boundary region.

To remove these boundary effects at the left region $(x \in [0, h))$, a diversity of methods have been developed during the past two decades. Among them the reflection method (Schuster [9]), the transformation method (Marron and Ruppert [12]), the boundary kernel method (Jones [16]), the pseudo-data method (Cowling and Hall [38]), the local linear method (Zhang and Karunamuni [28]).

As the boundary kernel density estimator could yield negative point estimates, Jones and Foster [22] propose much simpler nonnegative boundary corrected estimators which are analogues of the wide class of simple. Karunamuni and Alberts [25] proposed a new general method generates a class of boundary corrected estimators possess desirable properties such as local adaptivity and non-negativity, in addition to this work, Karunamuni and Alberts [26] constructed a new technique based on a data transformation that depends on the point of estimation. In a very exciting work, Zhang and al [30] expected a new method of boundary correction for kernel density estimation, their approach is an amount of generalized reflection method involving reflecting a transformation of the data.

In this paper, we focus on the boundary bias problem in the right side of the support $(1-h,1]$, when the true density supported with endpoints one, the kernel density estimator has the well-known boundary problem. More specifically, we assume that $f^{(j)}$, the j^{th} derivative of f, exists and is continuous on a neighborhood of x, $(j=0,1,2,3)$, with $f^{(0)}=f$, then for $x=1-ch, c\in[0,1[$,

$$\begin{aligned} E\left(f_R\left(x\right)\right) &= \int_{-c}^{1} k\left(t\right) f(x-th)dt \\ &= f(x)\int_{-c}^{1} k\left(t\right)dt - hf^{(1)}(x)\int_{-c}^{1} tk\left(t\right)dt + \frac{h^2}{2}f^{(2)}(x)\int_{-c}^{1} t^2k\left(t\right)dt + o\left(h^2\right) \\ &= f(x) - f(x)\int_{-1}^{-c} k\left(t\right)dt - hf^{(1)}(x)\int_{-c}^{1} tk\left(t\right)dt + \frac{h^2}{2}f^{(2)}(x)\int_{-c}^{1} t^2k\left(t\right)dt + o\left(h^2\right), \end{aligned}$$

therefore the value of bias of f_R is

$$-f(x)\int_{-1}^{-c}k(t)\,dt+hf^{(1)}(x)\int_{-c}^{1}tk(t)\,dt+\frac{h^2}{2}f^{(2)}(x)\int_{-c}^{1}t^2k(t)\,dt+o\left(h^2\right). \quad (3.1.4)$$

Similar computations give the variance expression,

$$\frac{f(x)}{nh}\int_{-c}^{1}k^2(t)\,dt+o\left(\frac{1}{nh}\right). \quad (3.1.5)$$

However, the usual bias is

$$\frac{h^2}{2}f^{(2)}(x)\int_{-1}^{1}t^2k(t)\,dt+o\left(h^2\right), \quad (3.1.6)$$

for (3.1.4) and (3.1.6), we see that f_R is not a consistent estimator of f and there exists an extra first order term of h. To correct this boundary problem, we construct a new approach, the basic technique of construction of the proposed estimator is kind of a generalized reflection method involving reflecting a transformation of the observed data. Then, a comparison of the boundary performance of our proposed estimator with the other kernel density estimators is carried out. It is well-known that a comparison between different methods is only meaningful with respect to their respective optimal performances. We have adopted this strategy in our comparison.

The rest of the paper is formulated as following. Section 3.2 introduces asymptotic properties of the proposed kernel estimator. Section 3.3 conducts Monte Carlo simulations and data analysis to compare the performance of our estimator, which is the main objective of this paper.

3.2 Transformation-reflection Kernel Density Estimation

Using transformation and reflection method in kernel density estimations improved bias at the boundary, but unless the first derivative of the density is 0, the estimator with reflection can still be much more severely biased at the boundary than in the interior. Marron and Ruppert [12] propose to transform the data to a density that has its first derivative equal to 0 at both boundaries. The transformation is selected from a parametric family, which is allowed to be quite general in our theoretical study. Zhang and al [30] combine those two methods to construct a new approach which correct the boundary problem at the left side of the support. We use this technique to correct the boundary problem at the right side. The proposed estimator defined as follow,

$$f_{n,TR}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{k\left(\frac{x-X_i}{h}\right) + k\left(\frac{x-2+\psi(X_i)}{h}\right)\right\}, \tag{3.2.1}$$

The transformation ψ is stated in the theorem 3.2.1, which exhibits the explicit forms of the bias, variance and mean squared error (MSE), under certain conditions on ψ.

Theorem 3.2.1 *Assume that $\psi^{(3)}$ exist and is continuous, where $\psi^{(i)}$ denote the i^{th} derivative of ψ. Further assume that $\psi^{-1}(1) = 1$ and $\psi^{(1)}(1) = 1$, where ψ^{-1} is the inverse function of ψ. Then for $x = 1 - ch, 0 \leq c < 1$, we have,*

$$Bias\left(f_{n,TR}\left(x\right)\right) = h\int_{-1}^{-c}(t+c)k\left(t\right)dt\left[2f^{(1)}\left(1\right) - f\left(1\right)\psi^{(2)}(1)\right] + \frac{h^2}{2}\mu_2 f^{(2)}(1)$$
$$- \frac{h^2}{2}\int_{-1}^{-c}(t+c)^2 k\left(t\right)dt \times \left[f\left(1\right)\psi^{(3)}(1) - 3\psi^{(2)}(1)\left[f^{(1)}\left(1\right) - f\left(1\right)\psi^{(2)}(1)\right]\right]$$
$$+ o\left(h^2\right), \tag{3.2.2}$$

and

$$Var\left(f_{n,TR}\left(x\right)\right) = \frac{f(1)}{nh}\left[\int_{-1}^{1} k^2\left(t\right)dt + 2\int_{-c}^{1} k\left(t\right)k\left(-\left(2c+t\right)\right)dt\right] + o\left(\frac{1}{nh}\right).$$

We shall choose the transformation ψ so that the first order term in the bias expansions (3.2.2) is zero. Assume that $f(1) > 0$, it is enough to let,

$$\psi^{(2)}\left(1\right) = 2f^{(1)}\left(1\right)/f\left(1\right). \tag{3.2.3}$$

So ψ should satisfy the following three conditions:

$C\left[1\right]$. *ψ is monotonically increasing.*

$C\left[2\right]$. $\psi^{(1)}\left(1\right) = 1$ *and* $\psi^{-1}\left(1\right) = 1$.

$C\left[3\right]$. $\psi^{(2)}\left(1\right) = 2f^{(1)}\left(1\right)/f\left(1\right)$.

The transformation function ψ, verify the conditions $C[1]$, $C[2]$ and $C[3]$, has the form:

$$\psi\left(x\right) = M - BM^2 + \left(1 - 2M + 3BM^2\right)x + \left(M - 3BM^2\right)x^2 + BM^2x^3, \tag{3.2.4}$$

where

$$M = f^{(1)}\left(1\right)/f\left(1\right), \tag{3.2.5}$$

and

$$B > 1/3. \tag{3.2.6}$$

For ψ be defined by (3.2.4) and for $x = 1 - ch, 0 \leq c < 1$, we have,

$$Bias\left(f_{n,TR}\left(x\right)\right) = \frac{h^2}{2}\left\{\mu_2 f^{(2)}(1) - 6\left[B+1\right]\frac{\left[f^{(1)}\left(1\right)\right]^2}{f\left(1\right)} \times \int_{-1}^{-c}(t+c)^2 k\left(t\right)dt\right\} + o\left(h^2\right). \tag{3.2.7}$$

Then the approximate form of mean squared error (MSE) is,

$$\begin{aligned} MSE\left(f_{n,TR}\left(x\right)\right) \sim \frac{h^4}{4}&\left(\mu_2 f^{(2)}(1) - 6\left[B+1\right]\frac{\left[f^{(1)}\left(1\right)\right]^2}{f\left(1\right)} \times \int_{-1}^{-c}(t+c)^2 k\left(t\right)dt\right)^2 \\ &+ \frac{f(1)}{nh}\left[\int_{-1}^{1} k^2\left(t\right)dt + 2\int_{-c}^{1} k\left(t\right)k\left(-\left(2c+t\right)\right)dt\right]. \end{aligned} \tag{3.2.8}$$

The mean integrated squared error $(MISE)$ of $f_{n,TR}\left(x\right)$ can be expressed as the sum of the integrated squared bias and the integrated variance for it,

$$MISE\left(f_{n,TR}\left(x\right)\right) = \int Bias^2\left(f_{n,TR}\left(x\right)\right)dx + \int Var\left(f_{n,TR}\left(x\right)\right)dx. \tag{3.2.9}$$

Estimation of ψ

In practice, the transformation ψ given by (3.2.4) is not available because it defined by unknown term M (3.2.5). We must replace M with a pilot estimator. Our proposed estimator (3.2.1) is not very sensitive to the accurate details of the pilot estimate of M, and therefore any appropriate estimate can be used. Note that M can be written as the derivative of $\log f(x)$ evaluated at $x = 1$, so M can be

estimated by,

$$M_n = \frac{\log f_R(1) - \log f_R(1-h)}{h}, \tag{3.2.10}$$

we now define,

$$\psi_n(x) = M_n - BM_n^2 + \left(1 - 2M_n + 3BM_n^2\right)x + \left(M_n - 3BM_n^2\right)x^2 + BM_n^2x^3, \tag{3.2.11}$$

as the estimator of $\psi(x)$.

The proposed new estimator

Our proposed new estimator of $f(x)$ is defined as, for $x = 1 - ch, 0 \leq c < 1$,

$$f_{n,TR,new}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{k\left(\frac{x - X_i}{h}\right) + k\left(\frac{x - 2 + \psi_n(X_i)}{h}\right)\right\},$$

where ψ_n is given by (3.2.11) with M replaced by M_n of (3.2.10).

Proof This proof starts by proving the bias of $f_{n,TR}$, we have

$$\begin{aligned} E(f_{n,TR}(x)) &= \frac{1}{h}E\left\{k\left(\frac{x - X_1}{h}\right) + k\left(\frac{x - 2 + \psi(X_1)}{h}\right)\right\} \\ &= \mathrm{I}_1 + I_2 \end{aligned}$$

By using change of variable and Taylor expansion for $x = 1 - ch, 0 \leq c < 1$, we can write,

$$\begin{aligned} I_1 &= \frac{1}{h}\int_0^1 k\left(\frac{x - y}{h}\right) f(y)\, dy \\ &= f(x)\int_{-c}^{1} k(t)\, dt - hf^{(1)}(x)\int_{-c}^{1} tk(t)\, dt + \frac{h^2}{2}f^{(2)}(x)\int_{-c}^{1} t^2k(t)\, dt \\ &\quad + o(h^2). \end{aligned}$$

Therefore,

$$
\begin{aligned}
I_2 = & \int_{-1}^{-c} k(t) \frac{f\left(\psi^{-1}(th-x+2)\right)}{\psi^{(1)}(\psi^{-1}(th-x+2))} dt \\
= & \int_{-1}^{-c} k(t) \left[\frac{f\left(\psi^{-1}(1)\right)}{\psi^{(1)}(\psi^{-1}(1))} \right. \\
& + (t+c)h \left[\frac{f^{(1)}\left(\psi^{-1}(1)\right)\psi^{(1)}(\psi^{-1}(1)) - f\left(\psi^{-1}(1)\right)\psi^{(2)}(\psi^{-1}(1))}{\left[\psi^{(1)}(\psi^{-1}(1))\right]^3} \right] \\
& + (t+c)^2 \frac{h^2}{2} \left[\frac{\psi^{(1)}(\psi^{-1}(1)) f^{(2)}\left(\psi^{-1}(1)\right) - f\left(\psi^{-1}(1)\right)\psi^{(3)}(\psi^{-1}(1))}{\left[\psi^{(1)}(\psi^{-1}(1))\right]^4} \right. \\
& - \frac{3\psi^{(2)}(\psi^{-1}(1))\left[f^{(1)}\left(\psi^{-1}(1)\right)\psi^{(1)}(\psi^{-1}(1))\right.}{\left[\psi^{(1)}(\psi^{-1}(1))\right]^5} \\
& \left. \left. - \frac{f\left(\psi^{-1}(1)\right)\psi^{(2)}(\psi^{-1}(1))\right]}{\left[\psi^{(1)}(\psi^{-1}(1))\right]^5} \right] dt \\
& + o(h^2).
\end{aligned}
$$

Using the condition $C\,[2]$ we have,

$$
\begin{aligned}
\frac{1}{h} E\left[k\left(\frac{x-2+\psi(X_1)}{h} \right) \right] = & \int_{-1}^{-c} f(1) k(t)\, dt + h\left[f^{(1)}(1) - f(1)\psi^{(2)}(1) \right] \int_{-1}^{-c} (t+c) k(t)\, dt \\
& + \frac{h^2}{2}\left[f^{(2)}(1) - f(1)\psi^{(3)}(1) - 3\psi^{(2)}(1)\left[f^{(1)}(1) - f(1)\psi^{(2)}(1) \right] \right] \\
& \times \int_{-1}^{-c} (t+c)^2 k(t)\, dt \\
& + o(h^2).
\end{aligned}
\tag{3.2.12}
$$

By the existence and continuity of $f^{(2)}$ near 1, we obtain, for $x = 1 - ch$,

$$\begin{aligned} f(1) &= f(x) + chf^{(1)}(x) + \frac{(ch)^2}{2} f^{(2)}(x) + o(h^2). \\ f^{(1)}(x) &= f^{(1)}(1) - chf^{(2)}(1) + o(h). \\ f^{(2)}(x) &= f^{(2)}(1) + o(1). \end{aligned}$$

So,

$$f(1) = f(x) + chf^{(1)}(1) - \frac{(ch)^2}{2} f^{(2)}(1) + o(h^2). \tag{3.2.13}$$

Now combining (I_1) and (I_2) and using the formula (3.2.13), we get,

$$\begin{aligned} Bias\left(f_{n,TR}(x)\right) = h \int_{-1}^{-c} (t+c)k(t)\,dt \left[2f^{(1)}(1) - f(1)\,\psi^{(2)}(1)\right] + \frac{h^2}{2}\mu_2 f^{(2)}(1) \\ - \frac{h^2}{2} \int_{-1}^{-c} (t+c)^2 k(t)\,dt \left[f(1)\,\psi^{(3)}(1) - 3\psi^{(2)}(1)\left[f^{(1)}(1) - f(1)\,\psi^{(2)}(1)\right]\right] \\ + o\left(h^2\right). \end{aligned}$$

The task now is to prove the variance of $f_{n,TR}$: observe that for $x = 1 - ch, 0 \leq c < 1$, we have,

$$\begin{aligned} Var\left(f_{n,TR}(x)\right) &= \frac{1}{nh^2} Var\left\{k\left(\frac{x - X_1}{h}\right) + k\left(\frac{x - 2 + \psi(X_1)}{h}\right)\right\} \\ &= J_1 + J_2, \end{aligned}$$

where,

$$
\begin{aligned}
J_1 &= \frac{1}{nh^2} E\left[k\left(\frac{x-X_1}{h}\right) + k\left(\frac{x-2+\psi(X_1)}{h}\right)\right]^2 \\
&= \frac{1}{nh^2}\left[\int_0^1 k^2\left(\frac{x-y}{h}\right) f(y)dy + \int_0^1 k^2\left(\frac{x-2+\psi(y)}{h}\right) f(y)dy\right] \\
&+ \frac{2}{nh^2}\int_0^1 k\left(\frac{x-y}{h}\right) k\left(\frac{x-2+\psi(y)}{h}\right) f(y)dy \\
&= J_{11} + J_{12}.
\end{aligned}
$$

Using a Taylor expansion, it can be shown that,

$$
\begin{aligned}
J_{11} &= \frac{1}{nh}\left[\int_{-c}^{1} k^2(t) f(x-th)dt + \int_{-1}^{-c} k^2(t)\frac{f(\psi^{-1}(th-x+2))}{\psi^{(1)}(\psi^{-1}(th-x+2))}dt\right] \\
&= \frac{1}{nh}\left[\int_{-c}^{1} k^2(t)\left[f(1)+o(1)\right]dt + \int_{-1}^{-c} k^2(t)\left(\frac{f(\psi^{-1}(1))}{\psi^{(1)}(\psi^{-1}(1))}+o(1)\right)dt\right] \\
&= \frac{f(1)}{nh}\mu_2 + o\left(\frac{1}{nh}\right),
\end{aligned}
$$

and,

$$
\begin{aligned}
J_{12} &= \frac{2}{nh}\int_{-c}^{1} k(t)\, k\left(\frac{x-2+\psi(x-th)}{h}\right) f(x-th)dt \\
&= \frac{2}{nh}\int_{-c}^{1} k(t)\, k\left(\frac{1-ch-2+1-(c+t)h)+o(h)}{h}\right) f(1-(c+t)h)dt \\
&= \frac{2f(1)}{nh}\int_{-c}^{1} k(t)\, k(-(2c+t))\, dt + o(\frac{1}{nh}).
\end{aligned}
$$

Similarly as in the proof of J_1, we get

$$J_2 = -\frac{1}{nh^2}\left[E^2\left(k\left(\frac{x-X_1}{h}\right)+k\left(\frac{x-2+\psi(X_1)}{h}\right)\right)\right]$$
$$= o\left(\frac{1}{nh}\right).$$

By adding up J_1 and J_2, we have the desired result for the variance.

3.3 Simulation study

To compare the performance of our proposed estimator against the other well known estimators, we divided this section into two parts, in the first, we introduce the results of simulated data and in the second part, we present some examples of real data. All computations were done by utilizing R software.

3.3.1 Simulated Data

In our simulation study reported in this part, we introduced the issue of potential quality of our proposed estimator per se form that of bandwidth selection. Throughout our study we use Epanechnikov kernel $k(t) = (3/4)(1-t^2)\,\mathbb{I}(-1 \leq t \leq 1)$, where $\mathbb{I}$ denote the indicatrice function.

Smoothing Parameter Selection

It is well known that the kernel estimation of the density depends crucially on the bandwidths. In our study, we used two methods of smoothing parameter selection which are the optimal bandwidth and the cross validation method.

Optimal Bandwidth

The popular bandwidth selector in kernel density estimation is due to Sheather and Jones [49]. This method adopts the asymptotic $MISE$ as criterion, defined by

$$AMISE \sim \frac{h^4}{4}\mu_2^2 \int \left[f^{(2)}(x)\right]^2 dx + \frac{1}{nh}\int k^2(t)dt, \tag{3.3.1}$$

the optimal bandwidth minimizing (3.3.1) is,

$$h_{opt} = \left\{ \int k^2(t)dt / n\mu_2^2 \int \left[f^{(2)}(x)\right]^2 dx \right\}^{1/5}.$$

Cross Validation Method

Rudemo [20] and Bowman [1] suggested known as unbiased cross-validation (UCV) in kernel density estimator, is surely the most popular and exceed studied one. The basic thought of this strategie, it purpose to estimate h the minimizer of $ISE(h)$. The minimisation measure is characterized by,

$$h_{ucv} = \arg\min_h UCV(h), \tag{3.3.2}$$

where

$$UCV(h) = \int f_{n,R}^2(x)\, dx - \frac{2}{n}\sum_{i=1}^{n} f_R(x_i). \tag{3.3.3}$$

Compared Estimators

We compare the performance of the kernel density estimator f_R, the transformation-reflection kernel density estimation $f_{n,TR}$, the boundary kernel estimator $f_{n,B}$ and the Jones and Foster estimator $f_{n,JF}$. The comparison is carried out with respect to the different densities.

The boundary kernel estimator is the general boundary corrected estimators define by Jones [16], which replace the standard kernel function by the modified version. The modified kernel function gave at the right boundary region based on the Epanechnikov kernel, by

$$k_B(t) = 12\frac{1-t}{(1+c)^4}\left(\frac{3c^2-2c+1}{2} - t(1-2c)\right)\mathbb{I}(-c \leq t \leq 1), \tag{3.3.4}$$

this kernel satisfies the following conditions,

$$\int_{-1}^{-c} k_B(t)\,dt = 0, \int_{-c}^{1} k_B(t)\,dt = 1, \int_{-c}^{1} tk_B(t)\,dt = 0, \int_{-c}^{1} t^2 k_B(t)\,dt < \infty, \tag{3.3.5}$$

the boundary kernel estimator is defined as,

$$f_{n,B}(x) = \frac{1}{nh}\sum_{i=1}^{n} k_B\left(\frac{x-X_i}{h}\right). \tag{3.3.6}$$

The Jones and Foster estimator that corrects for the possible negativity of the boundary kernel estimates has the following form,

$$f_{n,JF}(x) = f_{n,CN}(x)\exp\left(\frac{f_{n,B}(x) - f_{n,CN}(x)}{f_{n,CN}(x)}\right), \tag{3.3.7}$$

where,

$$f_{n,CN}(x) = \frac{1}{nh}\sum_{i=1}^{n} k_{CN}\left(\frac{x-X_i}{h}\right), \tag{3.3.8}$$

denote the cut-and-normalized density estimator introduced by Gasser and Muller [33] and by using the kernel function k_{CN} for the right boundary region truncated and normalized, ensuring integration to unity. For Epanechnikov kernel, it given by

$$k_{CN}(t) = \frac{1-t^2}{\int_{-c}^{1}\left(1-t^2\right)dt}\mathbb{I}(-c \leq t \leq 1) \tag{3.3.9}$$

Simulation Steps

We can compile the steps of simulation as follow,

Step 1. We simulate sample of size n with R repetition from the probability density f.

Step 2. We calculate h using the optimal bandwidth and the cross validation method.

Step 3. We estimate f by $f_R, f_{n,TR}, f_{n,B}$ and $f_{n,JF}$.

Step 4. We compute the $SBIS, VAR$ and MSE of each estimator.

Step 5. We graph the $MISE$ in the form of a boxplot.

For evaluating the performance of estimators at the boundaries, we tend to limit our attention to $n = 200$. We consider six distributions with bounded support $[0, 1]$. This set of distributions illustrated in Table 3.1, and for each distrubution we simulate $R = 1000$.

Results and discussions

For each density, we have calculated the squared bias ($SBIS$), variance (VAR) and mean squared error (MSE) of the estimators at the endpoint $x = 1$ using the

Table 3.1: Densities used in the simulation

Distribution		Density function
D_1	Beta $(3/2, 1)$	$(3/2)x^{1/2}$
D_2	Truncated Gamma $(2, 1)$	$x\exp(-x)/1 - 2\exp(-1)$
D_3	Truncated Normal $(0, 1)$	$\exp(-x^2/2)/\int_0^1 \exp(-t^2/2)dt$
D_4	$(1/2)$Beta $(3/2, 1)$ +$(1/2)$Beta $(1, 3/2)$	$0.75x^{1/2} + 0.75(1-x)^{1/2}$
D_5	Truncated Exponential (1)	$\exp(-x)/1 - \exp(-1)$
D_6	Truncated Beta $(5, 1)_{[1/2,1]}$	$160x^4/31$

two methods of smoothing parameter selection. The results are presented in Table 3.2 and Table 3.3.

Comparing between the estimators, we can see from the table 2, which includes simulated values of $SBIS$, VAR and MSE calculated by using the optimal bandwidth, that $f_{n,TR}$ had the smallest values of $SBIS$, VAR and MSE among the other estimators for all the cases considered, followed by $f_{n,B}$ and $f_{n,JF}$ estimators, while the f_R estimator is the worst among them, as to the Beta, Truncated Gamma and the last one Truncated Beta, but as to Truncated Normal, Mixture Beta and Truncated Exponontiel, the $f_{n,B}$ changed your position to the third place and $f_{n,JF}$ came in the second place.

From the table 3, which includes simulated values of $SBIS$, VAR and MSE

Table 3.2: The squared bias, variance and MSE values computed using the optimal bandwidth

Density		D_1	D_2	D_3	D_4	D_5	D_6
Estimator		$h = 0.2122$	$h = 0.1884$	$h = 0.2259$	$h = 0.2341$	$h = 0.2304$	$h = 0.2378$
f_R	$SBIS$	0.0954	0.0832	0.0221	0.0315	0.0148	0.0434
	VAR	0.0584	0.0466	0.0109	0.0132	0.0078	0.0248
	MSE	0.1538	0.1298	0.0331	0.0447	0.0226	0.0683
$f_{n,TR}$	$SBIS$	0.0205	0.0220	0.0096	0.0119	0.0078	0.0133
	VAR	0.0094	0.0077	0.0025	0.0030	0.0022	0.0048
	MSE	0.0299	0.0297	0.0121	0.0149	0.0100	0.0182
$f_{n,B}$	$SBIS$	0.0222	0.0262	0.0123	0.0149	0.0108	0.0163
	VAR	0.0156	0.0172	0.0073	0.0093	0.0065	0.0107
	MSE	0.0378	0.0433	0.0196	0.0242	0.0173	0.0270
$f_{n,JF}$	$SBIS$	0.0228	0.0267	0.0121	0.0143	0.0106	0.0165
	VAR	0.0173	0.0183	0.0069	0.0088	0.0062	0.0116
	MSE	0.0402	0.0451	0.0190	0.0231	0.0167	0.0282

Table 3.3: The squared bias, variance and MSE values computed using the cross validation method

Density		D_1	D_2	D_3	D_4	D_5	D_6
Estimator		$h = 0.1324$	$h = 0.0865$	$h = 0.0945$	$h = 0.2256$	$h = 0.1297$	$h = 0.1729$
f_R	$SBIS$	0.0946	0.0803	0.0292	0.0353	0.0209	0.0502
	VAR	0.0687	0.0583	0.0141	0.0162	0.0083	0.0289
	MSE	0.1633	0.1385	0.0433	0.0516	0.0292	0.0791
$f_{n,TR}$	$SBIS$	0.0463	0.0415	0.0223	0.0240	0.0146	0.0349
	VAR	0.0189	0.0157	0.0054	0.0055	0.0032	0.0092
	MSE	0.0652	0.0572	0.0278	0.0295	0.0178	0.0440
$f_{n,B}$	$SBIS$	0.0566	0.0511	0.0284	0.0298	0.0191	0.0444
	VAR	0.0331	0.0298	0.0144	0.0153	0.0102	0.0217
	MSE	0.0897	0.0809	0.0429	0.0452	0.0293	0.0661
$f_{n,JF}$	$SBIS$	0.0575	0.0518	0.0286	0.0299	0.0194	0.0449
	VAR	0.0320	0.0280	0.0146	0.0152	0.0114	0.0224
	MSE	0.0895	0.0798	0.0432	0.0451	0.0308	0.0674

calculated by using the unbasied cross validation method, it can be observe that $f_{n,TR}$ has good performance among the others. Moreover, the ranking from best to worst concerning the $SBIS$ is $f_{n,TR}$, $f_{n,B}$, $f_{n,JF}$, f_R for all densities, but concerning the MSE we can observe that $f_{n,B}$ change it's order to the third place and $f_{n,JF}$ came to the second place for the Beta, Truncated Gamma and Mixture Beta densities, but for the other densities Truncated Normal, Truncated Exponentiel and Truncated Beta $f_{n,B}$ and $f_{n,JF}$ came in the second, third place respectively.

Comparing two smoothing parameter for a given estimators, we also find that, in general, the optimal bandwidth tends to perform better.

3.3.2 Real Data

In this section, we apply our proposed estimators over two data sets. The densities of our data sets are assumed to have a compact support $S = [a, b]$. In our study, we mapped the original observation $X_i \in S$ onto the unit interval by the transformation $Y_i = (X_i - a)/(b - a)$.

The natural stands of the seedlings and saplings of Japanese black pines.

The first data set consists the data were collected on the position, height (cm) and age $(years)$ of the natural stands of the seedlings and saplings of 204 Japanese black pines in a $100m^2$ region. The data set can be found in the paper of Ogata Y. and Tanemura M. (1985). Table 3.4 shows the descriptive statistics of the data. We used bandwidth $h = 0.14$, which we choose subjectively. We have graphed the performance of our proposed estimator and the histogram of unknown real

Table 3.4: Descriptive statistics of the natural stands of the seedlings and saplings of Japanese black pines data.

Min	1st Qu	Median	Mean	3rd Qu	Max	kurtosis	skewness
3.10	10.20	15.40	31.92	46.52	150.2	6.70	1.91

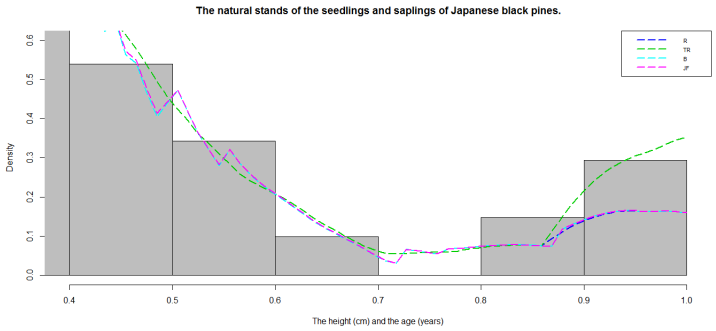

Figure 3.1: Density estimates of the natural stands of the seedlings and saplings of 204 Japanese black pines.

density function in figure 3.1.

From the figure 3.1 alone, one can see that the $f_{n,TR}$ is a good estimator of the true density removes a large part of the boundary effect and when we move to the interior, we remark that all the estimators close to the kernel density estimator. We can conclude that, $f_{n,TR}$ yield the best estimator of natural stands of the seedlings and saplings of Japanese black pines data and hence can be adequat for estimation these data.

The measure of the motor cortex neuron interspike of unstimulated monkey.

The second data set is the measres intervals of motor cortex neuron interspike (in ms) for an unstimulated monkey. The objects of the analysis were to estimate

Table 3.5: Descriptive statistics of the measrues intervals of motor cortex neuron interspike for an unstimulated monkey data.

Min	1st Qu	Median	Mean	3rd Qu	Max	kurtosis	skewness
2.00	20.00	29.50	36.49	49.25	104	1.08	3.87

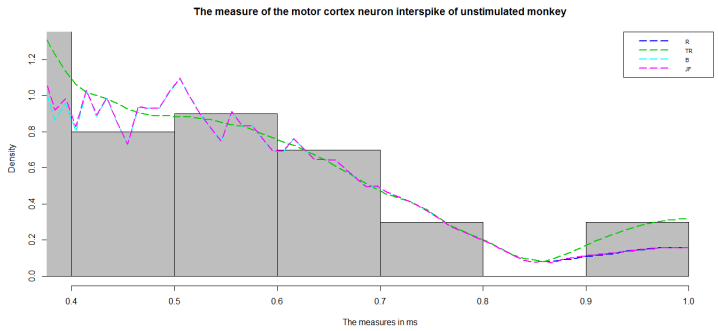

Figure 3.2: Density estimates of the measres intervals of motor cortex neuron interspike for an unstimulated monkey.

the firing rate prior to Stimulation and to characterize the time dependence. The data set can be found in the paper of Zeger, S.L. and Bahjat Qaqish (1988). The descriptive statistics of the data is given in Table 5. The bandwidth is chosen subjectively to be $h = 0.16$. The proposed estimators are plotted in figure 3.2, superimposed on the histogram of the data.

From figure 3.2 we can see that the $f_{n,TR}$ is closer to the empirical histogram of the density. That indicates, $f_{n,TR}$ is well covers the density of the measure of the motor cortex neuron interspike of unstimulated monkey data.

"

CHAPTER 4

Conclusions

The concept of probability density functions is a vital concept in statistics. Density function estimates represent a corner stone in a wide range of statistical analyses. Among the density function estimation techniques, the kernel method provides a simple and efficient way for estimating density functions, but this method is not consistent near the finite end points of their supports.

In other words, these effects seriously affect the performance of these estimators. In this thesis, we have studied the boundary effect in the kernel density estimation. We have mentioned some methods for correcting this effect and we propose a very intuitive and feasible kernel density estimator which reduces the bias.

The proposed estimator possesses a number of desirable properties, including the non-negativity of the estimator. Each estimator has certain advantages and works well at certain times. The proposed method seems to have inherited the best of both transformation and reflection methods and that improved boundary

effects near the points at left boundary region.

Bibliography

[1] BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **71**, 2, 353–360.

[2] COWLING, A. AND HALL, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation, *Journal of the Royal Statistical Society: Series B*, textbf58, 3, 551–563.

[3] ZAMBOM, A. Z. AND DIAS, R. (2013). A review of kernel density estimation with applications to econometrics, *International Econometric Review*, **5**, 1, 20–42.

[4] SILVERMAN, B. W. (1986). Density estimation for statistics and data analysis, *Chapman and Hall, London* , **175**, 0–412.

[5] KIM, C.; KIM, W. AND PARK B. U., (2003). Skewing and generalized jackknifing in kernel density estimation, *Taylor and Francis*.

[6] CLINE, D. B. H. AND HART, J. D., (1991). Kernel estimation of densities with discontinuities or discontinuous derivatives, *Statistics: A Journal of Theoretical and Applied Statistics*, **22** 1, 69–84.

[7] CHOI, E. AND HALL, P., (1998). On bias reduction in local linear smoothing, *Biometrika*, **85** 2, 333–345.

[8] PARZEN, E., (1962). On estimation of a probability density function and mode, *The annals of mathematical statistics*, **33** 3, 1065–1076.

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

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

[11] TERRELL, G. R.; SCOTT, D. W. AND OTHERS, (1980). On improving convergence rates for nonnegative kernel density estimators, *Annals of Statistics*, **8** 5, 1160–1163 .

[12] MARRON, J. S. AND RUPPERT, D., (1994). Transformations to reduce boundary bias in kernel density estimation, *Journal of the Royal Statistical Society: Series B (Methodological)*, **56** 4, 653–671.

[13] CHENG, M.-Y., (1997). Boundary aware estimators of integrated density derivative products, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59** 1, 191–203 .

[14] CHENG, M.-Y.; FAN, J.; MARRON, J. S.; AND OTHERS, (1997). On automatic boundary corrections, *The Annals of Statistics* **25** 4, 1691–1708.

[15] HIRUKAWA, M., (2010). Nonparametric multiplicative bias correction for kernel-type density estimation on the unit interval, *Statistics and Data Analysis* **54** 2, 473–495.

[16] JONES, M. C. , (1993). Simple boundary correction for kernel density estimation, *Statistics and computing*, **3** 3, 135–146.

[17] JONES,M. C. AND HENDERSON, D. A., (2007). Miscellanea kernel-type density estimation on the unit interval, *Biometrika*, **94** 4, 977–984.

[18] JONES, M. C.; LINTON, O. AND NIELSEN, J. P., (1995). A simple bias reduction method for density estimation, *Biometrika*, **82** 2, 327–338.

[19] ROSENBLATT, M., (1956). Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics*, pages 832–837.

[20] RUDEMO, M., (1982). Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics*, pages 65–78.

[21] CHENG, M.; CHOI,E. AND HALL, J. P., (2000). Skewing methods for two-parameter locally parametric density estimation, *International Statistical Institute (ISI) and the Bernoulli Society for Mathematical Statistics and Probability*, **6** 1, 169–182 .

[22] JONES, M. C. AND FOSTER, P. J., (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica*, 1005–1013.

[23] MALEC, P. AND SCHIENLE, M., (2014). Nonparametric kernel density estimation near the boundary. *Computational Statistics and Data Analysis*, 72, 57–76.

[24] HALL, P. AND PARK, B. U., (2002). New methods for bias correction at endpoints and boundaries, *The Annals of Statistics*, **30** 5, 1460–1479 .

[25] KARUNAMUNI, R. J. AND ALBERTS, T., (2005). On boundary correction in kernel density estimation, *Statistical Methodology*, **2** 3, 191–212.

[26] KARUNAMUNI, R. J. AND ALBERTS, T. , (2006). A locally adaptive transformation method of boundary correction in kernel density estimation, *Journal of Statistical Planning and Inference*, **136** 9, 2936–2960, .

[27] ZHANG, S., (2010). A note on the performance of the gamma kernel estimators at the boundary, *Statistics and probability letters*, **80** 7-8, 548–557.

[28] ZHANG, S. AND KARUNAMUNI, R. J., (1998). On kernel density estimation near endpoints, *Journal of statistical Planning and inference*, **70** 2, 301–316.

[29] ZHANG, S. AND KARUNAMUNI, R. J., (2000). On nonparametric density estimation at the boundary, *Journal of nonparametric statistics*, **12** 2, 197–221.

[30] ZHANG, S.; KARUNAMUNI, R. J. AND JONES, M. C., (1999). An improved estimator of the density function at the boundary, *Journal of the American Statistical Association*, **94** 448, 1231–1240.

[31] ZHANG, S. AND KARUNAMUNI, R. J. , (2010). Boundary performance of the beta kernel estimators, *Journal of Nonparametric Statistics*, **22** 1, 81–104.

[32] CHEN, S. X. , (1999). Beta kernel estimators for density functions, *Computational Statistics and Data Analysis*. **31** 2, 131–145.

[33] GASSER, T. AND MULLER, H.-G. , (1979). Kernel estimation of regression functions, *Springer* pages 23–68.

[34] GASSER,T.; MULLER, H.-G. AND MAMMITZSCH V., (1985). Kernels for nonparametric curve estimation, *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–252.

[35] WEN, K., WU, X. (2015). An improved transformation-based kernel estimator of densities on the unit interval. *Journal of the American Statistical Association*, 110(510), 773-783.

[36] CHEN, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics Data Analysis*, 31(2), 131-145.

[37] CHOI, E., HALL, P. (1998). On bias reduction in local linear smoothing. *Biometrika*, 85(2), 333-345.

[38] KOLAÂEK, J., KARUNAMUNI, R. J. (2009). On boundary correction in kernel estimation of ROC curves. *Austrian Journal of Statistics*, 38(1), 17-32.

[39] KARUNAMUNI, R. J., ZHANG, S. (2008). Some improvements on a boundary corrected kernel density estimator. *Statistics Probability Letters*, 78(5), 499-507.

[40] GEENENS, G. (2014). Probit transformation for kernel density estimation on the unit interval. *Journal of the American Statistical Association*, 109(505), 346-358.

[41] MORIYAMA, T. (2017). A new method of joint nonparametric estimation of probability density and its support. *arXiv preprint arXiv*:1704.08015.

[42] MYNBAEV, K., MARTINS-FILHO, C. (2010). Bias reduction in kernel density estimation via Lipschitz condition. *Journal of Nonparametric Statistics*, 22(2), 219-235.

[43] CHIU, S. T. (2000). Boundary adjusted density estimation and bandwidth selection. *Statistica Sinica*, 1345-1367.

[44] BAGAI, I., RAO, B. P. (1995). Kernel type density estimates for positive valued random variables. *SankhyÃÂ: The Indian Journal of Statistics, Series A*, 56-67.

[45] HIRUKAWA, M., SAKUDO, M. (2014). Nonnegative bias reduction methods for density estimation using asymmetric kernels. *Computational Statistics Data Analysis*, 75, 112-123.

[46] MYNBAEV, K., MARTINS-FILHO, C. (2019). Unified estimation of densities on bounded and unbounded domains. *Annals of the Institute of Statistical Mathematics*, 71(4), 853-887.

[47] *Cao, R., Cuevas, A., Manteiga, W. G.* (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics Data Analysis*, 17(2), 153-176.

[48] *Wand, M.P. and Jones, M.C.,* (1995): Kernel Smoothing. *London: Chapman and Hall*.

[49] *Sheather, S. J. and Jones, M. C.* (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 683-690.

APPENDIX A

R Software

R is a computer language not entirely unlike the S language developed at ATT Bell Laboratories by Rick Becker, John Chambers and Allan Wilks. The two languages are implemented quite differently, but bear enough superficial resemblance that users should be able to switch between the two with relative ease. Currently the software is undergoing active development. Discussion of the development process is carried out on the "r-devel" mailing list.

We have implemented R in what we hope is a very portable fashion and inway which requires relatively little in the way of machine resources. Implementations exist for many for many members of the Unix family of operating systems, including AIX, FreeBSD, GNU/Linux, HPUX, Irix, macOS, Solaris, and Tru64. Inaddition there is a version for Microsoft Windows (9x, ME, NT4, 2000, XP).

We have implemented R in what we hope is a very portable fashion and in way which requires relatively little in the way of machine resources. Implementations exist for many for many members of the Unix family of operating systems, including AIX, FreeBSD, GNU/Linux, HPUX, Irix, macOS, Solaris, and Tru64. In addition there is a version for Microsoft Windows (9x, ME, NT4, 2000, XP).

APPENDIX

B

Abbreviations and Notations

X	Random variable
$X_1, ..., X_n$	Sample of n observations
I_A	Indicator function of set A
$\mathbb{R}$	Set of real numbers
F	Distribution function
f	Propability density function
f_H	histogram estimate
f_n	naive estimator
f_R	kernel estimator
f_V	variable kernel estimate
f_{GW}	general weight function estimate
f_{Ref}	reflection method
f_{Tr}	transformation method
f_{LATr}	Locally Adaptive Transformation Estimator
f_{DT}	Double transformation Estimator
f_{CH}	Cowling and Hall Estimator
f_{Bou}	boundary kernel estimator
f_{LL}	local linear Estimator
f_{CN}	cut-and-normalized Estimator

k	Kernel function
K	Distribution of kernel function
h	Bandwidth or smoothing parameter
μ	Mean
$f^{(i)}$	The i th derivatives of f
ψ	Transformation function
$\psi^{(i)}$	The i th derivatives of ψ
iid	Independent and identically distributed
$[0, \infty)$	Positive interval
$E(X)$	Esperance of X
$Var(X)$	Variance of X
MSE	Mean Squared Error
$AMSE$	Asymptotic Mean Squared Error
$AMISE$	Asymptotic Mean Integrated Squared Error
UCV	Unbiased cross validation
CCV	Complete cross-validation.
MCV	Modified cross-validation
$MLCV$	Maximum likelihood cross-validation
$SBIS$	Squared Bias
PDF	Probability Density Function

Abstract

In this thesis, we study some boundary correction methods for kernel estimators of the density function and their statistical properties. The kernel estimators are not consistent near the finite end points of their supports. In other words, these effects seriously affect the performance of these estimators. To remove these boundary effects, various methods have been developed in the literature, the most used are the reflection method, the transformation and the local linear method. We combine transformation and reflection methods in order to introduce a new boundary correction estimator in the case of kernel estimation of the density function whose support is [0,1].

Résumé

Dans cette thèse, nous étudions certaines méthodes de correction aux bords pour les estimateurs à noyau de la fonction de densité et leurs propriétés statistiques. Les estimateurs à noyau ne sont pas cohérents près des extrémités finies de leurs supports. En d'autres termes, ces effets affectent sérieusement les performances de ces estimateurs. Pour supprimer ces effets de bords, différentes méthodes ont été développées dans la littérature, les plus utilisées sont la méthode de réflexion, la transformation et les méthodes linéaires locales. Nous combinons les méthodes de transformation et de réflexion afin d'introduire un nouvel estimateur de correction aux bords dans le cas de l'estimation à noyau de la fonction de densité dont le support est [0,1].