People 's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

THIRD CYCLE LMD FORMATION

A Thesis submitted in partial execution of the requirements of the degree of

DOCTOR IN MATHEMATICS

Suggested by

Mohamed Khidher University Biskra

Presented by

ALMI Nassima

Titled

On Kernel Inverse Distribution Function Estimation Near the Boundary.

Supersivor: **Pr. SAYAH Abdallah**

Examination Committee :

Pr. YAHIA Djabrane	University of Biskra	President
Pr. SAYAH Abdallah	University of Biskra	Supervisor
Pr. BENATIA Fateh	University of Biskra	Examiner
Pr. DJEFFAL El Amir	University of Batna	Examiner

MOHAMED KHIDER UNIVERSITY, BISKRA
FACULTY of EXACT SCIENCES, SIENCE of NATURE and LIFE
DEPARTMENT of MATHEMATICS

A thesis submitted for the fulfillment of the requirements of :

(The Doctorate Degree in Mathematics)

Option : Statistics

Presented by:
ALMI Nassima

Titré

On Kernel Inverse Distribution Function Estimation Near the Boundary.

Sur l'estimation à noyau de la fonction de distribution inverse près de la frontière .

October 4th, 2022

Supervisor: Pr. SAYAH Abdallah

Dedication

I dedicate this work

To the soul of my father, ***Sadek.***

To the heart of my mother, ***Fatima***

To my brothers: ***Nadir****,* ***Mohamed****,* ***Tarek*** *and* ***Walid***

To my sisters: ***Mofida, Kenza****,* ***Sana*** *and all my family*

To my dear friends ***Amira****,* ***Lamia*** *and* ***Hocine.***

Acknowledgments

At the end of this adventure, I express my sincere thanks to all those who, in one way or another, have contributed to the realization of this thesis.

First of all, I enthusiastically thank my supervisor, **Pr. Sayah Abdallah**, who give me the chance to learn about the kernel estimation area and for the wise advice that developed my thinking and for his patience, his availability during this job.

I would like also to thank warmly each members of the jury **Pr. YAHIA Djebrane**, and **Pr. BENATIA FATEH**, **Pr. DJEFFAL El Amir.** Who do me the great honored to accept evaluate this dissertation and for the valuable questions expected of them which incite me to widen my research from various perspectives.

From heart to heart, I thank my mother and brothers, sisters for their existence, really their existence made me here today. I do not forget to give back to my friends for their loyalty in friendship, which they have proof.

Thank you...

A. Nassima

Abstract

This work is about a nonparametric approach of both cumulative distribution and quantile function to improve boundary effects in the kernel estimation method. It is very often the case that the natural support of a distribution to be estimated is not the whole real line but an interval bounded on one or both sides. Hence, the kernel distribution estimator may not provide appropriate estimates of the distribution function at such points. To remove this effect, a variety of methods have been developed in the literature, the most widely used is the reflection, the convex combination, ... In this thesis, we introduce a new method of boundary correction when estimating both cumulative distribution and quantile function. Our technique based on a self elimination between the Bias and the estimator it self. we turned out that, with an adequate choice of the parameters of the two proposed estimators, the rate of convergence of two estimators will be faster than the existing kernel proposed.

Keywords: kernel distribution function estimation , Kernel inverse distribution function estimation, Optimal bandwidth, Boundary effect.

ملخص

يدور هذا العمل حول مقاربة غير معلمية لكل من التوزيع التراكمي والدالة الكمية لتحسين التأثير ات الحدودية في طريقة. تقدير النواة.

غالبًا ما يكون مجال تعريف دالة التوزيع المراد تقديره ليس مجال الحقيقي بأكمله ولكن مجال جزء منه محدد على أحد الجانبين أو كلاهما. ومن ثم ، قد لا يكون اداء مقدر توزيع النواة التقليدي مناسب لوظيفة التوزيع في هذه النقاط. لإزالة هذا التأثير ، تم تطوير مجموعة متنوعة من الأساليب و الطرق للعديد من الباحثين ، وأكثرها استخدامًا هو الانعكاس ، والتركيبة المحدبة ، ...في هذه الأطروحة ، نقدم طريقة جديدة لتصحيح الحدود عند تقدير كل من التوزيع التراكمي والكمي تعتمد تقنيتنا على القضاء الذاتي بين الانحياز والمقدر نفسه. لقد تبين لنا أنه مع الاختيار المناسب لمعلمات المقدرين المقترحين h, k ، فإن معدل تقاربااثنين من المقدرين سيكون أسرع من المقدرات الحالية الموجودة.

الكلمات المفتاحية : تقدير دالة توزيع النواة، تقدير دالة التوزيع العكسي، النطاق الترددي الامثل، تاثيرات الحدود .

Résumé

Ce travail traite d'une approche non paramétrique de la distribution cumulative et de la fonction quantile pour améliorer les effets de frontière dans la méthode d'estimation à noyau. Il arrive très souvent que le support naturel d'une distribution à estimer ne soit pas toute la droite réelle mais un intervalle borné d'un ou des deux côtés. Par conséquent, l'estimateur de distribution par noyau peut ne pas fournir d'estimations appropriées de la fonction de distribution à ces points. Pour supprimer cet effet, diverses méthodes ont été développées dans la littérature, les plus utilisées sont la réflexion, la combinaison convexe, ... Dans cette thèse, nous introduisons une nouvelle méthode de correction des limites lors de l'estimation à la fois de la distribution cumulative et de la fonction quantile. Notre technique est basée sur une auto-élimination entre le Bias et l'estimateur lui-même. nous nous sommes avérés qu'avec un choix adéquat de paramètres des deux estimateurs proposés, le taux de convergence de deux estimateurs seront plus rapides que le noyau existant proposé.

Mots-clés : Estimation de la function de distribution du noyau, Estimation de l'inverse de la function de distribution du noyau, Bande passante optimale, Effects de frontière.

Contents

List of Figures

List of Tables

List of Tables

1 Introduction

Nonparametric methods are gradually coming popular in satistical analysis of many fields problems, such as in Economics, Biology, and actuarial Science, this is because of the lack of information about the variable being analyzed and requires minimum assumptions like the continuity of the sampled population and it's quite powerful even if the sample sizes are small when compared with the parametric estimation method. Knowledge of the density function or distribution function, or their estimates, allows one to characterize the random variable more completely. Especially for the distribution function, we can derive some other characteristics of random variables from that, such as quantiles, survival function, hazard rate, etc. The kernel estimation method belongs to a general category of techniques for nonparametric curve estimation including nonparametric density estimators, nonparametric distribution estimators, and nonparametric quantile estimators. These estimators are now popular and in wide use with great success in statistical applications. Some results on kernel density estimation are due to Rosenblatt [39] and Parzen[35]. Good references in this area are Silverman [46], and Wand and Jones [56]. In the case of the estimation of the distribution and the inverse distribution function (quantiles function) have been proposed and studied extensively, references can be found in the work of Yamato,[61], Azzalini [4], and in the books of Galambos[18] and David [11].

Kernel estimation method depends on two parameters. The first one is called a bandwidth denoted h which controls the smoothness of the estimator indeed, a low value of h parameter implies a low degree of smoothing of the estimator. In contrast, a wide value of h leads to an over-smooth estimator. Several methods for choosing the bandwidth are discussed later, which allowed us to conclude that an adequate h is necessary for the good performance of the estimator. The second parameter is a kernel denoted k which plays a role of weight function. As far as the kernel function is concerned, a key parameter is its order which is related both to the number of its vanishing moments and to the number of existing derivatives for the underlying curve to be estimated. In Generally, the choice of kernel is relatively unimportant with respect to the choice of the smoothing parameter h, which determines the extent of the kernel on each side of the observation. In certain applicable restrictions for the convenience of theoretical developments, Epanechnikov [15]

propose a kernel that is optimal in the sense of integrated mean squared error (MISE), i.e. the kernel which minimizes the MISE, a more precise definition of the MISE will be given later. Rao 1983 it came to the conclusion that the choice of a kernel other than the optimal kernel only led to a slight loss of precision. Lall and al [28] defined that the choice of the kernel has a certain importance, but that its influence on the overall estimate is relatively low. It is however important to bring some precision to these conclusions.

The use of the classical form of kernel estimator causes the increase of the bias estimator, particularly in the so-called boundary region, near to end of support. In practical problems such a situation occurs often as many random variables considered in the problems of economic, technical or natural sciences are characterized by bounded support on one or both sides. In most situations, left boundary equals zero when the data under consideration are measurements of positive quantities. In different analyses, random variables with non-negative values are considered (duration of unemployment, the stock price, time of performing the specific technical operations, the amount of inventory in the warehouse, time of growing plants, and amount of atmospheric fall).

The problem of estimating a quantile function from observed data $X_1, \ldots, X_n$ of a continuous random variable X is typically solved by estimating the distribution function assuming that all observations are mutually independent and come from identical distributions. In the context of kernel distribution function estimation, the asymptotic properties of the classical estimator Nadaraya [32] in Interior region $[a + h, b - h[$ do not hold anymore for the points near the left $[a, a + h[$ or right $[a + h, b[$ end of the support$[a, a + h[$ when the density function has compact support $[a, b]$ where $a < b$. Hence, the kernel distribution estimator may not provide appropriate estimates of the distribution function at such points. The boundary problem in kernel distribution estimation is less severe than in kernel density estimation. This is due to the extra information $F(a) = 0$ and $F(b) = 1$. Kolacek and Karunamuni [26] considered the boundary problem in distribution function estimation in estimating ROC curves using the transformation method discussed in Zhang and al [63]. Tenreiro [51] proposed a boundary kernel method for correcting the boundary problem. However, Tenreiro [51] did not reveal the fact that there is no boundary problem in distribution function estimation if the density has a value of zero at the endpoints of the support. In his method, the boundary kernel k_c is constructed by truncating a density kernel at $[-c; c]$, and then normalizing it so that it integrates to 1 on $[-c; c]$. Realizing the fact that such boundary kernel corrects the boundary problem by shrinking the bandwidth to zero when data is near the boundary the resulting distribution estimates may have

high variability at such points. Zhang [64] develop a boundary distribution kernel method for correcting the boundary problem of the classical, which is continuous, non-decreasing, and does not have the aforementioned high variability problem of the estimator proposed in Tenreiro [51]. Tour and al [53] develop a new kernel estimator of the distribution function for heavy-tailed distributions based on the modified Champernowne transformation.

Some previous research has already studied nonparametric estimation of the inverse distribution. On the one hand, Azzalini [4] suggested estimating the CDF and then obtaining the quantile from its inverse function. On the other hand, Harrell and Davis [21] proposed an alternative quantile estimator, based on a weighted sum of sample observations. Later, Sheather and Marron [49] analyzed the existing kernel methods for quantile estimation and proposed a smoothing parameter. Most of the existing estimators suffer from either a bias or an inefficiency for high probability levels (p near 1). Inspired by Wand et al. [55]; Buch-Larsen et al [8] showed that for heavy-tailed distributions, the tail performance of the classical kernel density estimator could be significantly improved by using a tail flattening transformation. They used modified Champernowne distribution to estimate loss distributions in insurance which is categorically heavy-tailed distributions. Sayah et.al[43] produce a kernel quantile estimator for heavy-tailed distributions, which is based on the estimation of quantiles of the transformed variable so it can easily to be estimated using a classical approach of the kernel estimation and then taking the inverse transform, this idea was first used in the context of density estimation by Devroye and Gyorfi [13] for heavy-tailed observations.

The rest of this thesis is organized in two parts as follows.

The first part It is an introduction to the non-parametric estimation method, where some common approaches are presented in the distribution and inverse distribution function context and its asymptotic properties. Chapter (2).
Most estimators of both functions mentioned above have a problem with bias in the case when the data is near the boundary, chapter (3) deals with boundary effect where some recent methods of boundary correction have been discussed.
The second part contains our main results in order to reduce the bias in kernel distribution estimation and the inverse distribution context at the boundary region. In chapter (4), two kernel distribution function estimators are introduced and investigated in order to improve the boundary effects, we will restrict our attention to the right boundary. The theoretical properties of our estimators are established and their performance is evaluated by a simulation study and two real data applications. In chapter (5), we suggested an alternative estimator to the inverse distribution

kernel estimator and provided its asymptotic behavior when quantile near the boundary value. A simulation study and two real data applications were included to demonstrate the efficiency and reliability of our theoretical results.
Throughout this thesis, the following assumptions hold for f

- f is differentiable with bounded derivative $f^{(1)}$
- $f^{(1)}$ is continuous in the neighborhood of $Q(p)$ and $f^{(1)}(Q(p)) \neq 0$

Part I

Nonparametric Estimation

2 Nonparametric Estimation

The aim of this introductory part is to present the context in which the present dissertation takes place. The problem of estimating the inverse distribution function from observed data $X_1, \ldots, X_n$ assuming that all observations are mutually independent and come from identical distributions (iid) of a continuous random variable X is typically solved by estimating the distribution function according to the form

$$F^{-1}(p) = inf\{x \in R : F(x) \geq p\}, \qquad p \in]0, 1[.$$

Estimating the cumulative distribution function (CDF) is a fundamental goal in many fields in which analysts are interested in estimating the risk of occurrence of a particular event, for example As an effect of global warming, the insurance industry is increasingly exposed to extreme events such as hurricanes, hail storms and tornados, etc. Such events cause catastrophic losses. It is necessary to estimate the probability of such events and the probability of the payout exceeding certain amounts (such as $1,000,000$) in order for the insurance companies to determine the appropriate premiums. Denote by X the amount of the payout from an accident, the quantity of interest is $P(X > x)$, where x is a prespecified amount of payout. We assume that X is a random variable from a population with density f and CDF F where $F(x) = P(X \leq x)$ and the corresponding inverse function (quantiles) is $Q(p) = F^{-1}(p)$. Note that $Q(.)$ is the left-continuous inverse of F. several methods have been proposed to estimate CDF among them :

2.1 Empirical estimation method

2.1.1 Empirical distribution function estimator EDF

Let $X_1, X_2, ..., X_n$ be a data sample of a continuous random variable X. The most commonly used nonparametric estimation of a function F is an empirical distribution function (EDF) F_n, that puts mass $\frac{1}{n}$ at each data point x_i defined at some point x as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_i),$$

where $\mathbb{1}$ is the indicator function defined by

$$\mathbb{1}_{]-\infty,x]}(X_i) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x. \end{cases}$$

The EDF is most conveniently defined in terms of the order statistics of a sample. Suppose that the n sample observations are distinct and arranged in increasing order so that $X_{(1)}$ is the smallest and the $X_{(n)}$ is the largest. A formal definition of the E.D.F. $F_n(x)$ is

$$F_n(x) = \begin{cases} 0 & \text{if} \quad x < X_{(1)} \\ \dfrac{i}{n} & \text{if} \quad X_{(i)} \leq x < X_{(i+1)} \\ 1 & \text{if} \quad x \geq X_{(n)}. \end{cases}$$

Statistical properties of the EDF

Using properties of the binomial distribution, we get the following results.

▶ **Corollary 2.1.** The mean and the variance of $F_n(.)$ are

$$E(F_n(x)) = F(x) \text{ and } V(F_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

◀

The corollary shows that $F_n(.)$, the proportion of sample values less than or equal to the specified value x, is an unbiased estimator of $F(x)$ and shows that the variance of $F_n(x)$ tends to zero as n tends to infinity. Thus, using Chebyshev's inequality, we can show that $F_n(x)$ is a consistent estimator of $F(x)$.

▶ **Corollary 2.2.** For any fixed real value x, $F_n(x)$ is a consistent estimator of $F(x)$, or, in other words, $F_n(x)$ converges to $F(x)$ in probability. ◀

The convergence in probability is for each value of x individually, whereas sometimes we are interested in all values of x, collectively. A probability statement can be made simultaneously for all x, as a result of the following important theorems.

▶ **Theorem 2.3 (Glivenko-Cantelli Theorem).** $F_n(x)$ converge uniformly (Convergence almost-surely) to $F(x)$, that is

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s} 0.$$

◀

This theorem has been called the fundamental theorem of (nonparametric) statistics.

▶ **Theorem 2.4 (Dvoretsky-Kiefer-Wolfowitz).** For any $\varepsilon > 0$,

$$P\left(\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

◀

▶ **Theorem 2.5.** As $n \longrightarrow \infty$, the limiting probability distribution of the standardized $F_n(x)$ is standard normal, or

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \xrightarrow{L} N(0,1).$$

◀

Despite the good statistical properties of F_n, the known fact that smoothing can lose, figure (2.1), shows that For a discrete real random variable the distribution function (Poisson's law)is constant on any interval of empty intersection with the support of the law. It is therefore constant in pieces, F_n is a step function even in case F is continuous (Gaussian's law) and even when n is large, F_n loses smoothing, one could prefer a rather smooth estimate.

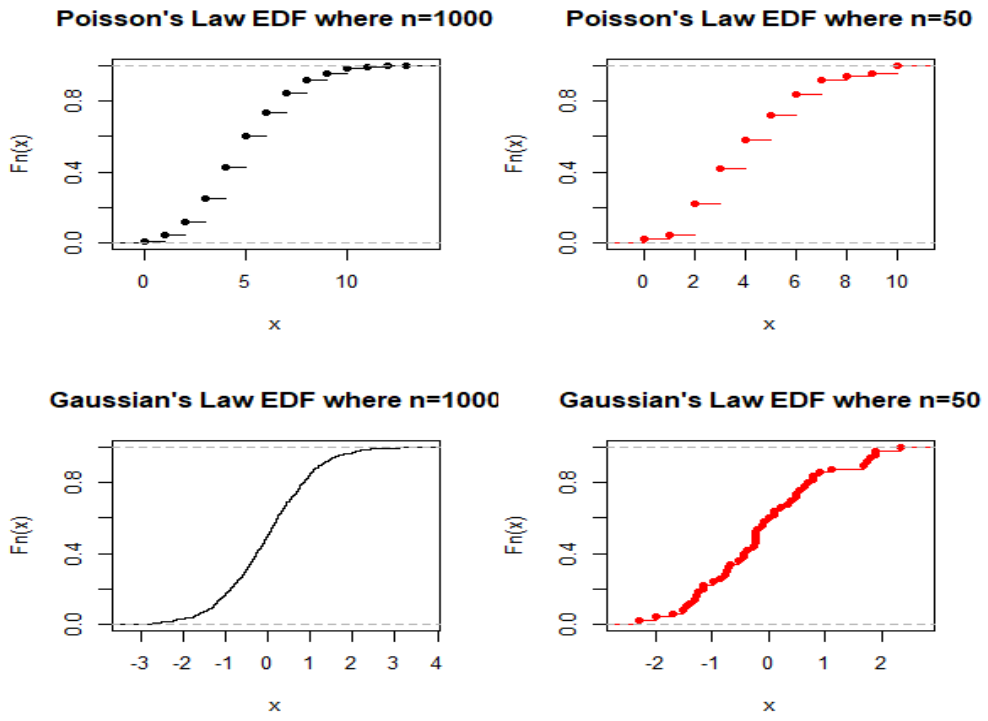

Figure 2.1: Smoothing of the empirical distribution function

2.1.2 Inverse of the empirical distribution function estimator

The corresponding estimator of $Q(p)$ is the p^{th} sample quantile which is given by

$$Q_n(p) = inf\{x \in R : F_n(x) \geq p\} = X_{[np]+1} \qquad p \in]0, 1[,$$

where [np] denotes an integral part of np.
When F is continuous, it is more natural to use a smooth random function as an estimator of F since there is a substantial lack of efficiency, caused by the variability of individual order statistics. Indeed, the choice of F_n does not always lead to the best estimator of F (see, Read [39]), which has shown that F_n is inadmissible with respect to the integrated square loss).
Different approaches to estimating sample quantiles through weighted order statistics have been proposed. A popular class of these estimators is called kernel quantile estimators.

2.2 Kernel estimation method

Nadaraya [32] proposed a smooth nonparametric alternative to the EDF estimator, namely, kernel distribution estimator (KDF) we denoted by $\widehat{F}_n$. This estimator is obtained by integrating the Rosenblatt-Parzen kernel density estimator, we denoted by $\widehat{f}_n$, that we briefly present in the following subsection.

2.2.1 Kernel density function estimator

It might seem natural to estimate the density f as the derivative of $F_n(x)$, but this estimator would be a set of mass points, not a density, and as such is not a useful estimate of $f(x)$. Instead, consider a discrete derivative. For some h small

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h},$$

we can write this as

$$\begin{aligned} f_n(x) &= \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{[x-h,x+h]}(X_i), \\ &= \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{[-1,1]}\left(\frac{X_i - x}{h}\right), \end{aligned}$$

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

where

$$w(t) = \begin{cases} \frac{1}{2} & |t| \leq 1 \\ 0 & |t| > 1 \end{cases}$$

where h is selected in such a way that $h := h_n$ ($h \longrightarrow 0$ and $nh \longrightarrow \infty$ as $n \longrightarrow \infty$) is the smoothing parameter, called the bandwidth, which controls the smoothness of the estimator

$f_n(x)$ is a special case of the Rosenblatt-Parzen estimator that is called the naive estimator. The naive estimator is not wholly satisfactory from the point of view of using density estimates for presentation. It follows from the definition that is not a continuous function but has jumped at the points $X_i - h$ and $X_i + h$ and has zero derivatives everywhere else. In figure (2.2), we plotted the performance of the naive estimator for the beta density function by giving two different values of h to illustrate the fact that the naive estimator is less smoothing. It is easy to

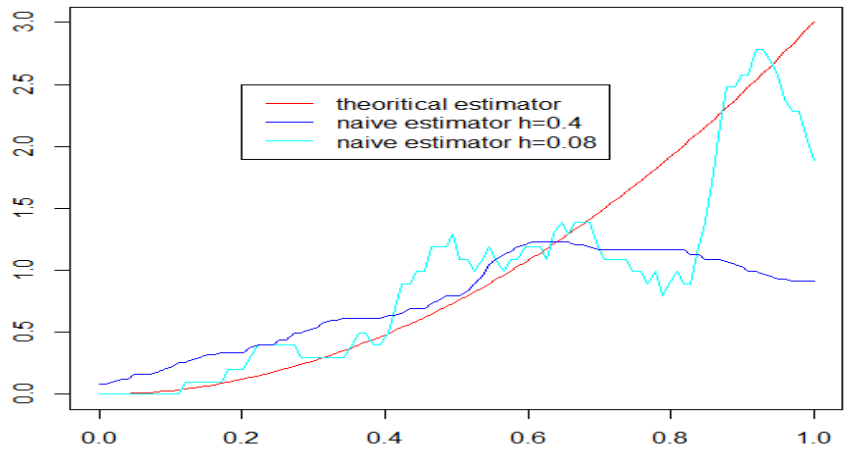

Figure 2.2 : Performance of the naive estimator.

generalize the naive estimator to overcome some of the difficulties discussed above. Replace the weight function w by a kernel function k. The only real restriction

on the kernel k is that its integration over the whole domain of the definition of x must be equal to one. One sometimes encounters other theoretical restrictions which are applied to k, such as

1. $k(-t) = k(t)$ hence k is a symmetric function.

2. $\mu_1 = 0$ and $\mu_2 < \infty$, where $\mu_j = \int_{-\infty}^{+\infty} t^j k(t)dt$.

However, these restrictions are mainly introduced in order to simplify theoretical developments. The nonparametric estimation of the density function of a sample can be seen as the cumulation of the functions k of each observation over the whole domain:

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

As an example, to compare the smoothness of the kernel density estimator with the naive estimator. We did an estimation based on $n = 200$ observations of the Beta distribution, we can see that the naive estimator is less smoothing than the kernel estimator for $k(t) = \frac{3}{4}(1 - t^2)\mathbb{1}_{[-1,1]}(t)$ and $h = 0.07$.

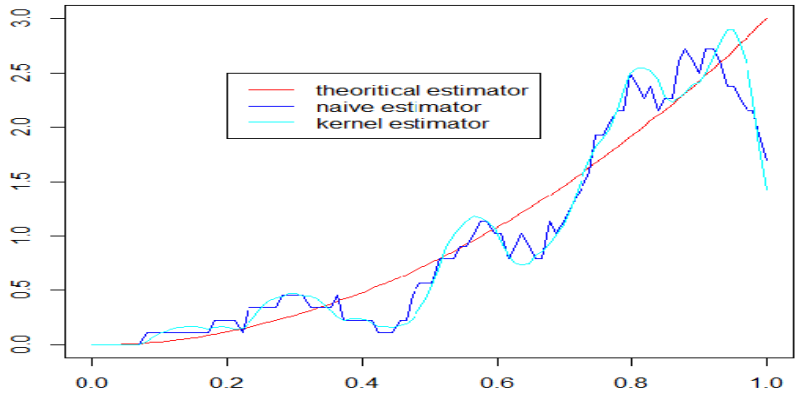

Figure 2.3: Comparison of the smoothness of the density estimators.

Optimal choice of kernel

The problem of the optimal choice of k consists in finding an optimal kernel under the constraint of positivity ($k \geq 0$). We recall the asymptotic properties of $\widehat{f_n}$

for $x \in \mathbb{R}$ we have

$$Bias(\widehat{f_n}(x)) = \frac{h^2}{2} f^2(x) \int_{-\infty}^{+\infty} x^2 k(x) dt + o(h^2),$$

and

$$Var(\widehat{f_n}(x)) = \frac{1}{nh} f(x) \int_{-\infty}^{+\infty} k(x) dx + o\left(\frac{1}{nh}\right),$$

then the mean square error is

$$Mse(\widehat{f_n}(x)) = \left(Bias(\widehat{f_n}(x)\right)^2 + Var(\widehat{f_n}(x)),$$

and the asymptotic mean integrated square error *Amise* is

$$Amise(\widehat{f_n}(x)) = \frac{1}{nh} \int_{-\infty}^{+\infty} k^2(t) dt + \frac{h^4}{4} (\mu_2(k))^2 \int_{-\infty}^{+\infty} \left(f^{(2)}(t)\right)^2 dt.$$

We note that the dependence of the *Amise* with respect to the kernel k is expressed by the intervention of its variance μ_2. An optimal kernel k^* is therefore a kernel that minimizes the functional μ_2

$$\mu_2(k^*) = min_{k \in \psi(k)} \mu_2(k) \ ,$$

where $\psi(k)$ denotes the set of positive kernels of order 1 satisfying the conditions

$$\int_{-\infty}^{+\infty} k(t) dt = 1, \int_{-\infty}^{+\infty} tk(t) dt = 0 \quad and \int_{-\infty}^{+\infty} t^2 k(t) dt < +\infty.$$

The solution of the problem is given by the following proposition.
Proposition(Tsybakov [54]) Let k be a kernel function, where $\mu_2(k) < \infty$, then k^* is

$$k^*(t) = \frac{3}{4}(1 - t^2) \mathbb{1}_{[-1,1]}(t).$$

We can consider the efficiency of each of the symmetrical kernels presented in table(2.1), compared with the Epanechnikov kernel. Efficiency is defined (see Silverman [46]) by :

$$eff(k) = \frac{C(k^*)}{C(k)},$$

where $C(k) = (\mu_2(k))^{\frac{2}{5}} \left(\int_{-\infty}^{+\infty} k^2(t)dt \right)^{\frac{4}{5}}$.

The problem of finding optimal kernels as minimizers of certain functionals was introduced into the theory of kernel density estimators by Epanechnikov [15]. Further results can be found in Gasser and Muller [19] and in Eddy [14] derive the optimal kernels for kernel estimators of the mode produce good results can be used, the following table presents the most frequently used kernels functions.

Kernel	Support	k(t)	Efficiency
Epanechnikov	$[-1,1]$	$\frac{3}{4}(1-x^2)$	1
Cosinus	$[-1,1]$	$\frac{\pi}{4}cos(\frac{\pi}{2}x)$	0,999
Biweight	$[-1,1]$	$\frac{15}{16}(1-x^2)^2$	0,994
Triweight	$[-1,1]$	$\frac{35}{32}(1-x^2)^3$	0,987
Triangulaire	$[-1,1]$	$1-\|x\|$	0,986
Gaussien	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$	0,946
Uniforme	$[-1,1]$	$\frac{1}{2}$	0,930
Double Epanechnikov	$[-1,1]$	$3\|x\|(1-\|x\|)$	0,816
Double Exponential	$\mathbb{R}$	$\frac{1}{2} \exp\{-\frac{1}{2}\|x\|\}$	0,759

Table 2.1: Usual kernel functions

In table 2.1, we can see that the efficiency values obtained are very close to 1 and that there is very little difference between different kernels based on asymptotic mean integrated square error, for this fact the choice of the kernel is less important in kernel estimation method.

▶ **Example 2.6.** An example is drawn in figure (2.3), where we show the performance of the kernel estimator by using two different kernel functions Epanechnikov and Gaussian kernels for a fixed value of $h = 1.4$ for Normal density

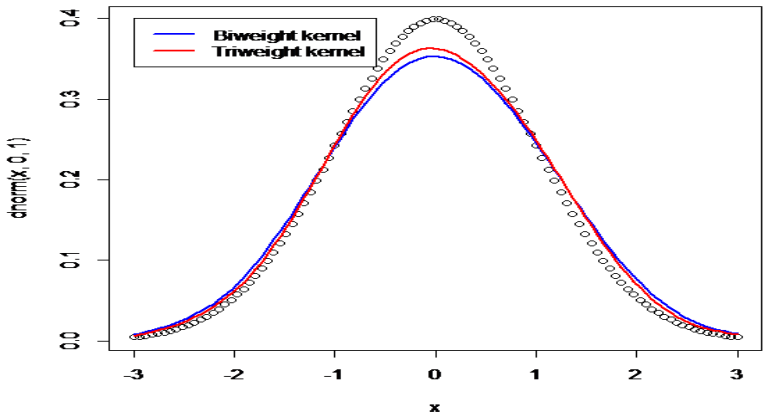

Figure 2.3: Influence of the kernel function to the performance of the kernel estimation.

2.2.2 Kernel distribution function estimator

As we can see in figure (2.2) the kernel density estimator is more smoothing than the naive estimator. the condition that k is a density function guarantees the existence of the primitive of the kernel k

$$K : \mathbb{R} \longrightarrow [0, 1]$$

i.e

$$K(t) = \int_{-\infty}^{t} k(y)dy,$$

then it is easy to construct a kernel estimator for the distribution function $\widehat{F_n}$ as:

$$\widehat{F_n}(x) = \int_{-\infty}^{x} \widehat{f_n}(y)dy,$$

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

using the substitution $y = \dfrac{t-X_i}{h}$ leads to

$$\widehat{F_n}(x) = \frac{1}{n}\sum_{j=1}^{n}\int_{-\infty}^{\frac{x-X_i}{h}} k(y)dy,$$

then the classical kernel distribution estimator of F at the point $x \in \mathbb{R}$ is defined as

$$\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right),$$

we assume that the kernel function k is a continuous density such that is bounded, and symmetric about zero $k(-t) = k(t)$. Thus k satisfies a kernel condition and the smoothing parameter h which tends to 0 as $n \to \infty$.

The estimate $\widehat{F_n}$ has been investigated by several authors, Nadaraya [32] has proved under mild conditions that $\widehat{F_n}$ has asymptotically unbiased and has the same variance as F_n with f is continuous Nadaraya [32], Winter [58], and Yamato [61] are obtains its uniform convergence to F with probability one, and without conditions on f Singh and al [47], Winter [59] also shows that checks the Chung-Smirnov property, that

$$\limsup_{n\to\infty}\left\{\left(\frac{2n}{\log\log n}\right)^{1/2}\sup_{x\in\mathbb{R}}\left|\widehat{F_n}(x) - F(x)\right|\right\} \leq 1 \quad ,$$

with probability 1. Watson and Leadbetter [57] proved the asymptotic normality of $\hat{F}_n$. Reiss [40] proves that the asymptotic relative inefficiency of F_n compared to $\widehat{F_n}$ tends rapidly to infinity as the sample size increases with an appropriate choice of kernel, e.g.

$$k(x) = \frac{9}{8}\left(1 - \frac{5}{3}x^2\right)\mathbb{1}_{[-1,1]}(t).$$

Falk [16], who has shown that the asymptotic performance of $\widehat{F_n}(x)$ is better than that of F_n in the sense of relative deficiency for appropriately chosen kernels and sufficiently smooth cdf's F.
Azzalini [4] derived also an asymptotic expression for the mean squared error Mse of $\widehat{F_n}(x)$ and determined the asymptotically optimal smoothing parameter, to have an Mse lower for F_n , and he obtained the asymptotic expressions for the mean integrated squared error Mise of $\widehat{F_n}(x)$. Some conditions verified in particular when the support of k is bounded and

$$\varphi(k) = 2\int_{-\infty}^{+\infty} xk(x)K(x)dx > 0 \quad .$$

Falk [16] provides a complete solution to this problem by establishing the representation of the relative inefficiency of F_n versus $\widehat{F_n}$ under the above conditions especially when the support of k is bounded. The number $\varphi(k)$ is introduced by Falk [16] as a measure of the asymptotic performance of the kernel k. But he shows that any square-integrable kernel does minimize φ. Then he uses the number

$$\rho(k) = \int_{-\infty}^{+\infty} k^2(y)dy \quad ,$$

, defined by Epanechnikov [15] as a measure of the performance of the kernel in density estimation. In the sense of ρ, the Epanechnikov kernel is the best but the Gaussian or Uniform kernels have very similar performance. Using the criterion ρ the Epanechnikov kernel is then by far the best of the three.
In the sense of mean integrated squared error Mise; the best kernel is the Uniform kernel although the performance of other kernels (Epanechnikov, Normal, Triangular) are, in practice, only slightly less good (Jones [23]. It is interesting to note that this is not the best kernel in the estimation of density.
The asymptotic expression of Mise. is also studied by SwanPoel [50]). For a continuous function f, he proves that the best kernel is the Uniform kernel.
Whereas for discontinuous f in a finite number of points, the Exponential kernel

$$k(x) = \frac{c}{2}exp(-c|x|) \qquad x \in \mathbb{R}.$$

for an arbitrary constant $c > 0$, $\widehat{F}_n(x)$ is again more efficiency than F_n for $h_n = o(n^{-1/2})$. However, $\widehat{F}_n(x)$ does not always provide a better estimate than F_n.

Indeed, in the case of a uniformly Lipschitz function F, Fernholz [17] obtains that

$$\sqrt{n}\| \widehat{F}_n(x) - F_n(x) \|_\infty \to 0 \quad a.s$$

and

$$\sqrt{n}\| \widehat{F}_n(x) - F(x) \|_\infty$$

and

$$\sqrt{n}\| F_n(x) - F(x) \|_\infty$$

have the same asymptotic distribution. In addition, Shirahata and Chu [48] show that under certain hypotheses on F the integrated square error

$$ISE = \int_{-\infty}^{+\infty} \left(\widehat{F}_n(x) - F(x)\right) dF(x)$$

for $\widehat{F}_n$ is almost certainly higher than that of F_n.

▶ **Example 2.7.** In figure (2.4), where we show the performance of the kernel estimator by using two different kernel functions Epanechnikov and Gaussian kernels for a fixed value of h.
In figure (2.5) we examine the performance of KDF for three different values of h for the Epanechnikov kernel. The data sample consists of 200 random numbers of a Beta distribution with parameters (1,3). ◀

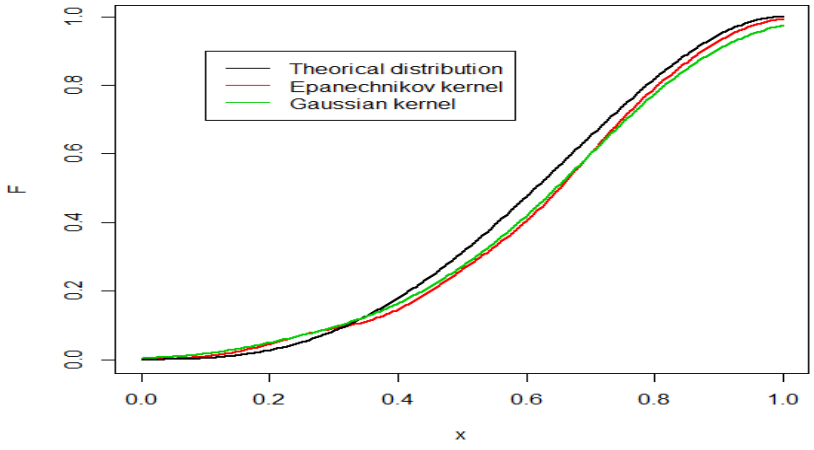

Figure 2.4: Influence of the kernel function to the performance of the KDF estimator.

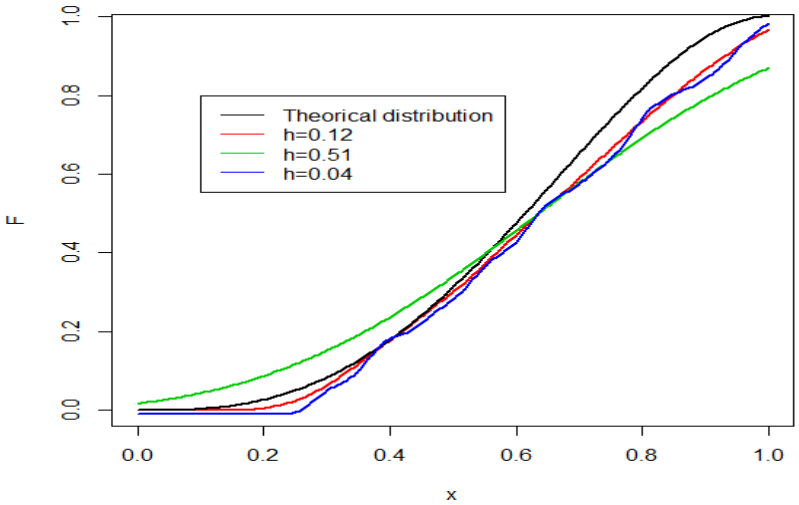

Figure 2.5: Influence of the bandwidth to the performance of the KDF estimator.

As we can see, the choice of the kernel function does not have a strong influence on the performance of the KDF, which confirms the results obtained in the case of the density kernel estimator (Table 2.1). While selecting an adequate bandwidth is essential for the good performance of the KDF estimator.

The problem of the choose an adequate bandwidth to ensure good performance of the KDFestimator consists in finding a minimization of a global measure of the error incurred when estimating $F(x)$ with $\widehat{F}_n(x)$ along $x \in \mathbb{R}$. A typical measure of performance for a KDFestimator is the mean integrated square error $Mise\ (h)$ for this fact we need to present some properties of the KDE estimator before we discuss the bandwidth selection method.

Statistical properties of KDF

Assume that k is symmetric and has a compact support $[-1, 1]$. Several properties of $\widehat{F}_n(x)$ are well known, we start with the evaluation of $E(\widehat{F}(x))$ at the point $x \in \mathbb{R}$:

$$\begin{aligned} E\left(\widehat{F}_n(x)\right) &= \int_{-\infty}^{+\infty} K\left(\frac{x-y}{h}\right) f(y)dy, \\ &= \int_{-\infty}^{x-h} 1f(y)dy + \int_{x-h}^{x+h} K\left(\frac{x-y}{h}\right) f(y)dy + \int_{x+h}^{+\infty} 0f(y)dy, \\ &= F(x-h) + h\int_{-1}^{1} K(t)f(x-ht)dt, \end{aligned}$$

we use the Taylor expansion we have

$$E\Big(\widehat{F}_n(x)\Big) = F(x) - hf(x) + \frac{1}{2}h^2 f^{(1)}(x) + h\int_{-1}^{1} K(t)\Big(f(x) - htf^{(1)}(x) + o(h)\Big)dt\ ,$$

$$= F(x) - hf(x) + \frac{1}{2}h^2 f^{(1)}(x) + hf(x) - h^2 f^{(1)}(x)\int_{-1}^{1} tK(t)dt\ ,$$

$$= F(x) + \frac{1}{2}h^2 f^{(1)}(x) - h^2 f^{(1)}(x)\left(\frac{1-\mu_2}{2}\right) + o\big(h^2\big).$$

Here, we notice that they yield an interesting formula for Bias

$$Bias(\widehat{F}_n(x)) = \frac{1}{2}F^{(2)}(x)\mu_2 h^2 + o(h^2). \tag{2.1}$$

For the variance, according to the definition we have

$$Var\Big(\widehat{F}_n(x)\Big) = \frac{1}{n}\left(\int_{-\infty}^{+\infty} K^2\Big(\frac{x-y}{h}\Big)f(y)dy - \left(\int_{-\infty}^{+\infty} K\Big(\frac{x-y}{h}\Big)f(y)dy\right)^2\right),$$

we are only dealing with the first term since the second is given in (2.1) Thus

$$\int_{-\infty}^{+\infty} K^2\left(\frac{x-y}{h}\right)\mathrm{d}y = \int_{-\infty}^{x-h} 1f(y)dy + \int_{x-h}^{x+h} K^2\Big(\frac{x-y}{h}\Big)f(y)dy$$

$$= F(x-h) + h\int_{-1}^{1} K^2(t)f(x-ht)dt,$$

$$= F(x) + hf(x)\left(\int_{-1}^{1} K^2(t)dt - 1\right) + o(h)$$

Since the expression of the Bias (2.1) can gives

$$E\left(K\left(\frac{x-X_i}{h}\right)\right) = F(x) + o(h)$$

Then the expression of $Var\left(\widehat{F}_n(x)\right)$ is

$$Var\left(\widehat{F}_n(x)\right) = \frac{1}{n}F(x)(1-F(x)) - \frac{h}{n}f(x)\left(1 - \int_{-1}^{1} K^2(t)dt\right) + o\left(\frac{h}{n}\right),$$

therefore

$$Var\left(\widehat{F}_n(x)\right) = \frac{1}{n}F(x)(1-F(x)) - \frac{h}{n}f(x)\varphi(k) + o\left(\frac{h}{n}\right). \tag{2.2}$$

The previous result shows that the asymptotic variance of $\widehat{F}_n(.)$ is of order $o\left(\frac{h}{n}\right)$ and it's smaller than the variance of the EDF. It is evident that for larger values of h, the quantity $hf(x)\varphi(k)$ increases, resulting in a smaller variance expression but a larger bias. This observation has important implications for choosing the bandwidth, i.e the choice of bandwidth h implies a variance-bias trade-off

- Larg h : $\widehat{F}_n$ is over-smoothing. Low Variance, high Bias,
- Small h : $\widehat{F}_n$ is under-smoother. High Variance, low Bias,

so we looking for h that

$$h = argmin\left(Mise(\widehat{F}_n(x))\right).$$

To obtain the Mean Squared Error (Mse) we combine (2.1) and (2.2), where

$$Mse\left(\widehat{F}_n(x)\right) = \left(Bias(\widehat{F}_n(x)\right)^2 + \left(Var(\widehat{F}_n(x)\right),$$

then we have

$$Mse\left(\widehat{F}_n(x)\right) = \frac{F(x)(1-F(x))}{n} - \frac{h}{n}f(x)(\varphi(k)) + \frac{h^4}{4}\left(F^{(2)}(x)\right)^2\mu_2^2 + o\left(\frac{h}{n} + h^4\right), \tag{2.3}$$

implicit that the expression of the asymptotic mean squared error $Amse\left(\widehat{F}_n(x)\right)$ is

$$Amse\left(\widehat{F}_n(x)\right) = \frac{F(x)(1-F(x))}{n} - \frac{h}{n}f(x)(\varphi(k)) + \frac{h^4}{4}\left(F^{(2)}(x)\right)^2\mu_2^2.$$

The bandwidth which minimizes the $Amise$ can be calculated by differentiating

the expression of the $Amise\left(\widehat{F}_n(x)\right)$, setting the equation to 0 and solving it for h. where

$$Amise\left(\widehat{F}_n(x)\right) = \int_{-\infty}^{+\infty} Amse\left(\widehat{F}_n(x)\right)dx, \tag{2.4}$$

The result is referred to

$$h_{opt}^{Amise} = Cn^{-\frac{1}{3}} = \left(\frac{\int_{-\infty}^{+\infty} V_F^2(x)dx}{2n\int_{-\infty}^{+\infty} B_F^2(x)dx}\right)^{\frac{1}{3}}, \tag{2.5}$$

where $B_F^2(x) = \frac{1}{2}\left(f^{(1)}(x)\right)^2\mu_2^2$ and $V_F^2(x) = 2f(x)\varphi(k)$.

Bandwidth selection in kernel distribution function estimation

In practice, to evaluate an optimal global bandwidth (2.5) we need to develop a method to replace the true distribution by her estimator. Several methods already exist to obtain different bandwidth selectors depending on the details of the procedure developed to minimize (2.4) without needing any additional estimate of distribution derivatives. Despite the great number of bandwidth selection techniques in other settings, for example in density or regression estimation Jones [23], Sheather and Marron [49] and Rio [41] . However in the distribution estimation context, only two popular methods have been investigated are plug-in and cross-validation methods.

Plug-in method

Because the constant C in equation (2.5) depends on the kernel function and the theoretical distribution function of the data unknown in practice, a plug-in estimation considers the bandwidth

$$h_{pl} = \hat{C}n^{-\frac{1}{3}}, \tag{2.6}$$

where $\hat{C}$ is estimated through the data sample. The way of obtaining $\hat{C}$ differs from one author to another.

Altman and Leger [3] consist in estimating nonparametric the unknown terms C, using Altman and Leger's notation, equation (2.5) can be written as:

$$h_{opt}^{Amise} = \left(\frac{\frac{1}{4}V_2}{B_3}\right)^{\frac{1}{3}} n^{\frac{-1}{3}}. \tag{2.7}$$

where $V_2 = \varphi(k)\int_{-\infty}^{+\infty}(f(x))^2dx$ and $B_3 = 0.25(\mu_2)^2\int_{-\infty}^{+\infty}(f^{(1)}(x))^2f(x)dx$.

So the plug-in bandwidth is

$$h_{AL} = \left(\frac{\frac{1}{4}\hat{V}_2}{\hat{B}_3} \right)^{\frac{1}{3}} n^{\frac{-1}{3}}, \tag{2.8}$$

where

$$\hat{V}_2 = \varphi(k) \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} \frac{1}{\alpha} k\left(\frac{x_i - x_j}{\alpha} \right),$$

and

$$\hat{B}_3 = 0.25(\mu_2(k))^2 \frac{1}{n^3 \alpha^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} k\left(\frac{x_i - x_j}{\alpha} \right) k\left(\frac{x_i - x_k}{\alpha} \right).$$

Polansky and Baker [36] plug-in Based also on Equation (2.6), using their notation can be written as

$$h_{BP} = \left(\frac{\varphi(k)}{-n\mu_2(k)^2 \hat{\psi}_2(g_2)} \right)^{\frac{1}{3}} \tag{2.9}$$

Polanski and Baker [36] developed an iterative method for calculating the plug-in bandwidth, which we detail below. Let $b > 0$ be an integer.
First step. Calculate $\hat{\Psi}_{2b+2}$ using the formula

$$\hat{\Psi}_{\mathrm{r}} = \frac{(-1)^{r/2} r!}{(2\hat{\sigma}(x_i))^{r+1} (r/2)! \pi^{1/2}},$$

where $\hat{\sigma}(x_i) = min\left(\hat{s}, \frac{Q_3 - Q1}{1.349} \right)$
with $\hat{\sigma}$ the sample standard deviation, and Q_1, Q_3 denoting the first and third quartile, respectively.
Second step. Begin from $j = b$ to $j = 1$, calculating $\hat{\Psi}_{2j}\left(\hat{g}_{2j}\right)$ where

$$\hat{g}_{2j} = \left(\frac{2L^{(2j)}(0)}{-n\mu_2(L)\hat{\Psi}_{2j+2}} \right)^{1/(2j+3)}$$

with

$$\hat{\Psi}_{2j+2} = \begin{cases} \hat{\Psi}_{2b+2} & \text{if} \quad j = b \\ \hat{\Psi}_{2j+2}\left(\hat{g}_{2j+2}\right) & \text{if} \quad j < b \end{cases}.$$

In practice, it is sufficient to consider $b = 2$ for most applications.

Cross-validation methods The cross-validation procedure is based on directly estimating the function MISE in Equation (2.4), and then selecting the bandwidth to minimize this function. Sarda [44] proposed to use

$$CV_S(h) = \frac{1}{n}\sum_{i=1}^{n}\left(F_n(x_i) - \widehat{F}_{-i}(x)\right)^2 dx,$$

where $\widehat{F}_{-i}(x)$ denotes the kernel estimator constructed from the data with observation x_i omitted.
In spite of the asymptotic optimality theorem proven in Sarda [44], this method does not provide good results in practice. Instead, the modified cross-validation proposal of Bowman et al [7] is also asymptotically optimal and works well in simulation studies and real cases. It consists in minimizing the function

$$CV_B(h) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\left(\mathbb{1}_{[0,+\infty[}(x - x_i) - \widehat{F}_{-i}(x)\right)^2 dx,$$

Lopez and al [31] $CV_B(h)$ is an unbiased estimator of MISE(h) plus an unknown constant that does not depend on h. They also demonstrate that minimization of $CV_B(h)$ leads to a bandwidth that is asymptotically equivalent to the bandwidth minimizing MISE(h).
Bowman and al [7] use a simulation study to compare this method with the plug-in one of Altman and Leger [3] . Better results are obtained, in general, with cross-validation. A drawback is the worse performance in terms of computational time, obviously, this is not really a drawback, for a real data situation, because the minimization process is carried out only once.

▶ **Example 2.8.** In figure (2.5), we show the performance of the kernel estimator by using three bandwidth selection methods, using a normal kernel and a standard normal distribution, in each case.

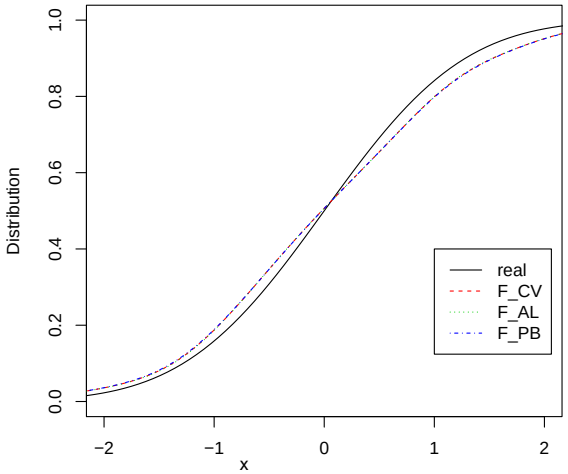

Figure 2.5: Distribution function estimation using the studied bandwidth selection methods. ◀

2.2.3 Kernel inverse Distribution Function Estimator

The main drawback to sample quantiles is that they experience a substantial lack of efficiency caused by the variability of individual order statistics in practical applications, we observe a finite number of samples. Indeed, Q_n is the inverse of the empirical distribution function.
Estimating the inverse distribution function has been treated extensively by several authors mention among them Parzen [35], Azzalini [4], Falk [16], Nadaraya [32], Yamato [61], Yang [62], Harrell and Davis [2.8], and Sheater and Marron [49]. As we see later, this type of kernel quantile estimator has a slower rate of convergence when p is a boundary point than when p is a fixed interior point. Indeed, An alternative estimator to p^{th} sample quantile based on the Nadaraya [32] estimator $\widehat{F}$ is the kernel quantile estimators given by

$$\widetilde{Q}_n(p) = inf\{x \in \mathbb{R} \setminus \widehat{F}_n(x) \geq p\}, \quad p \in [0,1] \tag{2.10}$$

Nadaraya [32] showed under some assumptions for k, f and h $\widetilde{Q}(p)$ has an asymptotic standard normal distribution. The almost sure consistency was obtained by Yamato [61]. Ralescu and Sun [39] obtained the necessary and sufficient conditions for the asymptotic normality of $\widetilde{Q}$.
Shankar [45] proved that for all $p \in]0,1[$ we have

$$Bias\Big(\widetilde{Q}_n(p)\Big) = \frac{h^2\Big(f^{(1)}(Q(p))\Big)^2}{2f^2(Q(p))} + o\big(h^2\big),$$

and

$$Var\Big(\widetilde{Q}_n(p)\Big) = \frac{p(1-p)}{nf^2(Q(p))} - \frac{h}{nf(Q(p))}\varphi(k) + o\left(\frac{h}{n}\right),$$

then the mean squared error of $\widetilde{Q}_n(p)$ is

$$\mathrm{Mse}\Big(\widetilde{Q}_n(p)\Big) = \frac{p(1-p)}{nf^2(Q(p))} + \frac{h^4\Big(f^{(1)}(Q(p))\Big)^2}{4f^2(Q(p))} - \frac{h}{nf(Q(p))}\varphi(k) + o\left(\frac{h}{n} + h^4\right).$$

▶ **Corollary 2.9.** The optimal bandwidth of $Amse\left(\widetilde{Q}_n(p)\right)$ is

$$h_{\text{opt}}\left(\widetilde{Q}_n(p)\right) = \left(\frac{f(Q(p))\varphi(k)}{n\left(f^{(1)}(Q(p))\right)^2\mu_2^2(k)}\right)^{\frac{1}{3}}.$$

◀

Parzen [35] proposed a version of the kernel quantile estimator as below:

$$\widehat{Q}_n(p) = \sum_{i=1}^{n}\left[\int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{1}{h}k\left(\frac{x-p}{h}\right)dx\right]X_{(i)}. \tag{2.11}$$

In practice, Yang [62] propose the following approximation to $\widehat{Q}_n(p)$ is often used:

$$\widehat{Q}_n^a(p) = \frac{1}{nh}\sum_{i=1}^{n} X_{(i)}k\left(\frac{\frac{i}{n}-p}{h}\right), \tag{2.12}$$

under suitable conditions on F, Falk [16] proposed the following kernel quantile estimator

$$\check{Q}_n(p) = \frac{1}{h}\int_0^1 Q_n(x)k\left(\frac{x-p}{h}\right)dx, \tag{2.13}$$

this kernel quantile estimator can then be approximated by $\widehat{Q}_n(p)$.
Yang [62]) provided the asymptotic normality property and the mean squared consistency of $\widehat{Q}_n(p)$ and proved that $\widehat{Q}_n(p)$ and $\widehat{Q}_n^a(p)$ are asymptotically equivalent in terms of mean square errors.
Falk [16] showed that the asymptotic performance of $\widehat{Q}_n(p)$ is better than that of the empirical sample quantile $Q_n(p)$ in terms of relative deficiency for appropriately chosen kernels and sufficiently smooth distribution functions.
Building on Falk [16], Sheater and al [49] gave the asymptotic mean squared error of $\widehat{Q}_n(p)$.

- If the second derivative of Q is continuous in a neighborhood of p and if F is not symmetric or F is symmetric but $p \neq \frac{1}{2}$ then Sheater and marron [49]

Provid the asymptotic properties of $\widehat{Q}_n(p)$ are

$$Bias(\widehat{Q}_n(p)) = \frac{h^2}{2}\left(Q^{(2)}(p)\right)\mu_2 + o(h^2),$$

and

$$Var(\widehat{Q}_n(p)) = \frac{p(1-p)}{n}\left(Q^{(1)}(p)\right)^2 - \frac{h}{n}\left(Q^{(1)}(p)\right)^2\varphi(k) + o\left(\frac{h}{n}\right).$$

Therefore

$$Amse\left(\widehat{Q}_n(p)\right) = \frac{p(1-p)}{n}\left(Q^{(1)}(p)\right)^2 + \frac{h^4}{4}\left(Q^{(2)}(p)\right)^2\mu_2^2 - \frac{h}{n}\left(Q^{(1)}(p)\right)^2\varphi(k)(k), \quad (2.14)$$

the optimal bandwidth for $Amse\left(\widehat{Q}_n(p)\right)$ is

$$h_{opt}\left(\widehat{Q}_n(p)\right) = \left(\frac{\left(Q^{(1)}(p)\right)^2\varphi(k)}{n\left(Q^{(2)}(p)\right)^2\mu_2^2}\right)^{\frac{1}{3}}. \quad (2.15)$$

- if F is symmetric and $p = \frac{1}{2}$ then the asymptotic mean squared error of $\widehat{Q}_n(p)$

$$Amse\left(\widehat{Q}_n(p)\right) = n^{-1}\left(Q^{(1)}\left(\tfrac{1}{2}\right)\right)^2\left[\frac{1}{4} - \frac{h}{2}\varphi(k) + \frac{1}{nh}\rho(k)\right],$$

where $\rho(k) = \int_{-\infty}^{\infty} k^2(x)dx$.

In this case, there is no single optimal bandwidth minimizing the Amse $\widehat{Q}_n(p)$.

Choice of the bandwidth We are interested in the choice of the smoothing parameter h of $\widehat{Q}(p)$ in the case where F is not symmetric or F is symmetric but $p \neq \frac{1}{2}$. Several data-based methods can be made to find the asymptotically optimal bandwidth in kernel quantile estimators for $\widehat{Q}_n$ given by (2.15). In practice, to evaluate an optimal global bandwidth, we need to develop a method to replace the

true derivatives of the quantile with her estimators. Note that

$$Q^{(1)}(p) = \frac{1}{f(Q(p))} \quad and \quad Q^{(2)}(p) = \frac{-f^{(1)}(Q(p))}{f^3(Q(p))},$$

can be estimated as follows (Jones [23]) :

$$\widehat{Q}_n^{(1)} = \frac{1}{h}\sum_{i=1}^{n} X_{(i)}\left(k\left(\frac{\frac{(i-1)}{n}-p}{h}\right) - k\left(\frac{\frac{(i)}{n}-p}{h}\right)\right),$$

and

$$\widehat{Q}_n^{(2)} = \frac{1}{h^2}\sum_{i=1}^{n} X_{(i)}\left(k^{(1)}\left(\frac{\frac{(i-1)}{n}-p}{h}\right) - k^{(1)}\left(\frac{\frac{(i)}{n}-p}{h}\right)\right).$$

Ali Al-Kenani [1] proposes a cross-validation method suitable for smoothing of kernel quantile estimators based on unbiased estimation of a mean integrated squared error curve of which the minimizing value determines an optimal bandwidth.
Note that when $h \longrightarrow 0$, $k(x) \longrightarrow \delta(x)$, where $\delta(.)$ Dirac delta function.
Now, from (2.12) when $h \longrightarrow 0$

$$\widehat{Q}_n(p) \longrightarrow \delta\left(\frac{i}{n} - p\right)X_{(i)},$$

where and thus a cross-validation function can be written as

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left\{\delta\left(\frac{i}{n} - p\right)X_{(i)} - \widehat{Q}_{-i}\left(\frac{i}{n}\right)\right\}^2 dp$$

The smoothing parameter h is then chosen to minimize this function. By subtracting a term that characterizes the performance of the true (p) we have

$$H(h) = \mathrm{CV}(h) - \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left\{\delta\left(\frac{i}{n} - p\right)X_{(i)} - Q\left(\frac{i}{n}\right)\right\}^2 dp$$

where the notation $\widehat{Q}_{-i}\left(\frac{i}{n}\right)$ with positive subscript denotes a kernel estimator based on a sample size of $n-1$. The proceeding arguments demonstrate that $CV(h)$ provides an asymptotic unbiased estimator of the true $MISE(h)$ curve for a sample size $n-1$.

► **Example 2.10.** we compare the performances of the two bandwidth selection methods through the Mse, by using Exponential (1) distribution for $n = 100$.

p	0.05	0.20	0.40	0.60	0.80	0.95
CV method	0.0006	0.0019	0.0058	0.0128	0.0373	0.1212
Sheater's method	0.0016	0.0021	0.0060	0.0138	0.0407	0.6668

we may conclude that in terms of Mise CV bandwidth selection method is more efficient than Sheather's method. ◄

3 Boundary correction problems in kernel estimation

Kernel estimation methods depend largely on the smoothing bandwidth, and very little depends on the type of kernel. It is well known that the performance of the classical kernel estimator at boundary points differs from the interior points even if we choose an adequate bandwidth due to so-called "boundary effects".

In order to deal with the boundary effects that occur in nonparametric, we note that the boundary problem in kernel density estimation has a non-consistency problem, in addition to the slow convergence problem of the bias (Gasser and al [20], Zhang and Karunamuni [63] and Bouredji and Sayah [6]). It is necessary to note that in the boundary region the estimator of the KDE is consistent. In such cases, modification of the KDF is needed to improve the Bias.

3.1 Boundary correction problems in kernel distribution estimation

It is very often the case that the natural support of a distribution to be estimated is not the whole real line but an interval bounded on one or both sides $[a, +\infty[$, $[-\infty, b[$ and $[a, b[$ where $a < b$. The boundary problem in kernel distribution estimation is less severe than in kernel density estimation, this is due to the extra information $F(a) = 0$, $F(b) = 1$ where $x \in [a, b]$. However, if we know that $f(a) = 0$ or $f(b) = 0$. Hence, the distribution kernel estimator $\widehat{F}_n(x)$ is free of boundary problems in such a case. In other cases, there has been intensive work in the literature about the Bias reduction in kernel distribution estimation, especially for the left boundary region in such a situation some methods are discussed among them

- **Generalized reflection method :** Kolăcek and Karunamuni [26] considered the boundary problem in distribution function estimation in estimating ROC curves using the transformation method discussed in Zhang and al [63]. The proposed estimator has the form

$$F_{n,roc}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - g_1(X_i)}{h}\right) + \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - g_2(X_i)}{h}\right), \tag{3.1}$$

where k is a kernel function with support $[-1;1]$, and $g_i, i = 1,2$ are two transformations nonnegative, continuous and monotonically increasing functions defined on $[0,+\infty[$, that need to be determined.
The trivial choice $g_1(y) = g_2(y) = y$ represents the classical reflection method estimator proposed by Horova and al [22]), other various improvements transformations can be found in Kolăcek and Karunamuni [26].
Under the assumption that $g^{-1}(0) = 1$ and $g^{(1)}(0) = 0$, where g^{-1} is the inverse function of g, the expectation value of the Bias and Variance of the estimator at $x = ch, 0 \leq c \leq 1$, are :

$$\begin{aligned} Bias\big(F_{n,roc}(x)\big) =& h^2\Bigg\{ f^{(1)}(0)\left(\frac{c^2}{2} + 2c\int_{-1}^{-c} K(t)dt - \int_{-c}^{c} tK(t)dt\right) \\ & - f(0)g_1^{(2)}(0)\int_{-1}^{c}(c-t)K(t)dt \\ & - f(0)g_2^{(2)}(0)\int_{-1}^{-c}(c+t)K(t)dt\Bigg\} + o\big(h^2\big), \end{aligned} \tag{3.2}$$

$$\begin{aligned} n\,\mathrm{Var}\big(F_{n,roc}(x)\big) =& F(x)(1-F(x)) + hf(0)\Bigg\{\int_{-1}^{c} K^2(t)dt \\ & - 2\int_{-1}^{c} K(t)K(t-2c)dt + \int_{-1}^{-c} K^2(t)dt\Bigg\} + o(h). \end{aligned}$$

- **A modified Champernowne transformation :** Tour and al [53] propose an estimator of heavy-tailed of F, based on ideas of the Generalized reflection method, and the work of Buch Larsen and al [8]. The transformation idea is based on transforming the original data by a new parametric transformation T, chosen by the modified Champernowne distribution function.
The modified Champernowne distribution is defined for $x \geq 0$ formulated as

$$T(x) = \frac{(x+c)^{\alpha} - e^{\alpha}}{(x+c)^{\alpha} + (M+c) - 2c^{\alpha}} \qquad x \geq 0, \tag{3.3}$$

with parameter $\alpha > 0, M > 0$ and $c \geq 0$.
Notice that $T_{a,M,0}(M) = 0.5$ this suggests that M can be estimated by the

empirical median (see Lehmann [29]).
In the other cases, $T(x)$ has a density given by

$$t(x) = \frac{\alpha(x+c)^{\alpha-1}((M+c)^{\alpha} - c^{\alpha})}{((x+c)^{\alpha} + (M+c) - 2c^{\alpha})^2} \quad x \geq 0,$$

the modified Champernowne distribution converges to a Pareto distribution in the tail:

$$t_{\alpha,M,c}(x) \longrightarrow \frac{\alpha((M+c)^{\alpha} - c^{\alpha})}{x^{\alpha+1}} \text{ as } x \longrightarrow \infty,$$

for more details about modified Champernowne distribution see for instance Buch Larsen and al [8].
The form of the estimator of the original data set, $X_1, X_2, \ldots, X_n$ is defined for $x = ch$, $0 \leq c \leq 1$ as,

$$F_a(x) = H_c(T(x)),$$

where

$$H_c(y) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{y - g(Y_i)}{h}\right) + \frac{1}{n}\sum_{i=1}^{n} K\left(-\frac{y + g(Y_i)}{h}\right),$$

thus $F_a(x)$ is a natural boundary continuation of the classical kernel distribution estimator.
then the expectation value of the Bias and variance are :

$$\begin{aligned}\text{Bias}(F_a(x)) = h^2\Bigg\{&\left(\frac{f}{T^{(1)}}\right)^{(1)}(0)\frac{1}{T^{(1)}(0)}\left(\frac{c^2}{2} + 2c\int_{-1}^{-c} K(t)dt - \int_{-c}^{c} tK(t)dt\right)\\ &-\frac{f(0)}{T^{(1)}(0)}g^{(2)}(0)\left(\int_{-1}^{c}(c-t)K(t)dt + \int_{-1}^{-c}(c+t)K(f)dt\right)\Bigg\} + o\left(h^2\right),\end{aligned} \tag{3.4}$$

and

$$\begin{aligned}\text{Var}(F_a(x)) = \frac{F(x)(1-F(x))}{n} + \frac{h}{n}\frac{f(0)}{T^{(1)}(0)}\Bigg\{&2\int_{-1}^{-c} K^2(t)dt - c\\ &+\int_{-c}^{c} K^2(t)dt - 2\int_{-1}^{c} K(t)K(t-2c)dt\Bigg\} + o\left(\frac{h}{n}\right).\end{aligned}$$

In the case where the support of the variable is $[a, b]$ we have

- **Boundary kernel method :** Tenreiro [51] proposed a boundary kernel method for correcting the boundary problem. However, Tenreiro [51] did not reveal the fact that there is no boundary problem in distribution function estimation if the density has a zero value at the endpoints of the support. In his method, the boundary kernel k_c is constructed by truncating a density kernel at $[-c, c]$, and then normalizing it so that it integrates to 1 on $[-c, c]$. Realizing the fact that such boundary kernel corrects the boundary problem by shrinking the bandwidth to zero when the data is near the boundary the resulting distribution estimates may have high variability at such points. Recently, Zhang and al [64] defined, the boundary distribution kernel estimator is defined for $x \in [a, b]$ as

$$F_B(x) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} K_c\left(\frac{x - X_i}{h}\right), & a \le x < a + h, c = (x-a)/h \\ \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), & a + h \le x \le b - h \\ \frac{1}{n}\sum_{i=1}^{n} K_c^*\left(\frac{x - X_i}{h}\right), & b - h \le x \le b, c = (b-x)/h \end{cases} \tag{3.5}$$

where k_c^* and k_c are the right and the left boundary kernel respectively and they must fulfill the following relation

$$\int_{-c}^{1} \frac{c+t}{c} k(t)dt = 1,$$

$$\int_{-1}^{c} \frac{c-t}{c} k(t)dt = 1.$$

Then the expectation value of the Bias and variance for $x = a + ch, \quad 0 \le c \le 1$ are

$$\text{Bias}(F_B(x)) = \frac{h^2}{2} f^{(1)}(a) \left[\int_{-1}^{c} (c-y)^2 k_c(y)dy - c^2 \right] + o\left(h^2\right) \tag{3.6}$$

and

$$\mathrm{Var}(F_B(x)) = \frac{2h}{n} f(a) \int_{-1}^{c} (c-y) k_c(y) K_c(y) dy + o\left(\frac{h}{n}\right).$$

For $x = b - ch, 0 \leq c \leq 1$,

$$\mathrm{Bias}(F_B(x)) = \frac{h^2}{2} f^{(1)}(b) \left[\int_{-c}^{1} (c+y)^2 k_c^*(y) dy - c^2 \right] + o\left(h^2\right) \tag{3.7}$$

and

$$\mathrm{Var}(F_B(x)) = \frac{2h}{n} f(b) \int_{-c}^{1} (c+y) k_c^*(y) \left[1 - K_c^*(y)\right] dy + o\left(\frac{h}{n}\right).$$

From (3.2), (3.6), (3.7) and (3.4) we turned out that the bias has been reduced to the second power of the bandwidth, while the bias of the kernel distribution function estimator has the first power of the bandwidth at the boundary, while the variance remains in the same order as the classical estimator.

3.2 Boundary correction problems in kernel inverse distribution estimation

It's well known that for high probabilities (0.95, 0.975 or 0.99), the classical estimators can be quite inefficient because have a large bias when p is close to 1. (see Wand and al [55], Jones and al [23], and reference therein).

- **Beta Kernel estimation** Harrell and Davis [21] or Park [34] suggest using the symmetric kernel, namely, the Beta-type kernel that as follows

$$HD_n(p) = \frac{\Gamma(n+1)}{\Gamma((n+1)p)\Gamma((n+1)(1-p)} \int_0^1 F_n^{-1}(y) y^{(n+1)p-1} (1-y)^{(n+1)(1-p)-1} dy,$$

where $F_n^{-1}(x)$ is the inverse of the empirical distribution function

$$F_n^{-1}(y) = \begin{cases} X_{(i)} & \text{if} \quad (i-1)/n < y \leq i/n \\ X_{(n)} & \text{if } 1 - 1/n < y < 1. \end{cases}$$

and Γ is the gamma function that is defined by.

$$\Gamma(k) = \int_0^{+\infty} x^{k-1} e^{-x} dx, \quad k > 0.$$

The $HD_n(p)$ estimator can be expressed as L-estimator

$$HD_n(p) = \sum_{i=1}^{n} w_{n,i}(p) X_{(i)},$$

where

$$w_{n,i}(p) = \frac{\Gamma(n+1)}{\Gamma((n+1)p)\Gamma((n+1)(1-p)} \int_{(i-1)/n}^{i/n} y^{(n+1)p-1}(1-y)^{(n+1)(1-p)-1} dy.$$

Notice that the expected value of the kth order statistic is given by

$$E\left(X_{(k)}\right) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} \int_0^1 Q(y) y^{k-1}(1-y)^{n-k} dy \ .$$

Asymptotic behavior of HD estimator
Harrel and Davis [21] show for F be an absolutely continuous distribution function with a strictly positive continuous density function f, such that

$$\int_{-\infty}^{+\infty} |x^{\alpha}| f(x) dx < \infty \text{ for some } \alpha > 0.$$

The HD estimator satisfies the same central limit theorem as does Q_n :

$$\sqrt{n}(HD_n(p) - Q(p)) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(Q(p))}\right), \quad \text{as } n \to \infty \text{ for } p \in]0,1[.$$

Based on simulation study Harrel and Davis [21] shows that Beta-kernel based estimators have very nice properties, both for light and heavy tails, for large or medium probability levels.

- **Beta Kernel estimation involving transform data :** In order to correct the bias problems in kernel quantile, Charpentier and Oulidi [9] suggested several nonparametric quantile estimators based on the beta-kernel and applied them to transform data by the generalized Champernowne distribution initially fitted to the data.
  Transforming observations Given a random variable Y, if H is a strictly increasing function, then the p-quantile of $H(Y)$ is equal to $H(Q(Y,p))$. Thus, an idea can be to transform initial observations $\{X_1,\dots,X_n\}$ into a sample $\{Y_1,\dots,Y_n\} = \{H(X_1),\dots,H(X_n)\}$ taking values in $[0,1]$, and then to use a beta-kernel based estimator, if $H : \mathbb{R} \longrightarrow (0,1)$. Then

$$Q_{ch,n}(p) = H^{-1}(Q_n(Y,p)),$$

  In theory, any transformation $H : \mathbb{R} \longrightarrow [0,1]$ should work. but Buch-Larsen and al [8] suggested to chose a transformation H such that $H(X)$ is closed to the uniform distribution. But since F is unknown, we need to find a distribution with nice goodness of fit properties, at least in tails (since we want to have a consistent estimate when p is close to one). And furthermore, since we want a standard procedure, we need a distribution that fits well losses and can be easily estimated. Thus, Buch-Larsen and al [8] suggested to set $Y_i = H(X_i)$ where H is a Champernowne distribution. Charpentier and Oulidi [9] use a monte Carlo study to explain the asymptotic behavior of $Q_{ch,n}$ in MSE criteria.

- **Champernowne transformation** Sayah and al. [43] proposed a new estimator of the quantile function, based on the modified Champernowne transformation noted by $TKQE$. The idea is to transform the initial data $\{X_1,\dots,X_n\}$ into $\{Z_1,\dots,Z_n\}$, where $Z_i := T(X_i)$, $i = 1,\dots,n$.
  This can be assumed to have been produced by a $(0,1)$-uniform rv Z.
  Thus, (2.11) yields the transformed kernel quantile estimator (TKQE)

$$\hat{Q}_{n,X}(p) := T^{-1}\Big(\hat{Q}_{n,Z}(p)\Big), \tag{3.8}$$

  where T^{-1} is the inverse of T (3.3) and

$$\hat{Q}_{n,z}(p) := \sum_{i=1}^{n} Z_{i,n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} K_h(z-p)dz \ .$$

The estimation procedure is described as follows:

1. Compute the estimates $(\hat{a}, \hat{M}, \hat{c})$ of the parameters of the modified Champernowne distribution (3.3).
   Then estimate the pair (α, c) which maximizes the log-likelihood function (see, Buch-Larsen and al. [8]):

$$l(\alpha, c) = n \log \alpha + n \log((M+c)^{\alpha} - c^{\alpha}) + (\alpha - 1) \sum_{i=1}^{n} \log(X_i + c)$$
$$- 2 \sum_{i=1}^{n} \log((X_i + c)^{\alpha} + (M+c)^{a} - 2c^{\alpha})$$

2. Transform the data $X_1, \ldots, X_n$ into $Z_1, \ldots, Z_n$ by

$$Z_i = T_{\bar{\alpha}, \bar{M}, \bar{c}}(X_i), \quad i = 1, \ldots, n$$

3. Using (2.11), calculate the kernel quantile estimator $\hat{Q}_{m,Z}(p)$ of the transformed data: $Z_1, \ldots, Z_n$.
4. The resulting TKQE of the original data $X_1, \ldots, X_n$ is given by

$$\hat{Q}_{n,x}(p) = T^{-1}_{a, \bar{X}, c}\left(\hat{Q}_{n,Z}(p)\right).$$

Then the Bias and the Variance of $\hat{Q}_{n,x}(p)$ are respectively

$$\text{Bias}\left(\hat{Q}_{n,x}(p)\right) = \frac{h^2}{2}\left[\left(T^{-1}\right)^{(2)}(Q_Z(p))\left(Q_Z^{(1)}\right)^2(p) + \left(T^{-1}\right)^{(1)}(Q_Z(p))Q_Z^{(2)}(p)\right]\mu_2(K) + o\left(h^2\right),$$

and

$$\text{Var}\left(\hat{Q}_{n,X}(p)\right) = \left(\left(T^{-1}\right)^{(1)}(Q_Z(p))Q_Z^{(1)}(p)\right)^2\left(\frac{p(1-p)}{n} - \frac{h}{n}\varphi(K)\right) + o\left(\frac{h}{n}\right),$$

where $\mu_2(K) := \int t^2 K(t)dt$, $\varphi(K) := 2\int tK(t)\left(\int_{-\infty}^{t} K(s)ds\right)dt$, $Q_Z^{(1)}$ and $Q_Z^{(2)}$ are the first and the second derivatives of Q_Z.

Part II

Main results

4 Nonparametric Kernel Distribution Function Estimation Near Endpoints

Abstract[1] In this paper, two kernel cumulative distribution function estimators are introduced and investigated in order to improve the boundary effects, we will restrict our attention to the right boundary. The first estimator uses a self-elimination between modify theoretical Bias term and the classical kernel estimator itself. The basic technique of construction the second estimator is kind of a generalized reflection method involving reflection a transformation of the observed data. The theoretical properties of our estimators turned out that the Bias has been reduced to the second power of the bandwidth, simulation studies and two real data applications were carried out to check these phenomena and are conducted that the proposed estimators are better than the existing boundary correction methods.
keywords : Boundary effects, Bias reduction, Cumulative distribution function, Kernel estimator.

4.1 Introduction

The cumulative distribution function F is used to determine the probability that a random observation X that is taken from an unknown population will be less than or equal to a certain x-value. Several approaches have been made to estimate this probability in this paper, we consider the classical kernel estimator F_n proposed by Nadaraya [32] defined for $X_1, X_2, ..., X_n$ a sample of a continuous real random variable by:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R} \tag{4.1}$$

such an estimator arises as an integral of kernel density estimator f_n which is introduced by Rosenblatt [42] and Parzen [35] that has the form:

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R} \tag{4.2}$$

where $h := h_n$ is a bandwidth that controls the smoothness of F_n and satisfying $h \to 0$ also $nh \to +\infty$ if $n \to +\infty$. The distribution function K is defined from a

[1] This chapter is a paper appeared in Advances in Mathematics: Scientific Journal 12:10 (2021), 3679–3697. Joint work with ALMI Nassima, SAYAH Abdallah.

kernel function k with the support $[-1, 1]$ as:

$$K(x) = \int_{-1}^{x} k(t)dt. \tag{4.3}$$

Many theoretical properties of F_n have been investigated among them, the uniform convergence of F_n to F with probability one, was proved by Winter [59] and Yamato [61], the asymptotic normality of F_n is established by Watson and Leadbetter [57] and an asymptotic expression for the mean squared error of F_n and the asymptotically optimal smoothing parameter proved by Azzalini [4]. These properties are satisfactory, but when the support of the variable is bounded kernel estimation may suffer. It is well known that F_n is a biased estimator near the boundary of its support, due to so-called boundary effects, this fact can be clearly seen by examining the behavior of F_n at interior points $]h, 1-h]$ and at the right boundary, $]1-h, 1]$. The value of Bias and Variance of F_n at interior points provided by Azzalini [4] are respectively:

$$\frac{1}{2}f^{(1)}(x)\mu_2(k)h^2 + o(h^2), \tag{4.4}$$

and

$$\frac{F(x)(1-F(x))}{n} + \frac{h}{n}f(x)\left(\int_{-1}^{1} K^2(t)dt - 1\right) + o\left(\frac{h}{n}\right), \tag{4.5}$$

where $\mu_2(k) = \int t^2 k(t)dt$ and $f^{(1)}$ denote the first derivative of f.

However, in the right boundary, we assume $x = 1 - ch$ where $0 \leq c < 1$, then the Bias and Variance of F_n at x are respectively:

$$-hf(1)\int_{-1}^{-c} K(t)dt + h^2 f^{(1)}(1)\left(\frac{c^2}{2} - \int_{-1}^{c} tK(t)dt + c\int_{-1}^{-c} K(t)dt\right) + o(h^2), \tag{4.6}$$

and

$$\frac{F(x)(1-F(x))}{n} + \frac{h}{n}f(1)\left(-c - 2\int_{-1}^{-c} K(t)dt + \int_{-1}^{c} K^2(t)dt\right) + o\left(\frac{h}{n}\right). \tag{4.7}$$

In the results, we can see that for densities taking value zero at the endpoints of the support the first order term in (4.6) disappears and the Bias converges to zero at the usual rate $o(h^2)$. Otherwise, the Bias of F_n is of order $o(h^2)$ at the interior instead is of order $o(h)$ near the right boundary points this is the boundary problem of the kernel distribution estimator. In order to correct this problem, many methods have been proposed for kernel estimation in regression and density function estimation, among them, reflection of data Silverman [46], pseudo-data method Cowlling [10] and also the boundary kernel method gasser and marron [20]. However, methods in kernel distribution function estimation are relatively few, this is due to the extra information $F(0) = 0$ and $F(1) = 1$. Karunamuni and al [26] considered this problem in estimating ROC curves using the transformation method, Tour and al [53] used a Champernowne transformation for heavy-tailed distributions in the left side of the support, and Tenreiro [51] and Zhang and al [64] proposed a boundary kernel method free of boundary problems. In this paper, we propose two estimators for kernel distribution function to improve the right boundary effects.

The rest of the paper is organized as follows. Notations and theoretical properties of the proposed estimators are introduced in Section 2. In Section 3 we support the theoretical results by simulation studies and two real data applications. The paper is finalized with some concluding remarks.

4.2 Assumptions and main results

For each result in this section, one at least of the following two assumptions will be used

- A_1 : F is twice continuously differentiable in a neighborhood of x and $f(1) \neq 0$.
- A_2: The kernel k is a probability density, nonnegative, bounded, symmetric, and has compact support $[-1, 1]$.

▶ **Remark 4.1.** If x is a point in the right boundary, we can write $x = 1 - ch$ where $c \in [0, 1[$ therefore we have $1 - ch > h$. ◀

4.2.1 Modify Bias of Kernel Estimator

In the context of Bias reduction in distribution estimation, our proposed estimator $\breve{F}_n$ consists to subtract the modify of the theoretical $Bias(F_n(x))$ term (4.6) from F_n itself when the data near the right boundary of the support for $x = 1 - ch$ defined by

$$\breve{F}_n(x) = F_n(x) + h\Psi(c)f_n(x) + h^2\alpha f_n^{(1)}(x), \tag{4.8}$$

where $f_n^{(1)}$ denote to the first derivative of kernel density estimator. Then the explicit form of our estimator is given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) + h\Psi(c)\left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)\right) + h^2\alpha\left(\frac{1}{nh^2}\sum_{i=1}^{n} k^{(1)}\left(\frac{x - X_i}{h}\right)\right),$$

where $k^{(1)}$ is the first derivative of kernel k, α is a positive constant and $\Psi(c)$ to be determined in the following proof in such a way the terms of h in the Bias vanish.

▶ **Theorem 4.2.** Under the above assumptions A_1 and A_2 we obtain at $x = 1 - ch$

$$Bias(\breve{F}_n(x)) = h^2 f^{(1)}(1)\phi(c) + o(h^2), \tag{4.9}$$

$$Var(\breve{F}_n(x)) = \frac{F(x)(1 - F(x))}{n} + \frac{h}{n}f(1)\rho(c) + o\left(\frac{h}{n}\right), \tag{4.10}$$

where

$$\phi(c) = \frac{c^2}{2} - \int_{-1}^{c} tK(t)dt + \int_{-1}^{-c} cK(t)dt - \int_{-c}^{1} ((t + c)\Psi(c) - \alpha)k(t)dt,$$

$$\rho(c) = c - \int_{-1}^{c} K^2(t)dt + 2\int_{-1}^{-c} K(t)dt - \int_{-c}^{1} \left(\Psi(c)k(t) + \alpha k^{(1)}(t)\right)^2 dt$$

$$-2\int_{-c}^{1} \left(\Psi(c)k(t) + \alpha k^{(1)}(t)\right)K(t)dt.$$

Additionally, it can be seen that the optimal bandwidths h^*_{opt} for minimizing Mse is :

$$h^*_{opt} = \left(\frac{f(1)\rho(c)}{4n\left(f^{(1)}(1)\phi(c)\right)^2} \right)^{\frac{1}{3}} ,$$

◀

Proof. For $x \in]1-h, 1]$, we have

$$E(\check{F}_n(x)) = E(F_n) + h\Psi(c)E(f_n(x)) + h^2\alpha E\left(f_n^{(1)}(x)\right).$$

We calculate each term separately

$$\begin{aligned} E(F_n(x)) &= \int_0^1 K\left(\frac{x-z}{h}\right) f(z)dz \\ &= h\int_c^{\frac{1}{h}-c} K(t)f(x-th)dt + h\int_{-c}^{c} K(t)f(x-th)dt, \end{aligned}$$

by using the remark (4.1) and the property $K(t) = 1-K(-t)$ on the first integration, we have

$$E(F_n(x)) = F(1-2ch) - h\int_{-1}^{-c} K(t)f(x+th)dt + h\int_{-c}^{c} K(t)f(x-th)dt,$$

depending on a Taylor expansion and some algebraic calculation, we have

$$E(F_n(x)) = F(x) - hf(1)\int_{-1}^{-c} K(t)dt + h^2 f^{(1)}(1)\left(\frac{c^2}{2} - \int_{-1}^{c} tK(t)dt + c\int_{-1}^{-c} K(t)dt\right) + o(h^2).$$

This is proof the relation(4.6).
By the same procedure, we have

$$E(f_n(x)) = f(1)\int_{-c}^{1} k(t)dt - hf^{(1)}(1)\int_{-c}^{1} (t+c)k(t)dt + o(h),$$

and

$$E\left(f_n^{(1)}(x)\right) = f^{(1)}(1)\int_{-c}^{1} k(t)dt + o(1).$$

At last, we combine all terms, we obtain

$$E(\breve{F}_n(x)) = F(x) + hf(1)\left(-\int_{-1}^{-c} K(t)dt + \Psi(c)\int_{-c}^{1} k(t)dt\right) + h^2 f^{(1)}(1)\left(\frac{c^2}{2} - \int_{-1}^{c} tK(t)dt + \int_{-1}^{-c} cK(t)dt - \int_{-c}^{1} (\Psi(c)(t+c) - \alpha)k(t)dt\right) + o(h^2),$$

therefore, $E(\breve{F}_n(x))$ can be improved the Bias by letting the terms in h, vanish if and only if we choice $\Psi(c)$ by

$$\Psi(c) = \frac{\int_{-1}^{-c} K(t)dt}{\int_{-c}^{1} k(t)dt}.$$

This completes the proof of expression (4.9).
On the other hand

$$\begin{aligned} Var\left(\breve{F}_n(x)\right) &= \frac{1}{n}E\left(K\left(\frac{x-X_i}{h}\right) + h\Psi(c)k\left(\frac{x-X_i}{h}\right) + \alpha h^2 k^{(1)}\left(\frac{x-X_i}{h}\right)\right)^2 \\ &- \frac{1}{n}\left(E\left(K\left(\frac{x-X_i}{h}\right) + h\Psi(c)k\left(\frac{x-X_i}{h}\right) + \alpha h^2 k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right)^2 \\ &= J_{11} + J_{12} + J_{13} + J_{14} + J_{15} + J_{16}, \end{aligned}$$

where

$$\begin{aligned} J_{11} &= \frac{1}{n}E\left(K^2\left(\frac{x-X_i}{h}\right)\right) - \frac{1}{n}\left(E\left(K\left(\frac{x-X_i}{h}\right)\right)\right)^2 \\ &= \frac{h}{n}\int_{-c}^{\frac{1}{h}-c} K^2(t)f(x-th)dt - \frac{1}{n}F^2(x) + o\left(\frac{h}{n}\right) \\ &= \frac{h}{n}\int_{-c}^{c} K^2(t)f(x-th)dt + \frac{h}{n}\int_{c}^{\frac{1}{h}-c} (1-K(-t))^2 f(x-th)dt - \frac{1}{n}F^2(x) + o\left(\frac{h}{n}\right), \end{aligned}$$

by Taylor expansion, we have

$$\begin{aligned} J_{11} &= \frac{F(x)(1-F(x))}{n} + \frac{h}{n}f(1)\left(-c - 2\int_{-1}^{-c} K(t)dt + \int_{-1}^{c} K^2(t)dt\right) + o\left(\frac{h}{n}\right), \\ &= Var(F_n(x)). \end{aligned}$$

This is proof of the relation (4.7).

$$
\begin{aligned}
J_{12} &= \frac{1}{n}E\left(h\Psi(c)\left(\frac{1}{h}k\left(\frac{x-X_i}{h}\right)\right)\right)^2 - \frac{1}{n}E^2\left(h\Psi(c)\left(\frac{1}{h}k\left(\frac{x-X_i}{h}\right)\right)\right) \\
&= \frac{h}{n}(\Psi(c))^2 f(1)\int_{-c}^{1} k^2(t)dt + o\left(\frac{h}{n}\right),
\end{aligned}
$$

$$
\begin{aligned}
J_{13} &= \frac{h^4\alpha^2}{n}E\left(\left(\frac{1}{h^2}k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right)^2 - \frac{1}{n}\left(E\left(\alpha h^2\left(\frac{1}{h^2}k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right)\right)^2 \\
&= \frac{h\alpha^2}{n}f(1)\int_{-c}^{1}\left(k^{(1)}(t)\right)^2 dt + o\left(\frac{h}{n}\right),
\end{aligned}
$$

$$
\begin{aligned}
J_{14} &= \frac{2}{n}h\Psi(c)\left(E\left(\frac{1}{h}K\left(\frac{x-X_i}{h}\right)k\left(\frac{x-X_i}{h}\right)\right) - E\left(\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right)E\left(k\left(\frac{x-X_i}{h}\right)\right)\right) \\
&= \frac{2}{n}h\Psi(c)f(1)\int_{-c}^{1} k(t)K(t)dt + o\left(\frac{h}{n}\right),
\end{aligned}
$$

$$
\begin{aligned}
J_{15} &= \frac{2\alpha}{n}h^2\left(E\left(\frac{1}{h^2}K\left(\frac{x-X_i}{h}\right)k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right. \\
&- \left. E\left(K\left(\frac{x-X_i}{h}\right)\right)E\left(\frac{1}{h^2}k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right) \\
&= \frac{2\alpha h f(1)}{n}\left(\int_{-c}^{1} k^{(1)}(t)K(t)dt\right) + o\left(\frac{h}{n}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
J_{16} &= \frac{2\alpha\Psi(c)h^3}{n}\left(E\left(\frac{1}{h}k\left(\frac{x-X_i}{h}\right)\frac{1}{h^2}k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right) \\
&- E\left(\frac{1}{h}k\left(\frac{x-X_i}{h}\right)\right)E\left(\frac{1}{h^2}k^{(1)}\left(\frac{x-X_i}{h}\right)\right) \\
&= \frac{2\alpha h\Psi(c)}{n}f(1)\int_{-c}^{1} k(t)k^{(1)}(t)dt + o\left(\frac{h}{n}\right).
\end{aligned}
$$

This completes the proof of expression (4.10). ■

4.2.2 Reflection Transformation Kernel Estimator

The technique of generalized reflection method involving reflecting a transformation of the observed data in kernel distribution estimation used by [26] when the data is near the left side of the support. Our proposed estimator $\widehat{F}_n$ developed this technique when the data near the right boundary of the support, given for $x \in]1-h, 1]$ by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - g(X_i)}{h}\right) + \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - 2 + g(X_i)}{h}\right), \tag{4.11}$$

where g is a transformation which is selected from a parametric family, we assume that verify:

- $H_1 : g$ is a continuous and monotonically increasing function from $[0, 1]$ to $[0, 1]$.

- $H_2 : g^{-1}$ exist and verify $g^{-1}(1) = 1$ and $g^{(1)}(1) = 1$ where g^{-1} and $g^{(1)}$ denoting respectively the inverse and the first derivative function of g.

It is clear that there are various possible choices available for the function g that satisfy the above assumptions. Based on extensive simulations, we choose the following transformation g which well adapts to various shapes of distributions and improve the Bias

$$g(t) = t - t(1-t)^2 \int_c^1 K(t)dt, \qquad c \in [0, 1[.$$

▶ **Theorem 4.3.** Under the above assumptions A_1, $A2$, H_1 and H_2, the asymptotic properties of our proposed estimator $\widehat{F}_n$ at $x = 1 - ch$ are

$$Bias(\widehat{F}_n(x)) = h^2\Gamma(c) + o(h^2), \tag{4.12}$$

and

$$Var(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n} + \frac{h}{n}f(1)\Omega(c) + o\left(\frac{h}{n}\right), \tag{4.13}$$

therefore, the value of h^{**}_{opt} which is the bandwidth that minimizes the Mse is

$$h_{opt}^{**} = \left(\frac{(f(1)\Omega(c))^4}{4n\Gamma(c)} \right)^{1/3},$$

where

$$\Gamma(c) = \frac{-c^2}{2} f^{(1)}(1) + \left(f^{(1)}(1) - g^{(2)}(1) f(1) \right) \left(-2c^2 + 2c \int_{-1}^{-c} K(t)dt - \int_{-c}^{c} K(t)(t+c)dt \right),$$

and

$$\Omega(c) = -c + \int_{-1}^{c} K^2(t)dt + \int_{-1}^{-c} K(t)(K(t)-2)dt + 2\int_{-c}^{1} K(t)K(-2c-t)dt.$$

◀

Proof. For $x \in]1-h, 1]$, we have

$$\begin{aligned} E(\widehat{F}_n(x)) &= E\left(K\left(\frac{x - g(X_i)}{h} \right) \right) + E\left(K\left(\frac{x - 2 + g(X_i)}{h} \right) \right) \\ &= \int_0^1 K\left(\frac{x - g(z)}{h} \right) f(z)dz + \int_0^1 K\left(\frac{x - 2 + g(X_i)}{h} \right) f(z)dz \\ &= I_1 + I_2, \end{aligned}$$

where

$$\begin{aligned} I_1 &= \int_0^1 K\left(\frac{x - g(z)}{h} \right) f(z)dz, \\ &= h \int_{-c}^{\frac{1}{h}-c} K(t) \frac{f(g^{-1}(x-th))}{g^{(1)}(g^{-1}(x-th))} dt, \\ &= h \int_{c}^{\frac{1}{h}-c} K(t) \frac{f(g^{-1}(x-th))}{g^{(1)}(g^{-1}(x-th))} dt + h \int_{-c}^{c} K(t) \frac{f(g^{-1}(x-th))}{g^{(1)}(g^{-1}(x-th))} dt, \end{aligned}$$

by using the property $K(t) = 1 - K(-t)$ on the first integration we have

$$I_1 = F(g^{-1}(1-2ch)) - h\int_{\frac{-1}{h}+c}^{-c} K(t)\frac{f(g^{-1}(x+th))}{g^{(1)}(g^{-1}(x+th))}dt + h\int_{-c}^{c} K(t)\frac{f(g^{-1}(x-th))}{g^{(1)}(g^{-1}(x-th))}dt,$$

we use a Taylor expansion of the function $F(g^1(.))$,

$$\begin{aligned} F\big(g^{-1}(1-2ch)\big) &= F(g^{-1}(1)) - 2hc\frac{f(g^{-1}(1))}{g^{(1)}(g^{-1}(1))} \\ &+ 2(ch)^2\left(\frac{f^{(1)}\big(g^{-1}(1)\big)g^{(1)}\big(g^{-1}(1)\big) - g^{(2)}\big(g^{-1}(1)\big)f\big(g^{-1}(1)\big)}{\big[g^{(1)}(g^{-1}(1))\big]^3}\right) + o(h^2). \end{aligned}$$

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

$$F(1) = F(x) + chf(x) + \tfrac{1}{2}(ch)^2 f^{(1)}(x) + o\big(h^2\big).$$

$$f(x) = f(1) - chf^{(1)}(1) + o(h)$$

$$f^{(1)}(x) = f^{(1)}(1) + o(1).$$

Therefore

$$F\big(g^{-1}(1-2ch)\big) = F(x) - chf(1) + \frac{3(ch)^2}{2}f^{(1)}(1) - 2(ch)^2\Big(g^{(2)}(1)f(1)\Big) + o(h^2).$$

Eventually, we obtain

$$\begin{aligned} I_1 &= F(x) - \frac{(ch)^2}{2}f^{(1)}(1) - hf(1)\int_{-1}^{-c} K(t)dt \\ &+ h^2(f^{(1)}(1) - f(1)g^{(2)}(1))\left(-2c^2 + \int_{-1}^{-c}(c-t)K(t)dt - \int_{-c}^{c}(c+t)K(t)dt\right) + o(h^2). \end{aligned}$$

Similar computation give I_2

$$\begin{aligned} I_2 &= \int_0^1 K\left(\frac{x-2+g(z)}{h}\right)f(z)dz \\ &= h\int_{-1}^{-c}\frac{f(g^{-1}(2-x+th))}{g^{(1)}g^{-1}(2-x+th)}K(t)dt, \end{aligned}$$

we use a Taylor expansion of the function $\frac{f(g^{-1}(.))}{g^{(1)}(g^{-1}(.))}$, we obtain

$$I_2 = hf(1)\int_{-1}^{-c} K(t)dt + h^2\left(f^{(1)}(1) - g^{(2)}(1)f(1)\right)\int_{-1}^{-c}(t+c)K(t)dt + o(h^2).$$

We combine I_1 and I_2 we obtain the expression of $Bias(\widehat{F}_n)$ (4.12).
To prove (4.13), note that

$$\begin{aligned} nVar(\widehat{F}_n) &= E\left(K\left(\frac{x-g(X_i)}{h}\right)+K\left(\frac{x-2+g(X_i)}{h}\right)\right)^2 \\ &- \left(E\left(K\left(\frac{x-g(X_i)}{h}\right)+K\left(\frac{x-2+g(X_i)}{h}\right)\right)\right)^2 \quad \text{where} \\ &= A_1 - A_2, \\ A_1 &= E\left(K\left(\frac{x-g(X_i)}{h}\right)+K\left(\frac{x-2+g(X_i)}{h}\right)\right)^2, \\ &= A_{11} + A_{12} + 2A_{13}, \end{aligned}$$

it can be shown that

$$\begin{aligned} A_{11} &= \int_0^1 K^2\left(\frac{x-g(z)}{h}\right)f(z)dz \\ &= h\int_c^{\frac{1}{h}-c} K^2(t)\frac{f(g^{-1}(x-th))}{g^{(1)}(g^{-1}(x-th))}dt + h\int_{-c}^{c} K^2(t)\frac{f(g^{-1}(x-th))}{g^{(1)}(g^{-1}(x-th))}dt \\ &= h\int_c^{\frac{1}{h}-c} (1-K(-t))^2\frac{f(g^{-1}(x-th))}{g^{(1)}(g^{-1}(x-th))}dt + h\int_{-c}^{c} K^2(t)\frac{f(g^{-1}(x-th))}{g^{(1)}(g^{-1}(x-th))}dt, \end{aligned}$$

by Taylor expansion, we have

$$A_{11} = F(x) + hf(1)\left(-c + \int_{-1}^{c} K^2(t) - 2\int_{-1}^{-c} K(t)dt\right) + o(h),$$

and similarly, we obtain

$$A_{12} = \int_0^1 K^2\left(\frac{x-2+g(z)}{h}\right) f(z)dz$$
$$= hf(1)\int_{-1}^{-c} K^2(t)dt + o(h),$$

and

$$A_{13} = hf(1)\int_{-c}^{1} K(t)K(-2c-t)dt + o(h).$$

we combine A_{11}, A_{12} and A_{13} to obtain A_1.
With the expression of the $Bias(\widehat{F}_n)$, we find:

$$A_2 = \left(E\left(K\left(\frac{x-g(X_i)}{h}\right) + K\left(\frac{x-2+g(X_i)}{h}\right)\right)\right)^2$$
$$= F^2(x) + o(h).$$

This completes the proof of expression (4.13) ■

4.3 Simulation study

A simulation study presented in this section to support the theoretical results of the proposed estimators, which was made through the comparison of the asymptotic properties of our estimators with the existing estimators summarized in the coming subsection. For each estimator, we evaluate the Bias and Mse at the right boundary from different distributions with support $[0, 1]$ are listed in the table (4.1). To be more specific, for each distribution we generated $\{X_1, X_2, ..., X_n\}$ a sample of size $n = 200$ and we did $r = 1000$ replication by using software R. Let $\hat{\theta}_i$ be estimator of θ based on the i^{th} generated random numbers of size n then Bias and Mse are estimated by

$$Bias(\hat{\theta}) = \frac{1}{r}\sum_{i=1}^{r}\left(\hat{\theta}_i(x) - \theta(x)\right),$$

$$Mse(\hat{\theta}) = \frac{1}{r}\sum_{i=1}^{r}\left(\hat{\theta}_i(x) - \theta(x)\right)^2.$$

We ran a cross-validation method [37] to choose bandwidth for the Epanechnikov kernel, the main reason for this choice is that it provides a fair basis for comparison among the different estimators without regard to bandwidth effects.

4.3.1 Existing estimators used in comparison

In this subsection, we briefly discuss existing distribution kernel estimators and propose important modifications.
For the first estimator (denote it by $\bar{F}_n$), inspired from the generalized reflection kernel distribution estimator Karunamuni et al [27], we find

$$\bar{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) + \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - 2 + X_i}{h}\right).$$

The second estimator (denote it by $\tilde{F}_n$), considers the boundary-modified kernel distribution function estimator suggested by Zhang et al [64].

$$\tilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_c\left(\frac{x - X_i}{h}\right),$$

where K_c is a kernel distribution function, and k_c satisfying

$$\int_{-c}^{1} \frac{c + x}{c} k_c(x)dx = 1,$$

for the Epanechnikov kernel we choice

$$k_c(t) = 12\frac{1 - t}{(1 + c)^4}\left(\frac{3c^2 - 2c + 1}{2} - t(1 - 2c)\right), -c \leq t \leq 1.$$

To account these estimators for different situations, we use distributions summarized in table (4.1), Note that the densities function D_4,D_5 and D_6 satisfies $f(0) = f(1) = 0$.

The simulation results measure the performance of the different estimators in the meaning of the Bias and Mse, are summarized in tables (4.2) and (4.3).

From Table (4.2), we can see that all the kernel distribution estimators previously mentioned have smaller Bias than the classical kernel distribution estimator F_n. Comparing among the kernel distribution estimators, we see that the reflection transformation estimator $\widehat{F}_n$ has a smaller Bias for the almost used distribution, except in the case of truncated exponential, the boundary distribution kernel estimator $\tilde{F}_n$ has an asymptotically smaller Bias when compared with our proposed estimator $\widehat{F}_n$. The comparison of the modify Bias of kernel estimator $\breve{F}_n$ depends on the choice of the positive constant α. When α is relatively small α=0.1 we can see that $\breve{F}_n$ has roughly the same Bias as F_n and when α increases gradually, $\breve{F}_n$

Table 4.1: Distributions used in the simulation study

	Description	Density for $x \in [0,1]$
D_1	Truncated Normal(0,1)	$exp(-x^2/2)/\int_0^1 exp(-x^2/2)dx.$
D_2	Truncated Exponential(3)	$3exp(-3x)/(1-exp(-3)).$
D_3	Truncated Exponential(0.02)	$(0.02)exp(-0.02x)/(1-exp(-0.02)).$
D_4	Truncated Beta$(2,2)_{[\frac{1}{3};1]}$	$4.05x(1-x)$
D_5	Kumaraswamy(4,2)	$8x^3(1-x^4)$
D_6	Beta(4,2)	$20x^3(1-x)$
D_7	Beta(3,1)	$3x^2$
D_8	Uniform(0,1)	1

Table 4.2: Bias values at x=1, Results are re-scaled by the factor 0.001.

	F_n	$\bar{F}_n$	$\tilde{F}_n$	$\widehat{F}_n$	$\breve{F}_n$		
				$\alpha =$	0.1	10	100
D_1	7.3783	2.8055	3.0703	2.7219	7.3780	5.7302	2.9829
D_2	5.27695	4.7483	8.70152	0.2531	5.2506	3.4679	0.3246
D_3	3.1702	2.19568	2.8583	2.1256	3.1621	3.1548	2.9564
D_4	6.2859	6.3277	6.6596	5.0277	6.2836	5.5245	5.0252
D_5	1.7211	1.6835	1.6731	1.6720	1.7211	1.7012	1.7005
D_6	3.5881	2.4585	2.4521	2.3023	3.5811	2.4012	2.3505
D_7	5.2351	3.6521	4.5231	1.6731	5.4587	4.6812	2.6802
D_8	0.1404	0.1306	0.1370	0.1285	0.1434	0.1374	0.1298

improve the performance of the estimator. For the other estimators in general, the boundary distribution kernel estimator $\tilde{F}_n$ has a second smaller Bias followed by the reflection estimator $\bar{F}_n$. From Table (4.3), our proposed estimator $\widehat{F}_n$ has an asymptotically smaller *Mse* when compared with the other estimators, which they organized in the sens of *Mse* by $\tilde{F}_n$ followed by $\bar{F}_n$ followed by $\breve{F}_n$ which is less than F_n for the almost used distribution.

4.4 Real data application

The aim of our applications is to compare the performance of the two proposed kernel distribution estimators given respectively in (4.8) and (4.11) using the cross-

Table 4.3: Mse values at x=1, Results are re-scaled by the factor 0.001.

	F_n	$\bar{F}_n$	$\tilde{F}_n$	$\widehat{F}_n$	$\breve{F}_n$		
				$\alpha =$	0.1	10	100
D_1	2.5926	1.8345	1.8321	1.8021	2.5912	2.3147	1.8745
D_2	1.7097	1.5795	1.5767	1.5710	1.7034	1.6524	1.6314
D_3	1.9258	1.9177	1.9124	1.9102	1.9258	1.9247	1.9235
D_4	1.8206	1.6904	1.7124	1.6814	1.8204	1.8045	1.7352
D_5	0.5641	0.5641	0.5639	0.5635	0.5641	0.5641	0.5638
D_6	2.2535	2.2012	2.1540	2.1201	2.2445	2.2354	2.1721
D_7	4.1521	3.2155	2.1325	0.1284	4.2354	3.4521	1.2572
D_8	0.4441	0.3897	0.3175	0.2210	0.4378	0.4102	0.3548

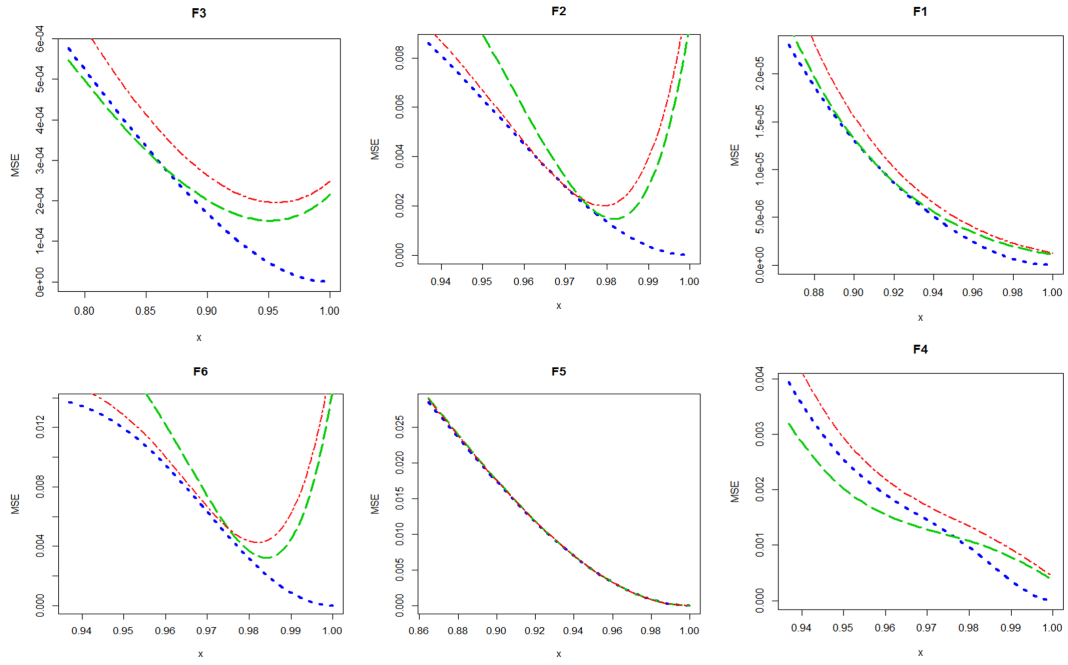

Figure 4.1: Mse of different estimators

validation method to bandwidth selection for two real data sets, in order to demonstrate its usefulness in practical application. The first data set X consists of the number of deaths due to COVID-19 recorded from february 29, 2020 to December 31,2020 in 50 states of the United States of America taken from www.nytimes.com, where $X_i \in [0, 3808]$. The second data set taken from [30] represents the failure

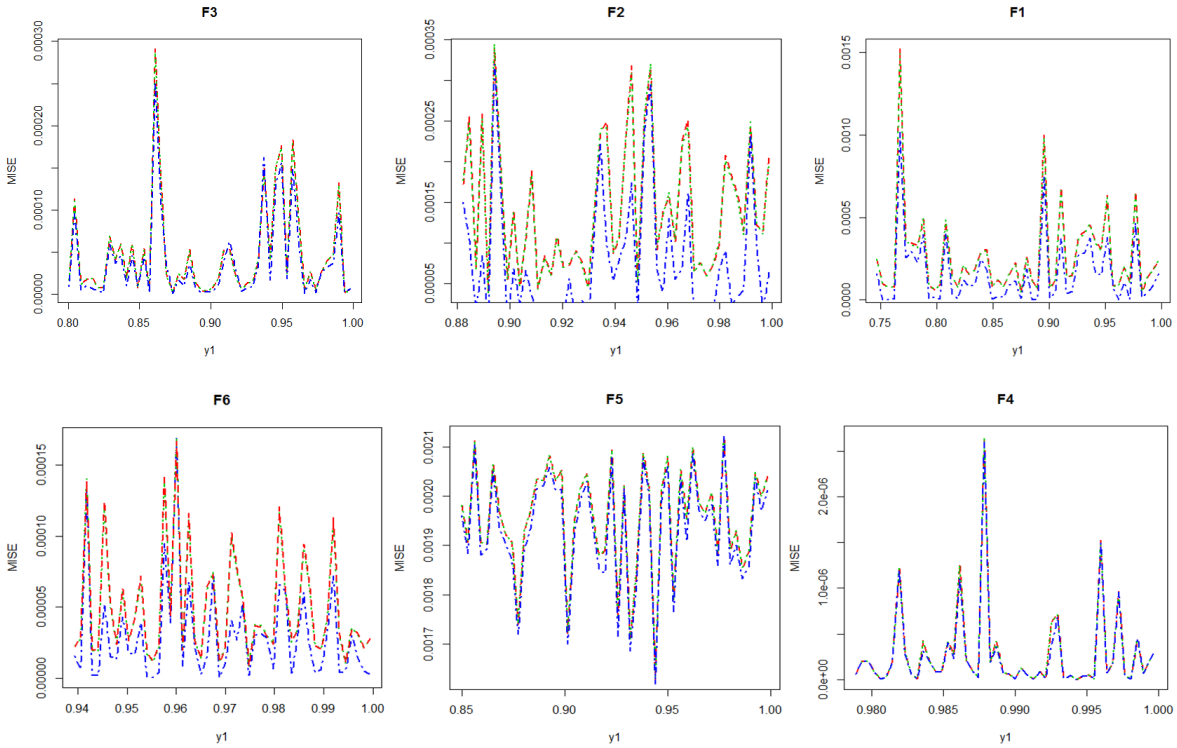

Figure 4.2: Mise of different estimators

times of the air conditioning system of an airplane, it consists of 30 observations in $[1.68, 6.81]$. For each data set we can be mapped onto the unit interval by the transformation $Z_i = (X_i - a)/(b - a)$, where $\{X_i\}$ a real observation in $[a, b]$. The table below gives a basic statistical description of the real data sets, a quick analysis of this table provides a preliminary insight concerning the distribution of data.

Table 4.4: Basic statistical description of real data sets

	Mean	Median	Skewness	Kurtosis	Std.error	Std.deviation
First data	0.2972	0.2578	1.0413	3.9265	0.0117	0.2058
Second data	0.5156	0.5263	-0.4167	3.0934	0.0181	0.1985

We have plotted the performance of our estimators and compared them to the previous mentioned estimators. In figure (4.3), we denote by red line to the classical estimator, green line to the modify Bias and bleu line to the reflection transformation, cyan line to boundary modified and pink line to reflection kernel distribution

estimator. We see that our estimators well distributed over $]1-h,1]$, the performance of $\check{F}_n$ estimator improves when the positive constant α is large in this graph we chose $\alpha = 10$. It is remarkably clear that our estimators remove the boundary effect and has improved the performance of the classical estimator when the data near the right boundary.

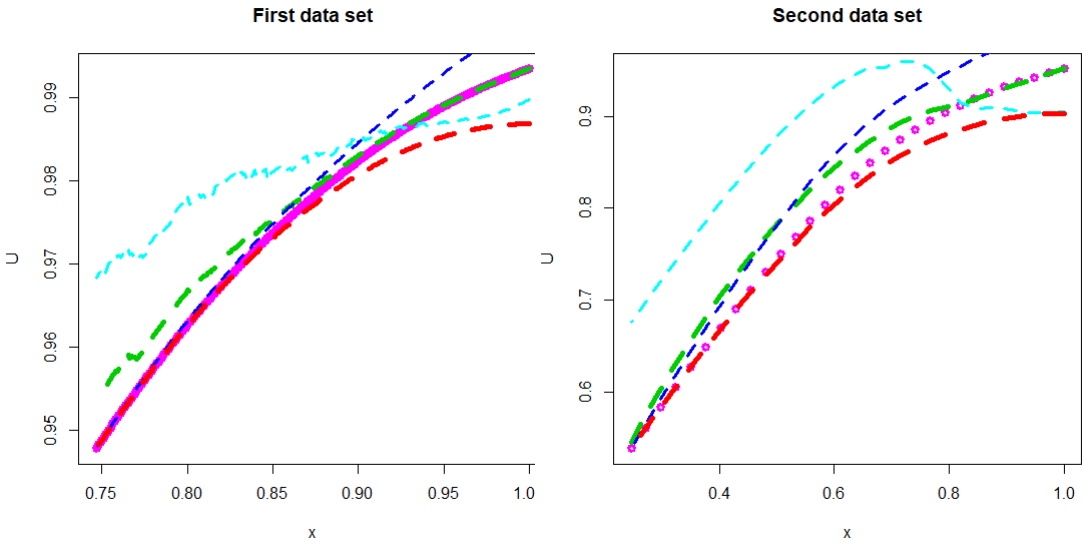

Figure 4.3: Performance of different estimators in real applications

4.5 Conclusions

The kernel method is an intuitive simple, and useful procedure, especially in density and distribution function estimation. When the support of the random variable is bounded, this procedure needs modification. In this paper, we proposed two new kernel distribution estimators to avoid the difficulties near the right boundary, by using two techniques that have been inspired from boundary correction methods. Depending on the theoretical and simulation results, it turned out that our proposed estimators have been reducing the Bias to the second power of the bandwidth, which is smaller than estimators have considered in this paper.

5 Estimating the Inverse Distribution Function at the Boundary

Abstract[1]. Most of the existing quantile estimators have problems of inefficiency in extreme quantiles. To solve this problem, In this paper, we suggested an alternative estimator and provided its asymptotic behavior when quantile near the boundary value. Simulation studies and two real data applications were included to demonstrate the efficiency and reliability of our theoretical results.
keywords: Kernel quantile estimation, Mean Square Error, Optimal Bandwidth, Boundary Quantiles.

5.1 Introduction

The estimation of population quantiles is of great interest when a parametric form for the underlying distribution is not available. It plays an important role in both statistical and probabilistic applications, namely: the goodness-of-fit, the computation of extreme quantiles and Value-at-Risk in insurance which are important measures of random performance, business and financial risk management, in reliability and medical studies, quantiles are adopted for characterize the survival distribution. Also, a large class of actuarial risk measures can be defined as functionals of quantiles (see, e.g. [12]).

Let $X_1, ..., X_n$ be independent and identically distributed with an unknown density $f(.)$ and absolutely continuous distribution function $F(.)$, while $X_{(1)} \leq ... \leq X_{(n)}$ denote the corresponding order statistics. The quantile function $Q(.)$ is defined to be the left-continuous inverse of $F(.)$ as follows:

$$Q(p) = \inf\{x : F(x) \geq p\} = F^{-1}(p), \ 0 < p < 1. \tag{5.1}$$

Indeed, to estimate a quantile function we need an estimator of the distribution function.

We recall two classical estimators. Traditionally, the estimator of the distribution function is the empirical function $F_n(.)$, which is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_i),$$

[1] This chapter is a paper appeared in Journal of Siberian Federal University Mathematics and Physics 15:4 (2022), 510–522.Joint work with ALMI Nassima, SAYAH Abdallah.

where the indicator function $\mathbb{1}_{]-\infty,x]}(X_i) = 1$ if $X_i \leq x$ and 0 otherwise.
Theoretical properties of $F_n(.)$ have been investigated by several authors, (see, e.g. [61], [40] and [16]). It is well known that $F_n(.)$ is less smoothing, this fact leads to the effort to obtain a smooth estimate. Rosenblatt [??], Parsen [35] and Nadaraya [??] introduced the kernel estimators of $f(.)$ and $F(.)$ at x by:

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

and

$$\tilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

respectively, where $h = h_n$ is the smoothing parameter (or the bandwidth) since it controls the amount of smoothness in the estimator, and satisfy $h := h_n \longrightarrow 0$ as $n \longrightarrow \infty$, $k(.)$ is a kernel function which is a predetermined density function symmetric about 0 and the function K is defined from a kernel k as

$$K(x) = \int_{-\infty}^{x} k(t)dt.$$

When the support of the variable is bounded, the asymptotic properties are not satisfactory when the data is near the endpoints of the support, due to so-called boundary problem. To remove this boundary problem several methods have been proposed, (see, e.g. [27], [51], [53], [64], [2], [52]). As a result, the corresponding estimators of the quantile function have been proposed and studied extensively, in references can be found in the books of Galambos [18] and David [11].

A basic estimator of $Q(.)$ is the empirical quantile or the sample quantiles which is given by

$$Q_n(p) = \inf\{x : F_n(x) \geq p\} = X_{([np])},$$

where $[.]$ denotes the integer part.

The corresponding estimator of the quantile function $Q = F^{-1}$ is then defined by

$$\tilde{Q}_n(p) = \inf\{x : \tilde{F}_n(x) \geq p\}, \quad 0 < p < 1. \tag{5.2}$$

Nadaraya [32] showed under some assumptions for k, f and h, $\tilde{Q}_n(p)$ has an asymptotic standard normal distribution. The almost sure consistency was obtained by Yamato [61]. Ralescu and Sun [38] obtained the necessary and conditions for

the asymptotic normality. Azzalini [4] obtains the asymptotic mean squared error of $\tilde{Q}_n(p)$:

$$AMSE\left(\tilde{Q}_n(p)\right) = \frac{h^4}{4}\left(\frac{Q^{(2)}(p)}{\left(Q^{(1)}(p)\right)^2}\right)^2 \mu_2^2 + \frac{p(1-p)}{n}\left(Q^{(1)}(p)\right)^2 - \frac{h}{n}Q^{(1)}(p)\psi(k), \tag{5.3}$$

where $\psi(k) = 2\int_{-\infty}^{\infty} tk(t)K(t)dt$, $\mu_2 = \int_{-\infty}^{\infty} t^2k(t)dt$ and $Q^{(1)}$, $Q^{(2)}$ are the first, the second derivative of Q respectively.
It can be seen that the optimal bandwidth for minimizing (5.3) has the form

$$\tilde{h}_{opt} = \left(\frac{\left(Q^{(1)}(p)\right)^5 \psi(k)}{n\left(Q^{(2)}(p)\right)^2 \mu_2^2}\right)^{1/3}.$$

Parzen [35] proposed a version of the kernel quantile estimator as below:

$$\hat{Q}_n(p) = \sum_{i=1}^{n}\left[\int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{1}{h}k\left(\frac{x-p}{h}\right)dx\right]X_{(i)}. \tag{5.4}$$

In practice, the following approximation to $\hat{Q}_n(p)$ is often used:

$$\hat{Q}_n^a(p) = \frac{1}{nh}\sum_{i=1}^{n} X_{(i)}k\left(\frac{\frac{i}{n}-p}{h}\right). \tag{5.5}$$

Under suitable conditions on F, Falk [16] proposed the following kernel type quantile estimator

$$\check{Q}_n(p) = \frac{1}{h}\int_0^1 Q_n(x)k\left(\frac{x-p}{h}\right)dx. \tag{5.6}$$

This kernel-type quantile estimate can then be approximated by $\hat{Q}_n(p)$.
Yang [62] provided the asymptotic normality property and the mean squared consistency of $\hat{Q}_n(p)$ and proved that $\hat{Q}_n(p)$ and $\hat{Q}_n^a(p)$ are asymptotically equivalent in terms of mean square errors. Falk [16] showed that the asymptotic performance of $\hat{Q}_n(p)$ is better than that of the empirical sample quantile $Q_n(p)$ in terms of relative deficiency for appropriately chosen kernels and sufficiently smooth distribution functions. Building on Falk [16], Sheater and al [49] gave the asymptotic mean

squared error of $\hat{Q}_n(p)$.
If the second derivative of Q is continuous in a neighborhood of p and if f is not symmetric or f is symmetric but $p \neq \frac{1}{2}$ then asymptotic mean squared error of $\hat{Q}_n(p)$ is

$$AMSE\left(\hat{Q}_n(p)\right) = \frac{p(1-p)}{n}\left(Q^{(1)}(p)\right)^2 + \frac{h^4}{4}\left(Q^{(2)}(p)\right)^2\mu_2^2 - \frac{h}{n}\left(Q^{(1)}(p)\right)^2\psi(k). \tag{5.7}$$

The optimal bandwidth for $AMSE\left(\hat{Q}_n(p)\right)$ is

$$\hat{h}_{opt} = \left(\frac{\left(Q^{(1)}(p)\right)^2\psi(k)}{n\left(Q^{(2)}(p)\right)^2\mu_2^2}\right)^{1/3}. \tag{5.8}$$

When F is symmetric and $p = 1/2$ then the asymptotic mean squared error of $\hat{Q}_n(p)$

$$AMSE\left(\hat{Q}_n(p)\right) = n^{-1}\left(Q^{(1)}(1/2)\right)^2\left[0.25 - 0.5h\psi(k) + (nh)^{-1}\rho(k)\right],$$

where $\rho(k) = \int_{-\infty}^{\infty} k^2(x)dx$.
But all these estimators have a large bias when p is close to the boundary. In order to correct the bias problems in the case of extreme quantiles, Harrell et al [21] and Park [34] suggest using an asymmetric kernel, namely the Beta-type kernel. In particular, in the case of heavy-tailed distributions and for the same aim, Charpentier et al [9] suggested several nonparametric quantile estimators based on the beta-kernel and applied them to transformed data. Sayah et al [43] propose a new approach based on the modified Champernowne distribution. The main objective of this paper is to propose a new estimator to improve the asymptotic problems of extreme quantiles.

The paper organised as follows: In section 5.2, we propose our estimator and drive its asymptotic properties. In section 5.3, a simulation study was conducted where we compare the performance of our proposed estimator with both the empirical and the classical quantile estimators at specific values of p. In section 5.4, we compare graphically the mentioned estimators by using two real data applications. The paper is finalized with a brief conclusion.

5.2 Main results

According to the work of Almi et al [2], our estimator is based on a self-elimination between the Bias $\hat{Q}_n(p)$ of the estimator from itself

$$\bar{Q}_n(p) = \hat{Q}_n(p) - \hat{Bias}\Big(\hat{Q}_n(p)\Big), \tag{5.9}$$

then the explicit form of our estimator is given by

$$\bar{Q}(p) = \frac{1}{h}\sum_{i=1}^{n}\left(\int_{\frac{i-1}{n}}^{\frac{i}{n}}\left(k\Big(\frac{x-p}{h}\Big) - \frac{1}{2}\mu_2 k^{(2)}\Big(\frac{x-p}{h}\Big)\right)dx\right)X_{(i)}.$$

The following theorem shows that the Bias of $\bar{Q}_n(p)$ is of order $O(h^4)$, while that of $\hat{Q}_n(p)$ is $O(h^2)$, and it gives the expressions for the bias and the variance of the proposed estimator

▶ **Theorem 5.1.** Assume that Q has four bounded, continuous derivatives in a neighborhood of p and the kernel function k is a continuous bounded density, symmetric about zero, then if $0 < h \longrightarrow 0, nh^4 \longrightarrow \infty$, for all fixed $p \in]0,1[$, we have

$$Bias\big(\bar{Q}_n(p)\big) = \frac{h^4}{24}Q^{(4)}(p)\big(\mu_4 - 6\mu_2^2\big) + o\big(h^4\big),$$

and

$$Var\big(\bar{Q}_n(p)\big) = \frac{p(1-p)}{n}\Big(Q^{(1)}(p)\Big)^2 - \frac{h}{n}\Big(Q^{(1)}(p)\Big)^2\Psi(k) + o\left(\frac{h}{n}\right) + o(1),$$

where

$$\Psi(k) = \psi(k) - \frac{1}{4}\mu_2^2\int_{-\infty}^{+\infty}\Big(k^{(1)}(t)\Big)^2dt - \mu_2\left(\int_{-\infty}^{+\infty}tk(t)\left(\int_{-\infty}^{t}k^{(2)}(t)dt\right)dt + \int_{-\infty}^{+\infty}tk^{(2)}(t)\left(\int_{-\infty}^{t}k(t)dt\right)dt\right),$$

$$\mu_4 = \int_{-\infty}^{\infty}t^4k(t)dt.$$ ◀

Proof. As a result, our proposed estimator is

$$\bar{Q}_n(p) \quad = \quad \hat{Q}_n(p) - \tfrac{1}{2}h^2\hat{Q}_n^{(2)}(p)\mu_2.$$

We can easily see that

$$E\left(\bar{Q}_n(p)\right) = E\left(\hat{Q}_n(p)\right) - \frac{1}{2}h^2\mu_2 E\left(\hat{Q}_n^{(2)}(p)\right),$$

by a Taylor expansion, we have

$$\begin{aligned} E\left(\hat{Q}_n(p)\right) &= \frac{1}{h}\int_0^1 Q(x)k\left(\frac{p-x}{h}\right)dx \\ &= \int_{-\infty}^{\infty} k(t)(Q(p-ht))dt \\ &= Q(p) + \tfrac{1}{2}h^2 Q^{(2)}(p)\mu_2 + \frac{h^4}{24}Q^{(4)}(p)\mu_4 + o\left(h^4\right). \end{aligned}$$

Moreover

$$\begin{aligned} E\left(\hat{Q}_n^{(2)}(p)\right) &= \frac{1}{h^3}\int_0^1 Q(x)k^{(2)}\left(\frac{p-x}{h}\right)dx \\ &= \frac{1}{h}\int_0^1 Q^{(2)}(x)k\left(\frac{p-x}{h}\right)dx \\ &= \int_{-\infty}^{\infty} k(t)Q^{(2)}(p-ht)dt \\ &= Q^{(2)}(p) + \frac{1}{2}h^2 Q^{(4)}(p)\mu_2 + o\left(h^2\right). \end{aligned}$$

Thus, we have

$$Bias\left(\bar{Q}_n(p)\right) = \frac{h^4}{24}Q^{(4)}(p)\left(\mu_4 - 6\mu_2^2\right) + o\left(h^4\right).$$

On the other hand

$$\begin{aligned} Var\left(\bar{Q}_n(p)\right) &= Var\left(\hat{Q}_n(p) - \frac{1}{2}h^2\hat{Q}_n^{(2)}(p)\mu_2\right) \\ &= Var\left(\hat{Q}_n(p)\right) + \frac{1}{4}h^4\mu_2^2 Var\left(\hat{Q}_n^{(2)}(p)\right) - h^2\mu_2 Cov\left(\hat{Q}_n(p), \hat{Q}_n^{(2)}(p)\right), \end{aligned}$$

the variance of $\hat{Q}_n(p)$ can be computed as

$$\begin{aligned} Var\left(\hat{Q}_n(p)\right) &= \frac{1}{n}\left(Q^{(1)}(p)\right)^2\left(-p^2+2\int_{-\infty}^{\infty}(p-ht)k(t)K(t)dt\right)+o\left(\frac{h}{n}\right) \\ &= \frac{p(1-p)}{n}\left(Q^{(1)}(p)\right)^2-\frac{h}{n}\left(Q^{(1)}(p)\right)^2\psi(k)+o\left(\frac{h}{n}\right) \end{aligned}$$

and

$$Var\left(\hat{Q}_n^{(2)}(p)\right)=\frac{1}{nh^3}\left(Q^{(1)}(p)\right)^2\int_{-\infty}^{\infty}\left(k^{(1)}(t)\right)^2dt+o\left(\frac{1}{nh^3}\right)+o(1).$$

Now we will calculate the third term on the right hand side of $Var\left(\bar{Q}_n(p)\right)$.
We have

$$\begin{aligned} Cov\left(\hat{Q}_n(p),\hat{Q}_n^{(2)}(p)\right) &= E\left(\frac{1}{h}\left(\int_0^1 Q_n(x)k\left(\frac{p-x}{h}\right)dx-\int_0^1 Q(x)k\left(\frac{p-x}{h}\right)dx\right)\right. \\ &\quad \left.\frac{1}{h^3}\left(\int_0^1 Q_n(x)k^{(2)}\left(\frac{p-x}{h}\right)dx-\int_0^1 Q(x)k^{(2)}\left(\frac{p-x}{h}\right)dx\right)\right) \\ &= \frac{1}{h^2}E\left(\left(\int_{-\infty}^{\infty}k(t)\left((p-ht)-\bar{F}_n(p-ht)\right)Q^{(1)}(p-ht)dt\right)\right. \\ &\quad \left.\left(\int_{-\infty}^{\infty}k^{(2)}(t)\left((p-ht)-\bar{F}_n(p-ht)\right)Q^{(1)}(p-ht)dt\right)\right), \end{aligned}$$

where $\bar{F}_n$ is the empirical distribution function according to n independent, uniformly on $[0,1]$ distributed random variables.
Furthermore

$$\begin{aligned} Cov\left(\hat{Q}_n(p),\hat{Q}_n^{(2)}(p)\right) &= \frac{1}{nh^2}\int_0^1\left(\int_{-\infty}^{\infty}k(t)\left((p-ht)-1_{]0,p-ht[}(y)\right)Q^{(1)}(p-ht)dt\right) \\ &\quad \left(\int_{-\infty}^{\infty}k^{(2)}(t)\left((p-ht)-1_{]0,p-ht[}(y)\right)Q^{(1)}(p-ht)dt\right)dy, \end{aligned}$$

$$\begin{aligned} Cov\Big(\hat{Q}_n(p), \hat{Q}_n^{(2)}(p)\Big) &= \frac{1}{nh^2}\Bigg(\int_0^1\Bigg(\int_{-\infty}^{\infty} k(t)((p-ht) - 1_{]0,p-ht[}(y)Q^{(1)}(p)dt\Bigg) \\ &\quad \Bigg(\int_{-\infty}^{\infty} k^{(2)}(t)\big((p-ht) - 1_{]0,p-ht[}(y)\big)Q^{(1)}(p)dt\Bigg)dy + o(h)\Bigg) \\ &= \frac{\Big(Q^{(1)}(p)\Big)^2}{nh^2}\int_0^1\Bigg(\int_{\frac{p-1}{h}}^{\frac{p-y}{h}} k(t)dt \int_{\frac{p-1}{h}}^{\frac{p-y}{h}} k^{(2)}(t)dt\Bigg)dy + o\Big(\frac{1}{nh}\Big) + o(1). \end{aligned}$$

By integration by part we find

$$\begin{aligned} Cov\Big(\hat{Q}_n(p), \hat{Q}_n^{(2)}(p)\Big) &= \frac{\Big(Q^{(1)}(p)\Big)^2}{nh^3}\int_0^1 y\Bigg(k\Big(\frac{p-y}{h}\Big)\int_{\frac{p-1}{h}}^{\frac{p-y}{h}} k^{(2)}(t)dt \\ &+ k^{(2)}\Big(\frac{p-y}{h}\Big)\int_{\frac{p-1}{h}}^{\frac{p-y}{h}} k(t)dt\Bigg)dy + o\Big(\frac{1}{nh}\Big) + o(1) \\ &= \frac{\Big(Q^{(1)}(p)\Big)^2}{nh^2}\Bigg(\int_{\frac{p-1}{h}}^{\frac{p}{h}} (p-ht)k(t)\Bigg(\int_{\frac{p-1}{h}}^{t} k^{(2)}(t)dt\Bigg)dt \\ &+ \int_{\frac{p-1}{h}}^{\frac{p}{h}} (p-ht)k^{(2)}(t)\Bigg(\int_{\frac{p-1}{h}}^{t} k(t)dt\Bigg)dt\Bigg) + o\Big(\frac{1}{nh}\Big) + o(1) \\ &= \frac{-1}{nh}\Big(Q^{(1)}(p)\Big)^2\Bigg(\int_{-\infty}^{\infty} tk(t)\Bigg(\int_{-\infty}^{t} k^{(2)}(t)dt\Bigg)dt \\ &+ \int_{-\infty}^{\infty} tk^{(2)}(t)\Bigg(\int_{-\infty}^{t} k(t)dt\Bigg)dt\Bigg) + o\Big(\frac{1}{nh}\Big) + o(1). \end{aligned}$$

By adding up all these terms we have the desired result for the variance of $\bar{Q}_n$. ■

► **Corollary 5.2.** Suppose that the conditions of previous theorem5.1 hold. The asymptotic mean squared error of $\bar{Q}_n(p)$ is given by

$$AMSE\big(\bar{Q}_n(p)\big) = \Big(\frac{h^4}{24}Q^{(4)}(p)\big(\mu_4 - 6\mu_2^2\big)\Big)^2 + \frac{p(1-p)}{n}\Big(Q'(p)\Big)^2 - \frac{h}{n}\Big(Q'(p)\Big)^2\Psi(k).$$

It can be seen that the optimal bandwidth for minimizing $AMSE\big(\bar{Q}_n(p)\big)$ is both of order $O\Big(n^{-1/7}\Big)$ and has the form

$$\bar{h}_{opt} = \left(\frac{72\left(Q^{'}(p)\right)^{2}\Psi(k)}{n\left(Q^{(4)}(p)\left(\mu_4 - 6\mu_2^2\right)\right)^{2}} \right)^{1/7},$$

and the associated asymptotic mean squared error is given by

$$AMSE_{opt}\left(\bar{Q}_n(p)\right) = \frac{p(1-p)}{n}\left(Q^{(1)}(p)\right)^{2} - 7\left(\frac{\left(\frac{1}{8}\left(Q^{(1)}(p)\right)^{2}\Psi(k)\right)^{8}}{\left(\left(\frac{1}{24}Q^{(4)}(p)\left(\mu_4 - 6\mu_2^2\right)\right)^{2}\right)} \right)^{1/7} n^{-8/7}.$$

◀

5.3 Simulation study

In this section, we report results of a Monte Carlo study which was conducted to compare the performance of our proposed estimator $\bar{Q}_n(p)$ with the classical $\hat{Q}_n(p)$ and the empirical quantile estimators $\tilde{Q}_n(p)$, by computing the Bias and Mse for specific values of p where $p \in \{0.025, 0.05, 0.10, 0.20, 0.40, 0.60, 0.80, 0.90, 0.95, 0.975\}$. It is well known that bandwidth plays a critical role in the kernel estimation, for this reason we use the optimal bandwidth for $AMSE\left(\hat{Q}_n(p)\right)$ on each p-values, by using triweight kernel $\frac{35}{32}(1-t^2)^3\mathbb{1}_{|t|\leq 1}$. In order to account for different cases, we generate a thousand samples of two sizes $n = 50$ and $n = 200$ from different distributions listed in the Table 1, results of the comparison are shown in Tables 2 to 11.
where ϕ^{-1} denote the Inverse of standard normal distribution.

After examining all tables, the classical estimator $\hat{Q}_n$ does not perform as well at near boundary points p =0.025 to 0.10 and p =0.90 to 0.975 as at interior points from $p = 0.20$ to $p = 0.80$, However, it can be observed that our proposed estimator $\bar{Q}_n$ produces lower Bias(MSE) for almost values of p specifically extreme values in all distributions considered, except for Weibull distribution in the case where p=0.05 the performance of the classical estimator is better than our estimator for the small size. Both estimators are more efficient than the empirical quantile estimator $\tilde{Q}_n$ in most situations.

Table 5.1: Distributions used in the simulation study

Distribution	Theorical quantille $Q(p)$
Weibull $(\frac{3}{2}, 1)$	$(-\log(1-p))^{\frac{2}{3}}$
Log-normal $(0, \frac{1}{2})$	$exp(\frac{1}{2}\phi^{-1}(p))$
Chi-square (1)	$(\phi^{-1}(\frac{p+1}{2}))^2$
Log-logistic $(1, 3)$	$\frac{1}{3}\left(\frac{p}{1-p}\right)$
Pareto $(3, 1)$	$\left(\frac{1}{1-p}\right)^{\frac{1}{2}} - 1$

Table 5.2: Bias(MSE) values for Weibull distribution, n=50, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	921.8(851.8)	67.49(8.130)	33.35(4.673)
0.050	872.0(760.4)	29.16(8.812)	32.21(10.97)
0.100	787.2(619.4)	17.64(15.67)	16.25(14.80)
0.200	642.2(412.4)	43.34(34.93)	26.63(24.63)
0.400	371.0(137.7)	22.15(20.78)	15.45(12.27)
0.600	297.0(267.9)	19.31(12.89)	9.298(3.918)
0.800	264.6(218.4)	19.11(6.300)	5.810(5.128)
0.900	307.7(125.4)	27.62(11.86)	12.41(11.31)
0.950	547.1(544.6)	96.32(47.93)	33.57(20.53)
0.975	261.9(159.2)	105.3(102.7)	98.05(65.27)

Table 5.3: Bias(MSE) values for Weibull distribution, n=200, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	91.84(37.23)	15.85(6.271)	8.210(3.764)
0.050	86.69(75.16)	18.75(6.188)	17.36(5.389)
0.100	78.19(61.14)	6.156(8.972)	4.025(7.439)
0.200	63.71(40.59)	5.627(8.628)	1.929(6.379)
0.400	36.60(33.39)	1.103(19.49)	0.137(17.56)
0.600	21.03(34.17)	0.955(31.91)	0.128(30.42)
0.800	63.66(405.3)	1.263(59.20)	0.134(57.26)
0.900	27.23(80.20)	2.177(10.92)	0.165(10.80)
0.950	62.74(60.91)	3.333(20.22)	0.318(20.04)
0.975	147.9(102.8)	87.32(78.69)	29.48(34.66)

Table 5.4: Bias (MSE) values for Log-normal distribution, n=50, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	692.5(874.8)	251.9(222.9)	224.6(183.1)
0.050	656.6(849.7)	167.9(205.2)	120.5(56.68)
0.100	478.1(496.2)	152.4(171.9)	98.47(22.09)
0.200	368.6(257.6)	148.6(249.4)	159.5(253.2)
0.400	124.0(254.2)	88.39(83.43)	75.51(68.43)
0.600	98.36(96.20)	48.36(47.50)	38.51(32.16)
0.800	99.57(95.47)	79.02(73.13)	65.15(60.21)
0.900	354.8(135.2)	281.6(156.2)	102.9(98.25)
0.950	97.22(48.21)	47.00(45.21)	31.42(27.52)
0.975	106.8(231.2)	94.27(194.2)	56.22(152.7)

Table 5.5: Bias(MSE) values for Log-normal distribution, n=200, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	525.8(973.5)	209.2(214.4)	146.1(143.0)
0.050	365.6(511.6)	119.1(147.2)	101.3(137.7)
0.100	278.1(437.0)	109.9(204.2)	83.91(129.3)
0.200	368.6(188.5)	27.7(28.46)	5.168(19.84)
0.400	124.0(254.2)	88.39(83.48)	68.43(75.71)
0.600	87.50(255.1)	8.891(31.66)	4.475(14.99)
0.800	486.8(410.2)	12.91(9.126)	9.848(7.638)
0.900	264.8(301.4)	18.98(35.20)	7.667(18.07)
0.950	686.7(261.6)	305.2(184.1)	76.70(156.6)
0.975	437.0(545.1)	568.0(315.5)	387.9(231.7)

Table 5.6: Bias (MSE) values for Chi-square distribution, n=50, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	57.08(41.67)	28.72 (25.07)	24.32 (20.59)
0.050	20.10(18.72)	1 3.46 (12.74)	8. 029 (6.447)
0.100	44.23(38.56)	26.27(24.28)	14.082(10.35)
0.200	94.08(97.85)	56.06(53.25)	19.084 (17.83)
0.400	88.94(84.51)	37.12(32.35)	32.35(20.47)
0.600	80.89(72.16)	29.18 (27.32)	11.00(10.31)
0.800	85.51(75.92)	30.83(27.46)	14.68(12.10)
0.900	96.79(95.75)	41.97(37.65)	7.484(5.908)
0.950	195.6(193.8)	185.6(176.1)	158.6(155.7)
0.975	196.4(195.5)	185.4(134.4)	153.0(144.2)

Table 5.7: Bias (MSE) values for Chi-square distribution, n=200, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	257.0(246.7)	150.2(122.4)	144.7(132.4)
0.050	18.07(15.84)	5.725(3.278)	0.487(0.222)
0.100	28.92(19.78)	16.43(12.70)	6.843(4.683)
0.200	9.408(8.851)	4.544(2.065)	2.201 (2.035)
0.400	13.29(13.23)	8.518(7.256)	6.235(5.524)
0.600	29.66(28.80)	9.882(9.504)	1.182(1.087)
0.800	40.78(36.63)	16.47(12.71)	4.830(2.333)
0.900	45.02(44.27)	20.38(19.23)	4.845(2.201)
0.950	46.00(43.11)	21.94(14.61)	3.487(2.516)
0.975	136.30(131.80)	115.31(112.40)	101.3(98.45)

Table 5.8: Bias (MSE) values for Log-logistic distribution, n=50, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	849.7(7.222)	229.6(5.246)	83.33(0.694)
0.050	987.4(9.751)	205.2(4.214)	56.68(0.321)
0.100	967.9(9.370)	171.9(2.95)	22.09(0.048)
0.200	188.8(35.66)	162.1(2.624)	19.84(0.039)
0.400	490.2(43.03)	284.6(8.102)	75.90(0.576)
0.600	487.4(57.24)	384.1(46.80)	221.6(4.914)
0.800	479.1(105.0)	279.3(77.91)	80.53(64.85)
0.900	462.4(234.6)	444.3(171.6)	130.9(127.7)
0.950	429.1(489.2)	198.1(392.7)	152.7(233.2)
0.975	107.2(1214)	92.13(848.8)	99.90(458.7)

Table 5.9: Bias (MSE) values for Log-logistic distribution, n=200, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	85.547(0.7305)	110,20(1.214)	38.083(0.145
0.050	334.21(111.69)	84.898(0.720)	14.621(0.021)
0.100	967,98(937.00)	55.155(0.304)	4.789(0.002)
0.200	921.69(849.51)	49.941(0.494)	8.084(0.006)
0.400	782.80(612.78)	80.428(0.646)	11.409(0.013
0.600	505.02(255.02)	182,61(3.334)	49.413(0.244
0.800	640.53(410.28)	569.59(32.44)	122.44(1.499
0.900	545.72(297.87)	404.6(99.81)	276.37(7.638)
0.950	3549.0(301.4)	2411.4(304,7)	990.34(98.07)
0.975	511.55(626.85)	409.28(475,1)	109.844(120.6)

Table 5.10: Bias (MSE) values for Pareto distribution, n=50, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	946.5(975.1)	257.7(421.4)	150.7(52.13
0.050	950.9(937.1)	261.0(654.2)	79.53(258.7)
0.100	45.83(583.7)	44.56(331.5)	39.39(250. 02)
0.200	48.95(96.10)	26.61(85.10)	12.73(83.8)
0.400	48.66(93.20)	39.63(93.00)	29.14(85.04)
0.600	48.01(158.2)	9.677(51.20)	4.853 (23. 55)
0.800	78.00(278.4)	26.99(258.4)	2.474(20.01)
0.900	457.7(673.5)	32.46(140.6)	19.51(76.42)
0.950	2439(1954)	308.8(531.1)	278.4(445.9
0.975	3135(3298)	848.8(1104)	87.25(300.1)

Table 5.11: Bias (MSE) values for Pareto distribution, n=200, Results are re-scaled by the factor 0.0001.

p	$\tilde{Q}$	$\hat{Q}$	$\bar{Q}$
0.025	503.7(763.5)	232.2(404.5)	105.0(43.85)
0.050	201.4(528.6)	159.3(401.7)	34.69(36.87)
0.100	9.509(280.6)	9.489(400.1)	9.109(36.76)
0.200	8.869(86.30)	1.693(39.96)	0.163(35.75)
0.400	10.71(58.61)	3.885(39.99)	2.917(36.72)
0.600	4.238(58.61)	1.421(40.11)	1.302(36.65)
0.800	72.47(256.0)	34.20(41.08)	5.521(36.73
0.900	472.3(301.7)	272.6(145.8)	96.73(101.2)
0.950	2037(1582)	1051(780.0)	803.2(386.6)
0.975	671.9(330.0	347.8(459.5)	187.5(221.4)

5.4 Application

In this section, we compare the performance of our proposed estimator with the empirical and the classical estimators by using the graphical representation of two real data sets. The first data set consist of 100 observations of breaking stress of carbon fibers (in Gba) given by Nichols and Padget [33] and the second data set consist of 63 observations related to the strength of carbon fibers tested under tension at gauge lengths of 10 mm, The data has been recently reported and analyzed by Bi and Gui [5] among others, the choice of the bandwidth bases to cross-validation method. The results are shown in Figures 1 and 2 respectively. It's remarkably clear that our newly proposed estimator is closer to the unknown quantile function as compared to both estimators the classical and the empirical kernel estimators, this yields that our estimator improves the performance of the classical estimator in extreme quantiles.

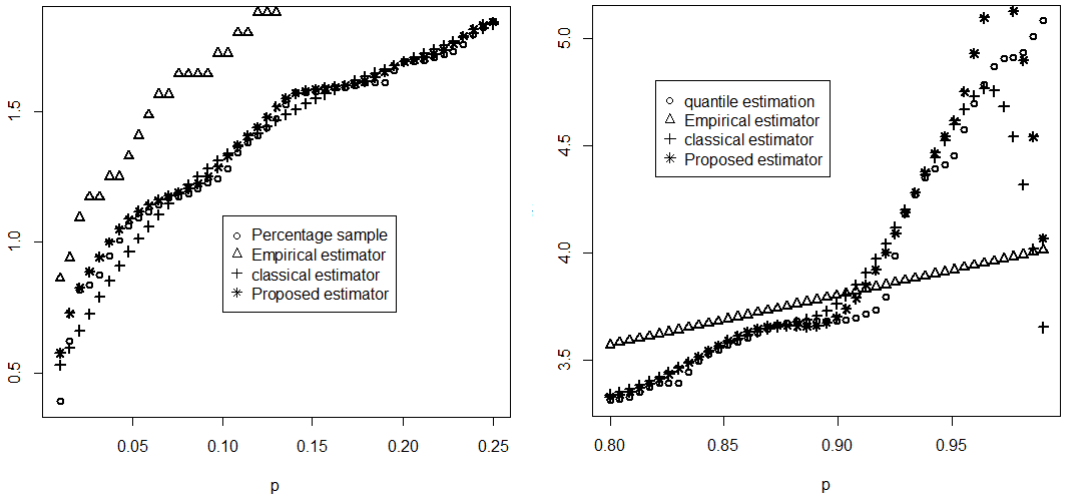

Figure 5.1: Performance of different estimators in real applications

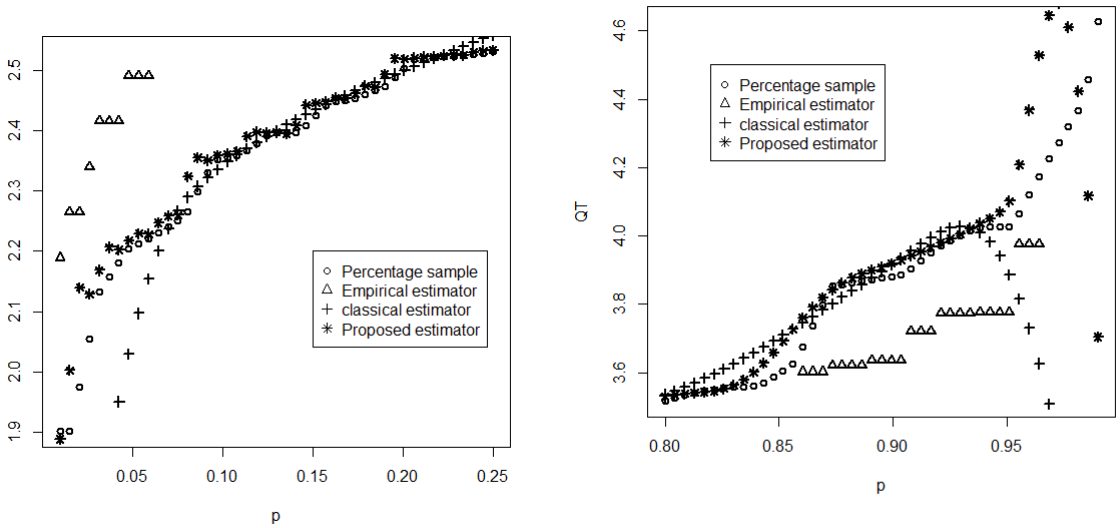

Figure 5.2: Performance of different estimators in real applications

5.5 Conclusion

This paper proposes a smooth estimator of the quantile function to improve the efficiency of the classical kernel estimator at extreme quantiles. Depending on the theory it turned out that the Bias of our proposed estimator is of fourth-order power of the bandwidth, while that of the classical is second-order. The numerical results are summarized in Tables 2 to 6 and Figures 1 and 2 conducted that our proposed estimator is better than both the classical and the empirical quantile estimators in the meaning of Bias and Mse for almost all p-values and specifically at extremes. These numerical results coincide with the theoretical results in Theorem (5.1).

6 Conclusions & Outlook

Kernel estimation methods are not well implemented when the data is near the boundary of the compressed support, even if we choose the appropriate bandwidth which we call the boundary effect. Several authors considered this problem in kernel density and regression estimates. Whereas, in the kenel distribution estimation and the inverse distribution estimation (quantile function) are relatively few, though these functions have found in numerous applications in econometrics, climatology, and hydraulics, among others.
In this thesis, we are interested to improve the performance of the classical estimators of both functions the distribution function Nadaraya [32] and the inverse distribution function Parzen [35] in the case when the data near the right boundary, we applied a new method based on self-elimination between the Bias and the estimator itself.

- Depending on the theoretical results it turned out that our proposed estimator reduces the order of bias from $o(h^2)$ to $o(h^4)$, while the variance remains at the same order as the existing estimators.
- the numerical results it is shown that the MISE of the proposed estimators is smaller than that of the used estimators in all distributions and for each sample size. Note that the reduction of the MISE is mainly due to the bias, and the variance parts for all estimators are very close.
- As a result we reveal the superior performance of the proposed estimator.

Bibliography

[1] Al-Kenani. A and Keming.Yu,New bandwidth selection for kernel quantile estimators, Journal of Probability and Statistics, (2012).

[2] Almi.N, Sayah.A, nonparametric kernel distribution function estimation near endpoints, Advances in Mathematics: Scientific Journal, 10(2021),3679-3697.

[3] Altman. N, and C. Leger, Bandwidth selection for kernel distribution function estimation, Journal of Statistical Planning and Inference(1995),46, 195–214.

[4] Azzalini. A, A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method, Biometrika, (1981) 68, 326–328.

[5] Bi. Q, Gui. W, Bayesian and classical estimation of stress-strength reliability for inverse Weibull lifetime models,Algorithms,(2017) 10,71.

[6] Bordji. H, Sayah. A, Bias correction at endpoints in kernel density estimation Doctoral dissertation, Université de mohamed kheider biskra(2021).

[7] Bowman. A, Hall. A, and T. Prvan, Bandwidth selection for the smoothing of distribution functions, Biometrika (1998) 85: 799–808.

[8] Buch-Larsen. T, Nielsen.J.P, Guillen. M and Bolancé, C, Kernel density estimation for heavy-tailed distributions using the Champernowne transformation, Statistics. (2005)(6)39, 503-518.

[9] Charpentier.A, A.Oulidi, Beta kernel quantile estimators of heavy-tailed loss distributions, Stat. Comput, 20(2010),35-55.

[10] Cowling.A et al On Pseudodata Methods for Removing Boundary Effects in Kernel Density Estimation, Journal of the Royal Statistical Society: Series B, 58 (1996), 551–563.

[11] David.H.A. , Order Statistics, 2nd Edition, New York: John Wiley.

[12] Denuit.M, J. Dhaene , Actuarial Theory for Dependent Risk: Measures, Orders and Models, Wiley, New York. (2005)

[13] Devroye. L. and Gyorfi, L, Nonparametric density estimation, The L1 View. New York: Wiley. (1985).

[14] Eddy.F, Optimum kernel estimators of the mode,The Annals of Statistics,(1980), 8870–882.

[15] Epanechnikov.A, Non-parametric estimation of a multivariate probability density, Theory of Probability & Its Applications, 114 (1969),153–158.

[16] Falk.M, Relative efficiency and deficiency of kernel type estimators of smooth distribution functions,Statist. Neerlandica, 37(1983),73-83.

[17] Fernholz. L Almost sure convergence of smoothed empirical distribution functions, Scandinavian Journal of Statistics, (1991),18, 255-262.

[18] Galambos.j, The asymptotic theory of extreme order statistics, Krieger,Malabar,Florida.(1978)

[19] Gasser, T, Muller, H. G, Kernel estimation of regression functions, Springer Berlin Heidelberg, 1979.

[20] Gasser. T and al: Kernels for Nonparametric Curve Estimation, Journal of the Royal Statistical Society. Series B (Methodological), 58 (1985), 238–252.

[21] Harrell. F.E, C. E.Davis, A New Distribution-Free Quantile Estimator, Biometrika, 69(1982),635-640.

[22] Horova.I, Kolacek. J., Zelinka, J., and El-Shaarawi, A. H. Smooth estimates of distribution functions with application in environmental studies, Advanced topics on mathematical biology and ecology,(2008), 122-127.

[23] Jones, M. C. The performance of kernel density functions in kernel distribution function estimation, Statistics and Probability Letters,(1990), 9 :129-132.

[24] Karunamuni.R J and Alberts.I T: A Generalized Reflection Method of Boundary Correction in Kernel Density Estimation, Canadian Journal of Statistics, 33 (2005), 497–509.

[25] Karunamuni.R J, zhang.S, Some Improvements on a Boundary Corrected Kernel Density Estimator, Statistics and Probability Letters, 78 (2008), 497–507.

[26] Koláček .J and Karunamuni.RJ., On Boundary Correction in Kernel Estimation of ROC Curves, Austrian Journal of Statistics, Statistics and Probability Letters, 38 (2009), 17–32.

[27] J.Koláček and RJ. Karunamuni, A Generalized Reflection Method for Kernel Distribution and Hazard Functions Estimation, Journal of Applied Probability and Statistics, 6 (2011), 73–85.

[28] Lall.U and Y.Moon and K.Bosworth, Kernel flood frequency estimators: bandwidth selection and kernel choice, Water Resources Research, 4 29,(1993),1003-1015.

[29] Lehmann. E. L, Theory of Point Estimation, Wadsworth and Brooks/Cole, Belmont, (1991).

[30] Linhartand. H, Zucchini. W: Model Selection, John Wiley and Sons, 9 (1986).

[31] López-de-Ullibarri.I, Bandwidth selection in kernel distribution function estimation,The Stata Journal, (2015)15,3,784–795.

[32] Nadaraya.E A, Some New Estimates for Distribution Functions, Theory of Probability and Its Applications, 9 (1964), 497–500.

[33] Nichols. MD, Padgett.WJ, A bootstrap control chart for Weibull percentiles, Qual Reliab Eng Int, 22 (2006),141-151.

[34] Park. C, Smooth nonparametric estimation of a quantile function under right censoring using beta kernels, Technical Report (TR 2006-01-CP), Department of Mathematical Sciences, Clemson University, (2006).

[35] Parzen . E, On Estimation of a Probability Density Function and Mode, The annals of mathematical statistics, 33 (1962), 1065–1076.

[36] Polanski.A, Baker.ER, Plug-in Bandwidth Selection for Kernel Distribution Function Estimates, Journal of Statistical Computation and Simulation,((2000)) 65, 63-80.

[37] Quintela.D and al, Nonparametric Kernel Distribution Function Estimation with kerdiest: an R Package For Bandwidth Choice and Applications, Journal of Statistical Software, 50 (2012), 1–21.

[38] Ralescu. S.S, Sun.S, Necessary and sufficient conditions for the asymptotic normality of perturbed sample quantiles, J. Statist. Plann. Inference, 35 (1993),55-64.

[39] Read.R.R The asymptotic inadmissibility of the sample distribution function. Ann. Math. Statist, 43 (1972), 89-95.

[40] Reiss. R D, Nonparametric Estimation of Smooth Distribution Functions, Scandinavian Journal of Statistics, 8 (1981), 116–119.

[41] Rio.del.AQ, Compariason of Bandwidth Selectors in Nonparametric Regression under Dependence, Computational Statistics Data Analysis,(1996). 21, 563-580.

[42] Rosenblatt. M, Remarks on Some Nonparametric Estimates of a Density Function, The Annals of Mathematical Statistics, 27 (1956), 832–837.

[43] Sayah. A, Y. Djebrane, A. Necir, Champernowne transformation in kernel quantile estimation for heavy-tailed distributions, Afrika Statistika, 5(2010),288-296.

[44] Sarda. P, Smoothing Parameter Selection for Smooth Distribution Function, Journal of Statistical Planning and Inference,(1993)35, 65-75.

[45] Shankar.B, An optimal choice of bandwidth for perturbed sample quantiles, master thesis (1998).

[46] Silverman.WR, Density Estimation for Statistics and Data Analysis, Chapman and Hall,London (1986).

[47] Singh.R. S, Gasser, T. and Prasad, B. Nonparametric estimates of distributions functions, Communication in Statistics Theory and Methods, (1983),12:2095-2108.

[48] Shirahata. S and Chu I. S, Integrated squared error of kernel-type estimator of distribution function, Annals of the Institute of Statistical Mathematics,(1992), 44: 579-591.

[49] S. J.Sheater, J. S. Marron, Kernel quantile estimtors, Journal of the American Statistical Associatio, 85 (1990),410-416.

[50] Swanepoel. J.W. H, Mean integrated squared error properties and optimal kernel when estimating a distribution function, Comm. Statist. Theory Methods,(1988),17,3785-3799.

[51] Tenreiro. C, Boundary Kernels for Distribution Function Estimation, REVSTAT Statistical Journal, 11 (2013), 169–190.

[52] Tenreiro. C, A note on boundary kernels for distribution function estimation, arXiv preprint arXiv:1501.04206, (2015).

[53] Tour. M and Sayah .A and Djebrane.Y: A Modified Champernowne Transformation to Improve Boundary Effect in Kernel Distribution Estimation, Afrika Statistika, 12 (2017), 1219–1233.

[54] Tsybakov. A,B Introduction à l'estimation non paramétrique,Springer Science & Business Media (41), (2003),

[55] Wand. M. P, Marron, J. S. and Ruppert, D, Transformations in density estimation, J. Amer. Statist. Assoc.(1991) 86 (414): 343-361, .

[56] Wand, Matt P and Jones, M Chris, Kernel smoothing, CRC press(1994).

[57] Watson. GS and Leadbetter. MR, Hazard Analysis II, Sankhyā: The Indian Journal of Statistics, Series A, 26 (1964), 101–116.

[58] Winter. B. B, Strong uniform consistency of integrals of density estimators, Canad.J.Statist,(1973), 1, 247-253.

[59] Winter.BB, Convergence Rate of Perturbed Empirical Distribution Functions, Journal of Applied Probability, 16 (1979), 163–173.

[60] Watson. G. S and Leadbetter. M. R, Hazard analysis II, Sankhyd Set. A, (1964),26, 101-116.

[61] Yamato. H, Uniform Convergence of an Estimator of a Distribution Function, Bulletin of Mathematical Statistics, 15 (1973), 69–78.

[62] Yang. S. S, A Smooth Nonparametric Estimator of a Quantile Function, Journal of the American Statistical Association, 80 (1985),1004-1011.

[63] Zhang. S, Karunamuni, R.J. and Jones, M.C., An improved estimator of the density function at the boundary, Journal of the American Statistical Association, (1999)94,12-31,

[64] Zhang. S, Zhong. L and Zhang. Z Estimating a Distribution Function at the Boundary, Austrian Journal of Statistics, 49 (2020), 1–23.

Appendix A: Abbreviations and Notations

The different abbreviations and symbols used throughout this thesis are explained below:

rv	*:random variable*
X	*:rv defined on* $(\Omega, \mathcal{A}, \mathcal{P})$, *population*
$(X_1, ..., X_n)$	*:samples of size n from X*
$(X_{1,n}, ...X_{n,n})$	*:order statistics pertaining to* $(X_1, ..., X_n)$
$X_{i,n}$	*:ith order statistics* $(i = 1, n)$
$E[X]$	*:expectation of (or mean of X)*
$Var(X)$	:variance of (X)
pdf	*probability density function*
df	*:distribution function*
f	*:pdf of X*
F	*:cumulative distribution function of X*
F_n	*:empirical df*
F^{-1}	*:gneralized inverse of F, quantile function*
$\mathbb{1}_A$	:indicator function of set A
[X]	:integer part of a real number
$o(.)$	$:f(x) = o(g(x))\ as\ x \to x_0 : f(x)/g(x) \to 0\ as\ x \to x_0$
$O(.)$	$:f(x) = O(g(x))\ as\ x \to x_0 : \exists M > 0, \lvert f(x)/g(x) \rvert \leq M as\ x \to x_0$
MSE	: Mean squared error
$MISE$	: Mean integrated squared error
AMISE	: Asymptotic Mean integrated squared error
iid	:independent identically distributed
inf A	:infinimum of set A

Appendix B: Useful R commands

The different function *used throughout this thesis are explained below:*

ecdf :	Compute an empirical cumulative distribution function.
kcdf :	Compute the nonparametric kernel estimate for cumulative distribution function
quantile :	Compute the sample quantiles corresponding to the given probabilities.
npquantile :	Computes smooth quantiles from a univariate unconditional kernel cumulative distribution estimate
kerdiest :	kerdiest-package
kde :	Kernel distribution function estimatorker
ALbw :	Plug-in bandwidth selection of Altman and Leger.
PBbw :	Plug-in bandwidth selection of Polansky and Baker.
CVbw :	Cross-validation bandwidth selection of Bowman, Hall and Prvan.
floor :	Returns the largest integers not greater than the corresponding elements.
sd :	Compute the standard deviation of the values in x

List of Publications and Communication

Articles in Refereed Journals

- *Nonparametric Kernel Distribution Function Estimation Near Endpoints*. *Advances in Mathematics: Scientific Journal 12:10 (2021), 3679–3697. Joint work with ALMI Nassima, SAYAH Abdallah.*

- *Estimating the Inverse Distribution Function at the Boundary*. *Journal of Siberian Federal University Mathematics and Physics 15:4 (2022), 510–522. Joint work with ALMI Nassima, SAYAH Abdallah.*

- *On Kernel Distribution Function Estimation Near Endpoints*. *Journal of Applied and Engineering Mathematics. Joint work with ALMI Nassima, SAYAH Abdallah. (Promise to publish and accept letter).*

- *Improved the Bias in Kernel Quantile Function Estimation*. *AIMS Mathematics. Joint work with SAYAH Abdallah, ALMI Nassima. 8:1(2022),1784-1799.*

- *A Novel Bias Reduction Method for Kernel Quantile Function Estimation at the Boundary*. *Journal of Science and Arts. Joint work with SAYAH Abdallah, ALMI Nassima.*

List of Publications and Communication

Communications

- *International Conference on Advances in Applied Mathematics, 17-20 December,2018,Tunisian Association of Applied and Industrial Mathematics , Sousse (Tunisie):* *On Kernel Distribution Function Estimation Near Endpoints.*

- *Second National Mathematics Seminar at the Freres Mentouri University Constantine (Algeria) 2 June, 2021:* On Kernel Distribution Function Estimation Near Endpoints.

- *Third National Mathematics Seminar at the Freres Mentouri University Constantine (Algeria) 26 May, 2022:* Nonparametric Kernel Distribution Function Estimation Near Endpoints.

- Applied Mathematics Days Mohamed Khider University Biskra. Algeria, a communication participante (December 13, 2022); Estimating the Inverse Distribution Function at endpoint.