Afrika Statistika

Afrika Statistika
Vol. 12(1), 2017, pages 1219–1233.
DOI: http://dx.doi.org/10.16929/as/2017.1219.101

ISSN 2316-090X

A modified Champernowne transformation to improve boundary effect in kernel distribution estimation

Madiha Tour, Abdallah Sayah and Djabrane Yahia

Laboratory of Applied Mathematics, Mohamed Khider University, Biskra, Algeria

Received : January 26, 2017; Accepted : June 6, 2017

Abstract. Kernel distribution estimators are not consistent near the boundary of its support. Several solutions to this problem have already been proposed. In this paper, we propose a new kernel estimation of the cumulative distribution function for heavy tailed distributions based on the method of the transformation of the data set with a modification of the Champernowne distribution and the generalized reflection method of boundary correction for kernel distribution estimation. The asymptotic bias, variance and mean squared error of the proposed estimator are given. Simulations are drawn to show that the proposed method perform quite well when compared with other existing methods.

Résumé. Les estimateurs à noyau de la fonction de distribution ne sont pas consistants aux bords de son support. Plusieurs solutions à ce problème ont déjà été proposées. Dans cet article nous proposons un nouveau estimateur à noyau de la fonction de distribution pour les distributions à queue lourde basé sur la méthode de la transformation de l'ensemble de donneés avec la distribution de Champernowne modifiée et la méthode de reflexion généralisée de correction de l'effet de bord pour l'estimation à noyau de la distribution. Le biais asymptotique, la variance et l'erreur quadratique moyenne de l'estimateur proposé sont donnés. Des simulations sont effectuées pour montrer que la méthode proposée se comporte assez bien par rapport à d'autres méthodes existantes.

Key words: Transformation; Boundary effect; Kernel distribution estimation; Heavy tailed distributions.
AMS 2010 Mathematics Subject Classification : 62G07; 62G20.

[†]Corresponding author Abdallah Sayah: sayahabdel@yahoo.fr
Madiha Tour : tourmadiha23@gmail.com
Djabrane Yahia : yahia_dj@yahoo.fr

1. Introduction

Let X be a real random variable (rv) with unknown continuous distribution function (cdf) F and density function f. Estimating the 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, risk quantification concentrates in the highest values of the domain of the distribution, where sample information is scarce and it is, therefore, necessary to extrapolate the behaviour of the cdf, even above the maximum observed. The most commonly used nonparametric estimate of a cdf is the empirical distribution function. It is known that the empirical distribution is an unbiased estimator of cdf. A nonparametric alternative for estimating the cdf is the kernel estimator. This is more efficient than the empirical distribution but it is, nevertheless, a biased estimator. Furthermore, both the empirical distribution and the kernel estimator of the cdf are not consistent near the boundary of its support. Although there is a vast literature on boundary correction in density estimation context, boundary effects problem in distribution function context has been less studied, in particularly for heavy tailed distributions. In this paper, we develop a new kernel type estimator of the cdf that removes boundary effects near the end points of the support.
Kernel smoothing has received a lot of attention in density estimation contex (see, e.g., Silverman, 1986, Wand and Jones, 1995). Specifically, let $X_1, ..., X_n$ be a sample of size $n \geq 1$ from the rv X. The popular nonparametric kernel estimator of f which is introduced by Rosenblatt (1956) and Parzen (1962) and has the form

$$\widehat{f}_n(x) = \frac{1}{nb} \sum_{i=1}^{n} k\left(\frac{x - X_i}{b}\right),$$

where $b := b_n$ is the bandwidth or the smoothing parameter ($b \longrightarrow 0$, as $n \longrightarrow \infty$) and k is a nonnegative symmetric kernel function such that it is bounded and has finite support. The kernel distribution function estimator $\widehat{F}_n(x)$ was proposed by Nadaraya (1964). Such an estimator arises as an integral of the Parzen-Rosenblatt kernel density estimator (see Reiss, 1981 and Tenreiro, 2013) and is defined for $x \in \mathbb{R}$, by

$$\widehat{F}_n(x) = \int_{-\infty}^{x} \widehat{f}_n(t)dt = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right), \tag{1}$$

where

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

is the integrated kernel. However, several properties of $\widehat{F}_n(x)$ have been investigated, Azzalini (1981) have derived an asymptotic expression for the mean squared error of $\widehat{F}_n(x)$, and determined also the asymptotically optimal smoothing parameter. Winter (1979) and Yamato (1973) proved the uniform convergence of $\widehat{F}_n(x)$ to $F(x)$ with probability one, the asymptotic normality of $\widehat{F}_n(x)$ is established by Watson and Leadbetter (1964).

The problems of boundary effect for kernel estimators with compact supports is well-known in regression and density function estimation and several modified estimators have been proposed in the literature (see Gasser and Müller, 1979, Karunamuni and Alberts, 2005, Zhang *et al.*, 1999, and references therein). A similar correction would be made for improve the

theoretical performance of the usual kernel distribution function estimator (1), at boundary. More specifically the performance of $\widehat{F}_n(x)$ at boundary points, for $x \in [0,b] \cup (a-b,a]$, $0 < a \leq \infty$, however differs from the interior points due to so-called "boundary effects" that occur in nonparametric curve estimation problems. The bias of $\widehat{F}_n(x)$ is of order $o(b)$ instead of $o(b^2)$ at boundary, while the variance of $\widehat{F}_n(x)$ is of order $o(\frac{b}{n})$. This fact can be clearly seen by examining the behavior of $\widehat{F}_n$ inside the left boundary region $[0,b]$. Let x be a point in the left boundary region, $x \in [0,b]$. The bias and variance of $\widehat{F}_n(x)$ at $x = sb$, $0 \leq s \leq 1$ are

$$Bias\left(\widehat{F}_n(x)\right) = bf(0)\int_{-1}^{-s} K(t)dt + b^2 f'(0)\left\{\frac{s^2}{2} + s\int_{-1}^{-s} K(t)dt - \int_{-1}^{s} tK(t)dt\right\} + o\left(b^2\right), \tag{2}$$

and

$$Var\left(\widehat{F}_n(x)\right) = \frac{F(x)(1-F(x))}{n} + \frac{b}{n}f(0)\left\{\int_{-1}^{s} K^2(t)dt - s\right\} + o\left(\frac{b}{n}\right). \tag{3}$$

To remove those boundary effects in kernel distribution estimator, a variety of methods have been developed in the literature. We briefly mention reflection of data (see, e.g., Silverman, 1986), transform of data (see, Marron and Ruppert (1994)), pseudo-data method (see Cowling and Hall, 1996) and also the boundary kernel method (Gasser et al., 1985 , Zhang and Karunamuni, 2000). For more details about this techniques one refers to Karunamuni and Alberts (2005), Karunamuni and Alberts (2005).

In this paper, we develop a new kernel type estimator for heavy tailed distributions functions that improved boundary effects near the points at left boundary region, i.e., for $x \in [0,b]$. This estimator is based on a new transformation on boundary corrected kernel estimator ideas of Koláček and Karunamuni (2009), Buch-Larsen *et al.* (2005), developed for boundary correction in kernel density estimation. The basic technique of construction of the proposed estimator is a kind of a generalized reflection method involving reflecting a transformation of the observed data, using two transformations. First, a transformation g which is selected from a parametric family, is applied to the data. Next, a transformation T is used. Specifically, our transformation T is based the little-known Champernowne distribution function, produces good results in all the studied situations and is easy to implement.

Theoretical properties of boundary kernel distribution estimator are introduced in Section 2. In Section 3 the proposed estimator is given and its bias and variance are computed. In Section 4, simulation studies are done to see the performance of the proposed estimator, and compare it with the "usual" and "boundary" distribution function estimators. The proofs are postponed in Section 5.

2. Boundary kernel distribution estimator

In order to deal with the boundary effects that occur in nonparametric regression and density function estimation, the use of boundary kernels is proposed and studied by authors such as Gasser and Müller (1979), Karunamuni and Alberts (2005). Next we extend this approach

to a distribution function estimator framework. This method of estimating combines the transformation and the reflection methods, consisting of three steps:

Step 1. Transform the initial data $X_1, ..., X_n$ to $g(X_1), ..., g(X_n)$, where g is a nonnegative, continuous, and monotonically increasing function from $[0, \infty)$ to $[0, \infty)$.

Step 2. Reflect $g(X_1), ..., g(X_n)$ around the origin, so we get $-g(X_1), ..., -g(X_n)$.

Step 3. The estimator of F is based on the enlarged data sample $-g(X_1), ..., -g(X_n)$, $g(X_1), ..., g(X_n)$. Then the boundary kernel distribution estimator of the distribution function for $x \in [0, b]$, is given by

$$\overline{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{K\left(\frac{x - g(X_i)}{b}\right) - K\left(-\frac{x + g(X_i)}{b}\right)\right\}, \tag{4}$$

where K is a distribution of the kernel function k as in (1). This estimator generates a class of boundary corrected estimators.

The bias and the variance of $\overline{F}_n$ are given by (Koláček and Karunamuni, 2009, page 22) under some assumptions on f and g. For $x = sb$, $0 \leq s \leq 1$, we have

$$\begin{aligned} Bias\left(\overline{F}_n(x)\right) = b^2 \Bigg\{ & f'(0)\left(\frac{s^2}{2} + 2s\int_{-1}^{-s} K(t)dt - \int_{-s}^{s} tK(t)dt\right) \\ & - f(0)\, g''(0)\left(\int_{-1}^{s}(s-t)K(t)dt + \int_{-1}^{-s}(s+t)K(t)dt\right)\Bigg\} + o\left(b^2\right), \end{aligned} \tag{5}$$

and

$$\begin{aligned} Var\left(\overline{F}_n(x)\right) = & \frac{F(x)\left(1 - F(x)\right)}{n} + \frac{b}{n} f(0)\left\{2\int_{-1}^{-s} K^2(t)dt - s\right. \\ & \left. + \int_{-s}^{s} K^2(t)dt - 2\int_{-1}^{s} K(t)K(t-2s)dt\right\} + o\left(\frac{b}{n}\right). \end{aligned} \tag{6}$$

Accordingly, the asymptotic mean squared error ($AMSE$) is

$$\begin{aligned} AMSE\left(\overline{F}_n(x)\right) = b^4 \Bigg\{ & f'(0)\left(\frac{s^2}{2} + 2s\int_{-1}^{-s} K(t)dt - \int_{-s}^{s} tK(t)dt\right) \\ & - f(0)\, g''(0)\left(\int_{-1}^{s}(s-t)K(t)dt + \int_{-1}^{-s}(s+t)K(t)dt\right)\Bigg\}^2 \\ & + \frac{F(x)\left(1 - F(x)\right)}{n} + \frac{b}{n} f(0)\left\{2\int_{-1}^{-s} K^2(t)dt - s\right. \\ & \left. + \int_{-s}^{s} K^2(t)dt - 2\int_{-1}^{s} K(t)K(t-2s)dt\right\}. \end{aligned} \tag{7}$$

Remark 1. Some discussion on the above choice of g and other various improvements that can be made would be appropriate here. It is possible to construct functions g that improve the bias under some additional conditions. For instance, if one examines the right

hand side of bias expansion, then it is not difficult to see that the coefficient of b^2 can be made equal to zero if g is appropriately chosen, (see Koláček and Karunamuni, 2009). Furthermore, functions satisfying conditions $g^{-1}(0) = 1$ and $g'(0) = 0$ are easy to construct. The trivial choice is $g(y) = y$, which represents the "classical" reflection method estimator. The following transformation adapts well to various shapes of distributions:

$$g(y) = y + \frac{1}{2} I_s y^2,$$

for $y \geq 0$ and $0 \leq s \leq 1$, where $I_s = \int_{-1}^{-s} K(t)\, dt$.

Remark 2. It is easy to see that for $x > b$, the estimator (4) reduces to (1) , which is the usual kernel distribution estimator. So (4) is a natural boundary continuation of the usual estimator.

3. The proposed estimator

We now have all the necessary tools to introduce our estimator for heavy tailed cdf F, based on ideas of Koláček and Karunamuni (2009), Buch-Larsen *et al.* (2005) and we insert a new transformation. We shall assume that the unknown cdf F has support $[0, \infty)$. The transformation idea is based on transforming the original data by a new parametric transformation T, chosen by modified Champernowne distribution function. The modified Champernowne distribution is defined on $x \geq 0$, and formulated as

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

with parameter $\alpha > 0$, $M > 0$ and $c \geq 0$, and its density is

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

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

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

For more details about the modified Champernowne distribution see for instance Buch-Larsen *et al.* (2005), Champernowne (1952).
The following steps describes the techniques using for obtain the proposed estimator of F.

Step 1. Estimate the parameters $(\widehat{\alpha}, \widehat{M}, \widehat{c})$ of the modified Champernowne distribution to obtain the transformation function. In the modified Champernowne distribution, we notice that $T_{\alpha,M,0}(M) = 0.5$. This suggests that M can be estimated as the empirical median of the data set Lehmann (1991). Then to estimate the pair (α, c) which maximizes the log likelihood function :

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

Step 2. Transform the initial data $X_1, ..., X_n$, with the transformation function,

$$Y_i = T(X_i), \quad i = 1, ..., n,$$

Y is a rv uniformly distributed in the interval $(0, 1)$.

Step 3. Calculate the boundary kernel distribution estimator of the transformed data, $Y_1, ..., Y_n$:

$$\widetilde{H}_n(y) = \frac{1}{n}\sum_{i=1}^{n}\left\{K\left(\frac{y - g(Y_i)}{b}\right) - K\left(-\frac{y + g(Y_i)}{b}\right)\right\}, \tag{8}$$

where g is the same transformation as in (4).

Step 4. The final form of our estimator of the original data set, $X_1, ..., X_n$ is defined as, for $x = sb$, $0 \leq s \leq 1$,

$$\widetilde{F}_n(x) := \widetilde{H}_n(T(x)). \tag{9}$$

Remark 3. The estimator $\widetilde{F}_n(x)$ is a natural boundary continuation of the usual kernel distribution estimator (1). Furthermore, it is important to remark here that the transform kernel distribution estimator (9) is nonnegative (provided K is nonnegative).

The next theorem establishes the bias and variance of the proposed estimator (9) .

Theorem 1. *Assume that $f'(.)$ and $g''(.)$ exist and are continuous. Further assume that $g^{-1}(0) = 1$ and $g'(0) = 0$, where g^{-1} is the inverse function of g, f' and g'' are the first and the second derivatives of f and g respectively. Then for $x = sb$, $0 \leq s \leq 1$ the bias and variance of $\widetilde{F}_n(x)$ are respectively*

$$\begin{aligned} Bias\left(\widetilde{F}_n(x)\right) = b^2 \Bigg\{ & \left(\frac{f}{T'}\right)'(0)\frac{1}{T'(0)}\left(\frac{s^2}{2} + 2s\int_{-1}^{-s} K(t)dt - \int_{-s}^{s} tK(t)dt\right) \\ & - \frac{f(0)}{T'(0)}g''(0)\left(\int_{-1}^{s}(s-t)K(t)dt + \int_{-1}^{-s}(s+t)K(t)dt\right)\Bigg\} + o\left(b^2\right), \end{aligned} \tag{10}$$

and

$$\begin{aligned} Var\left(\widetilde{F}_n(x)\right) = & \frac{F(x)(1 - F(x))}{n} + \frac{b}{n}\frac{f(0)}{T'(0)}\left\{2\int_{-1}^{-s} K^2(t)dt - s\right. \\ & \left. + \int_{-s}^{s} K^2(t)dt - 2\int_{-1}^{s} K(t)K(t-2s)dt\right\} + o\left(\frac{b}{n}\right). \end{aligned} \tag{11}$$

The asymptotic mean squared error is

$$\begin{aligned} AMSE\left(\widetilde{F}_n(x)\right) = b^4 \Bigg\{ & \left(\frac{f}{T'}\right)'(0)\frac{1}{T'(x)}\left(\frac{s^2}{2} + 2s\int_{-1}^{-s} K(t)dt - \int_{-s}^{s} tK(t)dt\right) \\ & - \frac{f(0)}{T'(0)}g''(0)\left(\int_{-1}^{s}(s-t)K(t)dt + \int_{-1}^{-s}(s+t)K(t)dt\right)\Bigg\}^2 \\ & + \frac{F(x)(1 - F(x))}{n} + \frac{b}{n}\frac{f(0)}{T'(0)}\left\{2\int_{-1}^{-s} K^2(t)dt - s\right. \\ & \left. + \int_{-s}^{s} K^2(t)dt - 2\int_{-1}^{s} K(t)K(t-2s)dt\right\}. \end{aligned} \tag{12}$$

Remark 4. By comparing expressions (2), (3), (10), and (11) we can see that both bias and variance are of the same order at boundary points. So the proposed estimator improved boundary effects in kernel distribution estimator since the bias at boundary points is of the same order as the bias at the interior points.

4. Simulation Studies

To compare the performance of our proposed estimator $\widetilde{F}_n$ against the boundary kernel estimator $\overline{F}_n$ and the usual $\widehat{F}_n$ estimator described by Nadaraya (1964), we made some simulation studies. We simulate data from three different heavy tailed distributions : Pareto type I, Pareto type II and Pareto type III. The distributions and the chosen parameters are listed in table 1.

Distribution	$F(x)$ for $x \geq 0$	Parameters
Pareto type I	$1-(1+x/\sigma)^{-\alpha}$	$(\sigma,\, \alpha) = (1, 1)$
Pareto type II	$1-\left(1+\frac{x}{\sigma}\right)^{-\alpha}$	$(\sigma,\, \alpha) = (1, 2)$
Pareto type III	$1-\left(1+\left(\frac{x}{\sigma}\right)^{\frac{1}{\gamma}}\right)^{-1}$	$(\sigma, \gamma) = (0.7, 1)$

Table 1. Distributions used in the simulation studies.

We measure the performance of the estimators by the error measures $AMSE$ and $AMISE$. The simulation is based on 1000 replications. In each replication the sample sizes: $n = 50$, $n = 200$ and $n = 400$ was used. For the kernel, we choosing the Epanechnikov kernel $k(t) = 3/4(1-t^2)I(|t| \leq 1)$, where $I(.)$ denotes the indicator function, has been observed in Silverman (1986), that this kernel possesses the maximum efficiency, in the sense that it produces the minimal $AMISE$. The choice of bandwidth is very important for the good performance of any kernel estimator. In all cases, we select the asymptotic optimal global bandwidth of the estimator $\overline{F}_n$ by minimizing the $AMISE$, because this is much more likely to be used in application and gave reliably good results. We have

$$b_{opt} = \left(\frac{2f\,(0)\,A(s)}{5\left[f'(0)B(s) - f\,(0)\,g''\,(0)\,C(s)\right]^2}\right)^{1/3} n^{-1/3},$$

where

$$A(s) := \left(2\int_{-1}^{s} K\,(t)\,K\,(t-2s)\,dt + s - 2\int_{-1}^{-s} K^2\,(t)\,dt - \int_{-s}^{s} K^2\,(t)\,dt\right),\ 0 \leq s \leq 1,$$

$$B(s) := \left(\frac{s^2}{2} + 2s\int_{-1}^{-s} K(t)dt - \int_{-s}^{s} tK(t)dt\right),\ 0 \leq s \leq 1,$$

and

$$C(s) := \left(\int_{-1}^{s} (s-t)K(t)dt + \int_{-1}^{-s} (s+t)K(t)dt\right),\quad 0 \leq s \leq 1.$$

The comparison is based on data simulated from the three distributions described in table 1. Firstly, for each value of $s \in \{0.35, 0.45, 0.55\}$ we have calculated the absolute bias, variance

and the $AMSE$ values of the three estimators and have displayed the results in tables 2, 3 and 4. Secondly, for different values of s we calculated the $AMISE$ values for each estimator over the whole boundary region $[0, b]$. The values of $AMISE$ are tabulated in table 5. The comparison show that the values of the $AMSE$ and the $AMISE$ were smallest in case of the proposed estimator, this is due to the fact that the proposed estimator is locally adaptive.

			Pareto type I			**Pareto type II**			**Pareto type III**		
	Est	s	.35	.45	.55	.35	.45	.55	.35	.45	.55
	$\widetilde{F}_n$		6.3049	6.3733	6.6385	14.830	16.289	17.605	4.1028	4.3000	4.5112
$\lvert Bias\rvert$	$\overline{F}_n$		26.904	28.668	31.311	28.840	31.802	34.524	26.783	28.780	31.482
	$\widehat{F}_n$		38.987	48.741	54.245	38.659	45.524	53.794	39.962	48.336	48.841
	$\widetilde{F}_n$		0.1042	0.2428	0.1299	0.2403	0.2759	0.1587	0.1183	0.3229	0.2383
Var	$\overline{F}_n$		0.3714	0.5461	0.4918	0.4391	0.5249	0.4619	0.3802	0.5384	0.4985
	$\widehat{F}_n$		0.9980	1.1071	1.3112	1.0359	1.3299	1.5280	0.9695	1.1331	1.3767
	$\widetilde{F}_n$		0.1439	0.2834	0.1740	0.4602	0.5412	0.4686	0.1352	0.3414	0.2587
$AMSE$	$\overline{F}_n$		1.0952	1.3679	1.4722	1.2709	1.5363	1.6539	1.0976	1.3667	1.4896
	$\widehat{F}_n$		2.5180	3.4828	4.2537	2.5305	3.4023	4.4219	2.5665	3.4695	3.7622

Table 2. Bias, Var and AMSE Values Over the Boundary Region for sample size n=50. Results are re-scaled by the factor 0.001.

			Pareto type I			**Pareto type II**			**Pareto type III**		
	Est	s	.35	.45	.55	.35	.45	.55	.35	.45	.55
	$\widetilde{F}_n$		2.0016	2.0856	2.1298	5.0417	5.3146	5.8175	0.8145	0.8062	0.8240
$\lvert Bias\rvert$	$\overline{F}_n$		10.489	11.443	12.596	11.811	12.531	13.786	10.623	11.250	12.468
	$\widehat{F}_n$		13.168	18.879	24.637	13.276	19.884	32.944	15.899	21.445	23.283
	$\widetilde{F}_n$		0.0611	0.0929	0.1029	0.0533	0.0879	0.0935	0.0593	0.1022	0.1247
Var	$\overline{F}_n$		0.0931	0.1309	0.1490	0.1025	0.1452	0.1647	0.0961	0.1344	0.1489
	$\widehat{F}_n$		0.1719	0.2346	0.2498	0.2329	0.2664	0.2778	0.1932	0.2092	0.2640
	$\widetilde{F}_n$		0.0651	0.0972	0.1075	0.0787	0.1161	0.1273	0.0600	0.1028	0.1253
$AMSE$	$\overline{F}_n$		0.2031	0.2618	0.3077	0.2420	0.3022	0.3547	0.2089	0.2609	0.3043
	$\widehat{F}_n$		0.3453	0.5910	0.8568	0.4091	0.6617	1.3631	0.4459	0.6691	0.8061

Table 3. Bias, Var and AMSE Values Over the Boundary Region for sample size n=200. Results are re-scaled by the factor 0.001.

4.1. Discussion and conclusion

For Pareto type I distribution, close examination of tables of $AMSE$ clear by shows that, the proposed estimator $\widetilde{F}_n$ and the boundary kernel distribution estimator $\overline{F}_n$ show the best performance, but the estimator $\widetilde{F}_n$ out performs the estimator $\overline{F}_n$ for all n. Also, in terms of

			Pareto type I			**Pareto type II**			**Pareto type III**		
	Est	s	.35	.45	.55	.35	.45	.55	.35	.45	.55
	$\widetilde{F}_n$		0.9937	1.0213	1.0449	2.5560	2.7109	2.9718	0.7311	0.7523	0.7888
$\|Bias\|$	$\overline{F}_n$		6.5212	7.2401	7.9143	7.2751	7.7788	8.5954	6.6332	7.1435	7.9339
	$\widehat{F}_n$		8.7913	10.2731	16.296	11.057	17.428	20.437	7.4034	12.070	14.831
	$\widetilde{F}_n$		0.0336	0.0450	0.0629	0.0255	0.0401	0.0407	0.0281	0.0422	0.0357
Var	$\overline{F}_n$		0.0424	0.0559	0.0759	0.0512	0.0705	0.0790	0.0448	0.0630	0.0644
	$\widehat{F}_n$		0.0719	0.1025	0.1113	0.0926	0.1066	0.1420	0.0768	0.0956	0.1193
	$\widetilde{F}_n$		0.0345	0.0461	0.0640	0.0320	0.0474	0.0496	0.0286	0.0428	0.0363
$AMSE$	$\overline{F}_n$		0.0849	0.1084	0.1386	0.1042	0.1310	0.1529	0.0888	0.1141	0.1274
	$\widehat{F}_n$		0.1492	0.2081	0.3769	0.2149	0.4103	0.5597	0.1316	0.2413	0.3393

Table 4. Bais, Var and AMSE Values Over the Boundary Region for sample size n=400. Results are re-scaled by the factor 0.001.

			Pareto type I			**Pareto type II**			**Pareto type III**		
	Est	s	.35	.45	.55	.35	.45	.55	.35	.45	.55
	$\widetilde{F}_n$		0.2015	0.1062	0.1014	0.2353	0.1148	0.0317	0.3175	0.2504	0.1805
$n = 50$	$\overline{F}_n$		0.5882	0.5437	0.4142	0.4024	0.3262	0.2826	0.3998	0.3456	0.2947
	$\widehat{F}_n$		1.3379	1.2503	1.3200	0.7429	0.7232	0.6875	0.8923	0.7979	0.8473
	$\widetilde{F}_n$		0.0405	0.0241	0.0101	0.0262	0.0178	0.0065	0.0450	0.0396	0.0337
$n = 200$	$\overline{F}_n$		0.0748	0.0657	0.0619	0.0523	0.0468	0.0432	0.0495	0.0444	0.0398
	$\widehat{F}_n$		0.2028	0.1927	0.1799	0.1281	0.1045	0.0995	0.1542	0.1249	0.1216
	$\widetilde{F}_n$		0.0160	0.0103	0.0072	0.0088	0.0036	0.0078	0.0157	0.0133	0.0108
$n = 400$	$\overline{F}_n$		0.0258	0.0225	0.0229	0.0195	0.0157	0.0135	0.0177	0.0156	0.0133
	$\widehat{F}_n$		0.0806	0.0729	0.0639	0.0512	0.0374	0.0383	0.0509	0.0469	0.0453

Table 5. AIMSE Values Over the Boundary Region. Results are re-scaled by the factor 0.001.

$AMISE$ for each sample size, the $AMISE$ of the estimator $\widetilde{F}_n$ is smaller than that of $\overline{F}_n$. the performance of usual kernel distribution estimator $\widehat{F}_n$ is worse than the performance of the estimator $\widetilde{F}_n$. For the Pareto type II distribution, much the best, in terms of both $AMSE$ and $AMISE$, is the proposed estimator $\widetilde{F}_n$. Next much the worst, although with performance, is the usual kernel distribution estimator $\widehat{F}_n$. For Pareto type III distribution, the estimator $\widehat{F}_n$ also is overall clearly the worst. The proposed estimator and boundary kernel distribution estimator have rather different performances in this case. Clearly best in terms of $AMSE$ and $AMISE$ terms is the estimator $\widetilde{F}_n$.

The main results of our simulation studies is that the proposed estimator is recommended to improved boundary effect for heavy tailed distributions. We see that overall $\widetilde{F}_n$ is the best choice among the three estimators considered. Indeed, the performance of boundary kernel distribution estimator $\overline{F}_n$ is very disappointing, and this estimator can not be recommended for use. The usual kernel distribution estimator $\widehat{F}_n$ is clearly the worst estimator for the three heavy tailed distribution considered. This is clearly due to the boundary effect. In conclusion,

the proposed method for estimating the cdf that is suitable for heavy-tailed distribution and improves the classical kernel estimator and boundary corrected kernel estimator of cdf.

5. Proofs

*Proof (**of** (2)).* For $x = sb$, $0 \leq s \leq 1$, using the property $K(t) = 1 - K(-t)$, $-s \leq t \leq s$, and a Taylor expansion of order 1. First note that

$$Bias\left(\widehat{F}_n(x)\right) = E\widehat{F}_n(x) - F(x),$$

then,

$$\begin{aligned} E\widehat{F}_n(x) &= EK\left(\frac{x - X_i}{b}\right) \\ &= \int_0^{\infty} K\left(\frac{x-z}{b}\right) f(z)dz. \end{aligned}$$

To calculate the mean of $\widehat{F}_n$, we used the change of variable $t = (x - z)/b$, we have

$$\begin{aligned} E\widehat{F}_n(x) &= b\int_{-1}^{s} K(t)\, f((s-t)b)dt \\ &= b\int_{-1}^{-s} K(t)\, f((s-t)b)dt + b\int_{-s}^{s} (1 - K(-t))\, f((s-t)b)dt \\ &= b\int_{-1}^{-s} K(t)\, f((s-t)b)dt + F(2sb) - b\int_{-s}^{s} K(t)\, f((s+t)b)dt. \end{aligned}$$

Using a Taylor expansion of order 2 on the function $F(.)$ we have

$$F(2sb) = F(0) + f(0)2sb + f'(0)2s^2b^2 + o\left(b^2\right).$$

By the existence and continuity of $f'(.)$ near 0, we obtain for $x = sb$

$$\begin{aligned} F(0) &= F(x) - f(x)sb + \frac{1}{2}f'(x)\, s^2b^2 + o(b^2) \\ f(x) &= f(0) + f'(0)\, sb + o(b) \\ f'(x) &= f'(0) + o(1). \end{aligned}$$

Therefore,

$$F(2sb) = F(x) + f(0)sb + \frac{3}{2}f'(0)\, s^2b^2 + o\left(b^2\right).$$

We obtain

$$\begin{aligned} Bias\left(\widehat{F}_n(x)\right) &= b\int_{-1}^{-s} K(t)\left\{f(0) + f'(0)(s-t)b + o(b)\right\}dt + f(0)sb + \frac{3}{2}f'(0)\, s^2b^2 + o\left(b^2\right) \\ &\quad -b\int_{-s}^{s} K(t)\left\{f(0) + f'(0)(s+t)b + o(b)\right\}dt \\ &= b\left\{f(0)s + f(0)\int_{-1}^{-s} K(t)\,dt - f(0)\int_{-s}^{s} K(t)\,dt\right\} + b^2\left\{\frac{3}{2}f'(0)\, s^2\right. \\ &\quad \left. + f'(0)\int_{-1}^{-s}(s-t)K(t)\,dt - f'(0)\int_{-s}^{s}(s+t)K(t)\,dt\right\} + o\left(b^2\right). \end{aligned}$$

From the symmetry of k, one can write $K(x) = 1/2 + r(x)$, where $r(x) = -r(-x)$ for all x such that $|x| \leq 1$. Thus $\int_{-s}^{s} K(t)dt = s$ and after some algebra we obtain the bias expression as

$$Bias\left(\widehat{F}_n(x)\right) = bf(0)\int_{-1}^{-s} K(t)\,dt + b^2 f'(0)\left\{\frac{s^2}{2} + s\int_{-1}^{-s} K(t)dt - \int_{-1}^{s} tK(t)dt\right\} + o\left(b^2\right).$$

This completes the proof of expression (2).

*Proof (**of** (3)).* Observe that for $x = sb$, $0 \leq s \leq 1$, we have

$$\begin{aligned} Var\left(\widehat{F}_n(x)\right) &= \frac{1}{n^2} Var\left\{\sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right)\right\} \\ &= \frac{1}{n} E\left\{K\left(\frac{x - X_i}{b}\right)\right\}^2 - \frac{1}{n}\left\{E\left\{K\left(\frac{x - X_i}{b}\right)\right\}\right\}^2 \\ &=: I_1 - I_2, \end{aligned}$$

where

$$\begin{aligned} I_1 &= \frac{1}{n} E\left\{K\left(\frac{x - X_i}{b}\right)\right\}^2 \\ &= \frac{1}{n}\int_0^{\infty} K^2\left(\frac{x - z}{b}\right) f(z)dz \\ &= \frac{b}{n}\int_{-1}^{s} K^2(t)\, f((s-t)b)dt \\ &= \frac{b}{n}\int_{-1}^{-s} K^2(t)\, f((s-t)b)dt + \frac{b}{n}\int_{-s}^{s} K^2(t)\, f((s-t)b)dt. \\ &=: I_{11} + I_{12}. \end{aligned}$$

It can be shown that

$$\begin{aligned} I_{11} &= \frac{b}{n}\int_{-1}^{-s} K^2(t)\, f((s-t)b)dt \\ &= \frac{b}{n}\int_{-1}^{-s} K^2(t)\left\{f(0) + o(1)\right\}dt. \end{aligned}$$

We use the identity $K(t) = 1 - K(-t)$ and similarly as in the last proof we obtain I_{12}

$$\begin{aligned} I_{12} &= \frac{b}{n}\int_{-s}^{s} K^2(t)\, f((s-t)b)dt \\ &= \frac{b}{n}\int_{-s}^{s} \left(1 - 2K(-t) + K^2(-t)\right) f((s-t)b)dt \\ &= \frac{b}{n}\int_{-s}^{s} f((s-t)b)dt - 2\frac{b}{n}\int_{-s}^{s} K(t)\, f((s+t)b)dt + \frac{b}{n}\int_{-s}^{s} K^2(t)\, f((s+t)b)dt \\ &= \frac{F(2sb)}{n} - 2\frac{b}{n}\int_{-s}^{s} K(t)\left\{f(0) + o(1)\right\}dt + \frac{b}{n}\int_{-s}^{s} K^2(t)\left\{f(0) + o(1)\right\}dt \\ &= \frac{F(x)}{n} - f(0)s\frac{b}{n} + \frac{b}{n}f(0)\int_{-s}^{s} K^2(t)\,dt + o(\frac{b}{n}), \end{aligned}$$

and now combine I_{11} and I_{12} to obtain I_1. With the expression obtained for the bias we get the expression for I_2 as

$$\begin{aligned} I_2 &= \frac{1}{n}\left\{E\left\{K\left(\frac{x - X_i}{b}\right)\right\}\right\}^2 \\ &= \frac{1}{n}\left\{E\widehat{F}_n(x)\right\}^2 \\ &= \frac{1}{n}F^2(x) + o(\frac{b}{n}). \end{aligned}$$

Finally, we obtain the variance of the estimator $\widehat{F}_n(x)$ as

$$\begin{aligned} Var\left(\widehat{F}_n(x)\right) &= I_1 - I_2 \\ &= \frac{F(x)}{n} + \frac{b}{n}f(0)\left\{\int_{-1}^{s} K^2(t)\,dt - s\right\} - \frac{1}{n}F^2(x) + o(\frac{b}{n}) \\ &= \frac{F(x)\left(1 - F(x)\right)}{n} + \frac{b}{n}f(0)\left\{\int_{-1}^{s} K^2(t)\,dt - s\right\} + o(\frac{b}{n}). \end{aligned}$$

This completes the proof of expression (3).

*Proof (**of Theorem 1**).* We have $X_1, ..., X_n$ are independent identically distributed rv's with density f and cdf F. the Transform kernel distribution estimator of $F(x)$ is

$$\widetilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{K\left(\frac{T(x) - g(T(X_i))}{b}\right) - K\left(-\frac{T(x) + g(T(X_i))}{b}\right)\right\},$$

where $T(\cdot)$ is the transformation function. Let the transformed variable $Y_i = T(X_i)$, have distribution H:

$$H(y) = F\left(T^{-1}(T(x))\right) = F(x),$$

and the density of $H(y)$ as

$$h(y) = \frac{f\left(T^{-1}(y)\right)}{T'\left(T^{-1}(y)\right)},$$

so the boundary kernel distribution estimator of $H(y)$ is

$$\widetilde{H}_n(y) = \frac{1}{n}\sum_{i=1}^{n}\left\{K\left(\frac{y-g(Y_i)}{b}\right) - K\left(-\frac{y+g(Y_i)}{b}\right)\right\}.$$

The transform kernel distribution estimator can be expressed by :

$$\widetilde{F}_n(x) = \widetilde{F}_n\left(T^{-1}(T(x))\right) = \widetilde{H}_n(y),$$

implying the bias of the transform kernel distribution estimator is

$$\begin{aligned} Bias\left(\widetilde{F}_n(x)\right) &= Bias\left(\widetilde{F}_n\left(T^{-1}(T(x))\right)\right) \\ &= Bias\left(\widetilde{H}_n(T(x))\right) \\ &= b^2\left\{h'(T(0))\left(\frac{s^2}{2} + 2s\int_{-1}^{-s}K(t)dt - \int_{-s}^{s}tK(t)dt\right)\right. \\ &\left. - h(T(0))g''(0)\left(\int_{-1}^{s}(s-t)K(t)dt + \int_{-1}^{-s}(s+t)K(t)dt\right)\right\} + o\left(b^2\right), \end{aligned}$$

note that

$$h(T(x)) = \frac{f(x)}{T'(x)},\ h'(T(x)) = \left(\frac{f(x)}{T'(x)}\right)'\frac{1}{T'(x)},$$

then

$$h(T(0)) = \frac{f(0)}{T'(0)},\ h'(T(0)) = \left(\frac{f}{T'}\right)'(0)\frac{1}{T'(0)},$$

which are used to find the mean of the transform kernel distribution estimator

$$\begin{aligned} Bias\left(\widetilde{F}_n(x)\right) = b^2\left\{\left(\frac{f}{T'}\right)'(0)\frac{1}{T'(0)}\left(\frac{s^2}{2} + 2s\int_{-1}^{-s}K(t)dt - \int_{-s}^{s}tK(t)dt\right)\right. \\ \left. - \frac{f(0)}{T'(0)}g''(0)\left(\int_{-1}^{s}(s-t)K(t)dt + \int_{-1}^{-s}(s+t)K(t)dt\right)\right\} + o\left(b^2\right). \end{aligned}$$

By the same idea we calculated the variance

$$\begin{aligned} Var\left(\widetilde{F}_n(x)\right) &= Var\left(\widetilde{F}_n\left(T^{-1}(T(x))\right)\right) \\ &= Var\left(\widetilde{H}_n(T(x))\right) \\ &= \frac{H(y)(1-H(y))}{n} + \frac{b}{n}h(T(0))\left\{2\int_{-1}^{-s}K^2(t)dt - s\right. \\ &\left. + \int_{-s}^{s}K^2(t)dt - 2\int_{-1}^{s}K(t)K(t-2s)dt\right\} + o\left(\frac{b}{n}\right). \\ &= \frac{F(x)(1-F(x))}{n} + \frac{b}{n}\frac{f(0)}{T'(0)}\left\{2\int_{-1}^{-s}K^2(t)dt - s\right. \\ &\left. + \int_{-s}^{s}K^2(t)dt - 2\int_{-1}^{s}K(t)K(t-2s)dt\right\} + o\left(\frac{b}{n}\right). \end{aligned}$$

This completes the proof of Theorem 1.

Acknowledgement. We thank the editor and the referee for their constructive and useful comments that led to a much improved paper.

References

Azzalini, A., 1981. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 326-328.

Buch-Larsen, T., Nielsen, J.P., Guillén, M. and Bolancé, C., 2005. Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics*. **39**(6), 503-516.

Champernowne, D.G., 1952. The graduation of income distributions. Econometrica: *Journal of the Econometric Society*. 591-615.

Cowling, A. and Hall, P., 1996. On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society*. Series B (Methodological), 551-563.

Gasser, T. and Müller, H.G., 1979. Kernel estimation of regression functions. In Smoothing techniques for curve estimation, 23-68. *Springer Berlin Heidelberg*.

Gasser, T., Muller, H.G. and Mammitzsch, V., 1985. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society*. Series B (Methodological), 238-252.

Karunamuni, R.J. and Alberts, T., 2005. A generalized reflection method of boundary correction in kernel density estimation. *Canadian Journal of Statistics*. **33**(4), 497-509.

Karunamuni, R.J. and Alberts, T., 2005. On boundary correction in kernel density estimation. *Statistical Methodology*, **2**(3), 191-212.

Koláček, J. and Karunamuni, R.J., 2009. On boundary correction in kernel estimation of ROC curves. *Austrian Journal of Statistics*. **38**(1), 17-32.

Lehmann, E.L., 1991. Theory of Point Estimation. Wadsworth and Brooks/Cole, Belmont.

Marron, J. S., and Ruppert, D., 1994. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society*. Series B (Methodological), 653-671.

Nadaraya, E.A., 1964. Some new estimates for distribution functions. *Theory of Probability and Its Applications*. **9**(3), 497-500.

Parzen, E., 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*. **33**(3), 1065-1076.

Reiss, R.D., 1981. Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*. 116-119.

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

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

Tenreiro, C., 2013. Boundary kernels for distribution function estimation. *REVSTAT–Statistical Journal*, **11**(2), 169-190.

Wand, M.P. and Jones, M.C., 1995. *Kernel Smoothing*. London: Chapman and Hall.

Watson, G.S., and Leadbetter, M.R., 1964. Hazard analysis II. Sankhyā: *The Indian Journal of Statistics*. Series A, 101-116.

Winter, B.B., 1979. Convergence rate of perturbed empirical distribution functions. *Journal of Applied Probability*. **16**(1), 163-173.

Yamato, H., 1973. Uniform convergence of an estimator of a distribution function, Bulletin of Mathematical Statistics. **15**, 69-78.

Zhang, S., Karunamuni, R.J., and Jones, M.C., 1999. An improved estimator of the density function at the boundary. *Journal of the American Statistical Association.* **94**(448), 1231-1240.

Zhang, S., and Karunamuni, R.J., 2000. On nonparametric density estimation at the boundary. *Journal of nonparametric statistics.* **12**(2), 197-221.