ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **10** (2021), no.12, 3679–3697
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.10.12.10

NONPARAMETRIC KERNEL DISTRIBUTION FUNCTION ESTIMATION NEAR ENDPOINTS

Nassima Almi and Abdallah Sayah[1]

ABSTRACT. 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.

1. INTRODUCTION

The cumulative distribution function F used to determine the probability that a random observation X that taken from 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

[1]*corresponding author*
2020 *Mathematics Subject Classification.* 62G05, 62G20.
Key words and phrases. Boundary effects, Bias reduction, Cumulative distribution function, Kernel estimator.
Submitted: 30.11.2021; *Accepted:* 16.12.2021; *Published:* 20.12.2021.

by Nadaraya [9] defined for $X_1, X_2, ..., X_n$ a sample of a continuous real random variable by:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}, \tag{1.1}$$

such an estimator arises as an integral of kernel density estimator f_n which is introduced by Rosenblatt [13] and Parzen [10] that has the form:

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}, \tag{1.2}$$

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

$$K(x) = \int_{-1}^{x} k(t)dt. \tag{1.3}$$

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 [18] and Yamato [19], the asymptotic normality of F_n is established by Watson and Leadbetter [17] and an asymptotic expression for the mean squared error of F_n and the asymptotically optimal smoothing parameter proved by Azzalini [1]. These properties are satisfactory, but when the support of the variable is bounded kernel estimation may suffer. It is well know 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 right boundary $]1 - h, 1]$.

The value of Bias and Variance of F_n at interior points provided by Azzalini [1] are respectively:

$$\frac{1}{2}f^{(1)}(x)\mu_2(k)h^2 + o(h^2), \tag{1.4}$$

and

$$\frac{F(x)(1 - F(x))}{n} + \frac{h}{n}f(x)\left(\int_{-1}^{1} K^2(t)dt - 1\right) + o\left(\frac{h}{n}\right), \tag{1.5}$$

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:

$$-hf(1)\int_{-1}^{-c} K(t)dt + h^2 f^{(1)}(1)\left(\frac{c^2}{2} - \int_{-1}^{c} tK(t)dt + c\int_{-1}^{-c} K(t)dt\right) + o(h^2), \tag{1.6}$$

and

$$\frac{F(x)(1-F(x))}{n} + \frac{h}{n}f(1)\left(-c - 2\int_{-1}^{-c} K(t)dt + \int_{-1}^{c} K^2(t)dt\right) + o\left(\frac{h}{n}\right). \tag{1.7}$$

In the results, we can see that for densities taking value zero at the endpoints of the support the first order term in (1.6) 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 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 [14], pseudo-data method [2] and also the boundary kernel method [3]. 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 et al [6] considered this problem in estimating ROC curves using the transformation method, Tour et al [16] used a Champernowne transformation for heavy tailed distributions in the left side of the support and Tenreiro [15] and Zhang et al [20] proposed a boundary kernel method free of boundary problem. 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.

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

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 (1.6) from F_n itself when the data near the right boundary of the support for $x = 1 - ch$ defined by

$$\breve{F}_n(x) = F_n(x) + h\Psi(c)f_n(x) + h^2\alpha f_n^{(1)}(x), \tag{2.1}$$

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

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

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

$$Bias(\breve{F}_n(x)) = h^2 f^{(1)}(1)\phi(c) + o(h^2), \tag{2.2}$$

$$Var(\breve{F}_n(x)) = \frac{F(x)(1 - F(x))}{n} + \frac{h}{n}f(1)\rho(c) + o\left(\frac{h}{n}\right), \tag{2.3}$$

where

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

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

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

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

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

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

We calculate each term separately

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

by using the remark 2.1, relation 1.3 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 relaion(1.6).

By the same procedure, we have

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

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

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

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

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

This completes the proof of expression (2.2).

On the other hand

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

where

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

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 (1.7).

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

$$
\begin{aligned}
J_{13} &= \frac{h^4\alpha^2}{n} E\left(\left(\frac{1}{h^2}k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right)^2 - \frac{1}{n}\left(E\left(\alpha h^2\left(\frac{1}{h^2}k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right)\right)^2 \\
&= \frac{h\alpha^2}{n} f(1)\int_{-c}^{1}\left(k^{(1)}(t)\right)^2 dt + o\left(\frac{h}{n}\right), \\
J_{14} &= \frac{2}{n}h\Psi(c)\left(E\left(\frac{1}{h}K\left(\frac{x-X_i}{h}\right)k\left(\frac{x-X_i}{h}\right)\right)\right. \\
&\quad \left. -E\left(\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right)E\left(k\left(\frac{x-X_i}{h}\right)\right)\right) \\
&= \frac{2}{n}h\Psi(c)f(1)\int_{-c}^{1}k(t)K(t)dt + o\left(\frac{h}{n}\right), \\
J_{15} &= \frac{2\alpha}{n}h^2\left(E\left(\frac{1}{h^2}K\left(\frac{x-X_i}{h}\right)k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right. \\
&\quad \left. - E\left(K\left(\frac{x-X_i}{h}\right)\right)E\left(\frac{1}{h^2}k^{(1)}\left(\frac{x-X_i}{h}\right)\right)\right) \\
&= \frac{2\alpha h f(1)}{n}\left(\int_{-c}^{1}k^{(1)}(t)K(t)\,dt\right) + o\left(\frac{h}{n}\right),
\end{aligned}
$$

and

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

This completes the proof of expression (2.3). □

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 [6] when the data near the left side of the support. Our proposed estimator $\hat{F}_n$ developed this technique when the data near

the right boundary of the support, given for $x \in]1-h,1]$ by

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-g(X_i)}{h}\right) + \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-2+g(X_i)}{h}\right), \tag{2.4}$$

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

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

Theorem 2.2. *Under the above assumptions A_1, $A2$, H_1 and H_2, the asymptotic properties of our proposed estimator $\hat{F}_n$ at $x = 1 - ch$ are*

$$Bias(\hat{F}_n(x)) = h^2\Gamma(c) + o(h^2), \tag{2.5}$$

and

$$Var(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n} + \frac{h}{n}f(1)\Omega(c) + o\left(\frac{h}{n}\right), \tag{2.6}$$

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

$$\Gamma(c) = \frac{-c^2}{2}f^{(1)}(1) + \left(f^{(1)}(1) - g^{(2)}(1)f(1)\right)$$

$$\cdot \left(-2c^2 + 2c \int_{-1}^{-c} K(t)dt - \int_{-c}^{c} K(t)(t+c)dt \right),$$

and

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

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

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

where

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

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

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

$$+ h \int_{-c}^{c} K(t) \frac{f(g^{-1}(x-th))}{g^{(1)}\left(g^{-1}(x-th)\right)} dt,$$

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

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

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

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

Therefore

$$\begin{aligned} F\left(g^{-1}\left(1-2ch\right)\right) = F(x) - chf(1) + \frac{3(ch)^2}{2} f^{(1)}(1) \\ - 2(ch)^2 \left(g^{(2)}(1) f(1)\right) + o(h^2). \end{aligned}$$

Eventually, we obtain

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

Similar computation give I_2,

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

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 $Bias(\hat{F}_n)$ (2.5).

To prove (2.6), note that

$$\begin{aligned} nVar(\hat{F}_n) &= E\left(K\left(\frac{x-g(X_i)}{h}\right) + K\left(\frac{x-2+g(X_i)}{h}\right)\right)^2 \\ &\quad - \left(E\left(K\left(\frac{x-g(X_i)}{h}\right) + K\left(\frac{x-2+g(X_i)}{h}\right)\right)\right)^2 \\ &= A_1 - A_2, \end{aligned}$$

where

$$\begin{aligned} A_1 &= E\left(K\left(\frac{x-g(X_i)}{h}\right) + K\left(\frac{x-2+g(X_i)}{h}\right)\right)^2, \\ &= A_{11} + A_{12} + 2A_{13}. \end{aligned}$$

It can be shown that

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

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 $Bias(\hat{F}_n)$, 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 (2.6). □

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 asymptotic properties of our estimators with the existing estimators summarized in the coming subsection. For each estimator, we evaluate the Bias and Mse at right boundary from different distributions with support $[0, 1]$ are listed in table 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 cross-validation method [11] to choose bandwidth for Epanechnikov kernel, the main reason for this choice is that it provides a fair basis for comparison among the different estimators without regards to bandwidth effects.

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 $\bar{F}_n$), inspered from the generalized reflection kernel distribution estimator (Karunamuni et al [7]), 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 [20]

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

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 Epanechnikov kernel we choice

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

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

TABLE 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$(2,2)_{\left[\frac{1}{3};1\right]}$	$4.05x(1-x)$
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

The simulation results measure the performance of the different estimators in the meaning of the Bias and Mse, are summarized in tables 1.2 and 1.3.

TABLE 2. Bias values at x=1, Results are re-scaled by the factor 0.001.

	F_n	$\bar{F}_n$	$\tilde{F}_n$	$\hat{F}_n$	$\breve{F}_n$		
				$\alpha =$	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

TABLE 3. Mse values at x=1, Results are re-scaled by the factor 0.001.

	F_n	$\bar{F}_n$	$\tilde{F}_n$	$\hat{F}_n$	$\breve{F}_n$		
				$\alpha =$	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

From Table 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 $\hat{F}_n$ 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 $\hat{F}_n$. The comparison of the modify Bias of kernel estimator $\breve{F}_n$ depend to 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, $\breve{F}_n$

improve the performance of the estimator. For the other estimators in generally, the boundary distribution kernel estimator $\tilde{F}_n$ has second smaller Bias followed by the reflection estimator $\bar{F}_n$. From Table 3, our proposed estimator $\hat{F}_n$ has an asymptotically smaller Mse when compared with the other estimators, which they organised 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. Real data application

The aim of our applications is to compare the performance of the two proposed kernel distribution estimators given respectively in (2.1) and (2.4) using the cross-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 [8] represents the failure 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. 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 (1), 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 $\breve{F}_n$ estimator improves when the positive constant α is large in this graph we chose α=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.

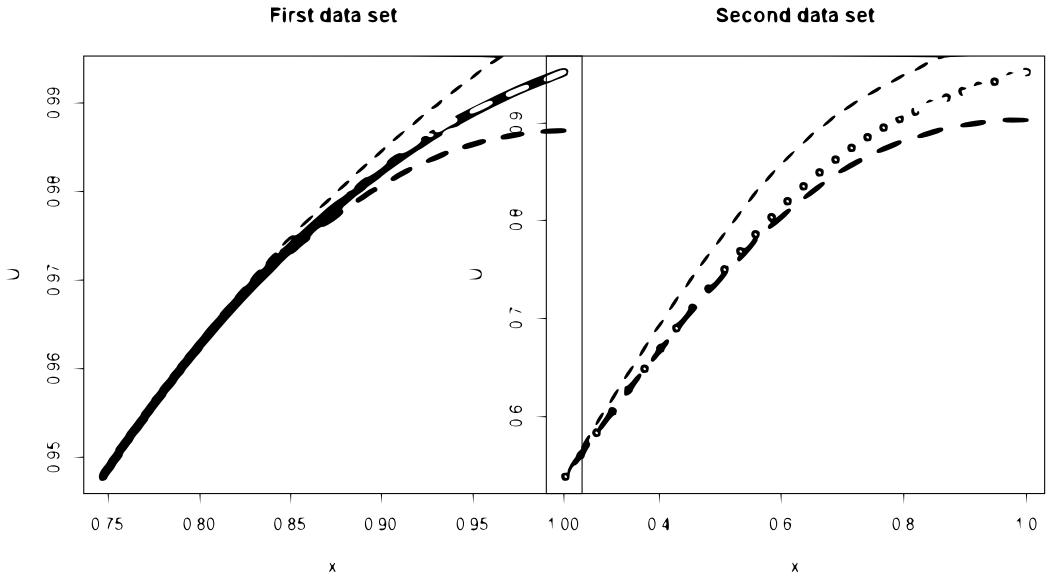

FIGURE 1. Performance of different estimators in real applications

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.

REFERENCES

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

[2] A. COWLING 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.

[3] T. GASSER ET AL: *Kernels for Nonparametric Curve Estimation*, Journal of the Royal Statistical Society. Series B (Methodological), **58** (1985), 238–252.

[4] R J. KARUNAMUNI AND I T. ALBERTS: *A Generalized Reflection Method of Boundary Correction in Kernel Density Estimation*, Canadian Journal of Statistics, **33** (2005), 497–509.

[5] R J. KARUNAMUNI AND S. ZHANG: *Some Improvements on a Boundary Corrected Kernel Density Estimator*, Statistics and Probability Letters, **78** (2008), 497–507.

[6] J.KOLÁČEK AND RJ. KARUNAMUNI: *On Boundary Correction in Kernel Estimation of ROC Curves, Austrian Journal of Statistics*, Statistics and Probability Letters, **38** (2009), 17–32.

[7] 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.

[8] H. LINHARTAND W. ZUCCHINI: *Model Selection*, John Wiley and Sons, **9** (1986).

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

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

[11] D. QUINTELA 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.

[12] R.D. REISS: *Nonparametric Estimation of Smooth Distribution Functions*, Scandinavian Journal of Statistics, **8** (1981), 116–119.

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

[14] WR. SILVERMAN: *Density Estimation for Statistics and Data Analysis*, Chapman and Hall,London (1986).

[15] C. TENREIRO: *Boundary Kernels for Distribution Function Estimation*, REVSTAT Statistical Journal, **11** (2013), 169–190.

[16] M. TOUR AND A. SAYAH AND Y. DJEBRANE: *A Modified Champernowne Transformation to Improve Boundary Effect in Kernel Distribution Estimation*, Afrika Statistika, **12** (2017), 1219–1233.

[17] GS. WATSON AND MR. LEADBETTER: *Hazard Analysis II*, Sankhyā: The Indian Journal of Statistics, Series A, **26** (1964), 101–116.

[18] BB. WINTER: *Convergence Rate of Perturbed Empirical Distribution Functions*, Journal of Applied Probability, **16** (1979), 163–173.

[19] H. YAMATO: *Uniform Convergence of an Estimator of a Distribution Function*, Bulletin of Mathematical Statistics, **15** (1973), 69–78.

[20] S. ZHANG AND L. ZHONG AND Z. ZHANG: *Estimating a Distribution Function at the Boundary*, Austrian Journal of Statistics, **49** (2020), 1–23.

DEPARTMENT OF MATHEMATICS.
UNIVERSITY OF MOHAMED KHIDER.
BISKRA.
ALGERIA.
Email address: almi.nassima@gmail.com

DEPARTMENT OF MATHEMATICS.
UNIVERSITY OF MOHAMED KHIDER.
BISKRA.
ALGERIA.
Email address: abdallah.sayah@univ-biskra.dz