Journal of Sciences, Islamic Republic of Iran 34(4): 349 - 361 (2023) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Beta Kernel Estimator for a Cumulative Distribution Function with Bounded Support

B. Mansouri*, A. Rastin, H. Allah Mombeni

Department of Statistics, Shahid Chamran University of Ahvaz, Ahvaz,Islamic Republic of Iran

Received: 20 November 2023 / Revised: 25 December 2023 / Accepted: 3 February 2024

Abstract

Kernel estimation of the cumulative distribution function (CDF), when the support of the data is bounded, suffers from bias at the boundaries. To solve this problem, we introduce a new estimator for the CDF with support (0,1) based on the beta kernel function. By studying the asymptotic properties of the proposed estimator, we show that it is consistent and free from boundary bias. We conducted an extensive simulation to illustrate the performance of the proposed estimator. The results demonstrate the superiority of the proposed estimator over other commonly used estimators. As an application, we use the estimated CDF for nonparametric simulation. Using a numerical study, we show that the performance of the kernel probability density function (PDF) estimation in which a large sample simulated from the estimated CDF is employed can be noticeably improved. We also use the proposed estimator to estimate the CDF of the household health cost in Iran in 2019.

Keywords: Nonparametric estimation, Kernel estimator, Boundary bias, Bootstrap, Household cost.

Introduction

The usual estimator of the CDF is the empirical distribution function (EDF). Although the Glivenko-Cantelli theorem (1-2) proves the uniform convergence of the EDF to the true CDF, EDF is not smooth. This drawback restricts the application of EDF in many fields (see (3) or (4)). An alternative estimator that provides a smooth estimate is the ordinary kernel estimator (OKE), which is proposed by authors such as Watson and Leadbetter (5) or Nadaraya (6). The asymptotic properties of kernel estimators, including their uniform strong convergence and asymptotic normality, have been investigated by many authors (see (7) and the references therein for a short review of the literature in this field).

Despite the good performance of the kernel estimators with unbounded support, they suffer from the well-known boundary bias when the support of the data is bounded. There are various boundary correction methods for the kernel PDF estimator in the literature. See (8) for an overview of the methods available in this field and their classification. While studies are mainly developed for the kernel PDF estimator, there are few boundary correction methods for the kernel CDF estimator, such as the boundary kernel estimator (BKE) (9-10), the reflection method ((11-12)), the asymmetric kernels ((7) , (13-16)).

The idea of using asymmetric kernels in the context of PDF estimation and non-parametric regression has been discussed in (17-18). Chen (17) proposed a beta kernel PDF estimator for densities defined on (0,1). He shows that his proposed estimator is free from boundary bias when the data support is compact. His idea was that to remove the boundary bias, the support of the kernel

* Corresponding author: Tel:+989166043957; Email: b.mansouri@scu.ac.ir

should coincide with the support of the observations. Charpentier et al. (19) extended Chen's idea to estimate the PDF of a copula since the support of a copula is $(0,1) \times (0,1)$. For some recent studies on beta kernel estimators, see (20-22). However, Zhang and Karunamuni (23) showed that the beta kernel PDF estimator has high variance in the boundary region, so they concluded that the estimator is only free from the boundary bias problem, but it is not free from the boundary problem. Informally speaking, the boundary region includes the design points that are located near the endpoints and the interior region includes the design points located far from the endpoints of the support of the data.

The good performance of asymmetric kernel CDF estimators motivated us to extend the analysis for estimating a CDF with support on $(0,1)$ by the beta kernel. However, our simulations showed that the beta kernel introduced by Chen (17) performs poorly in estimating the CDF at the boundary points. We will discuss this point later. To improve the performance of Chen's beta kernel, especially at the boundary region, we use a small but very effective change in it and introduce a new version of the beta kernel to establish a boundary-free bias estimator for the CDF.

In this paper, we introduce a beta kernel estimator (BTKE) for the CDF and investigate its asymptotic properties, such as the bias, variance, mean squared error (MSE) and mean integrated squared error (MISE). We derive the convergence rate of the proposed estimator and obtain the optimal bandwidth by minimizing the MISE. The numerical studies show that the proposed estimator performs better than other competing estimators. Nonparametric estimation of the CDF has many applications: Goodness-of-fit test and copula estimation (24), estimating receiver operating characteristic (*ROC*) curve ((11) and (25-26)), and estimating survival functions (27) are just examples, to name a few. Another use of the estimated CDF is for simulation and bootstrapping (28). To show the usefulness of our proposed estimator, we conducted a numerical study to compare the performance of the beta kernel PDF estimation in the ***R*** package ***ks*** (29) when the PDF is estimated using two approaches. In the first approach, the PDF is estimated from the raw data. In the second approach, however, the raw data are used for estimating the CDF by BTKE. Then we generate a large sample from the estimated CDF and eventually estimate PDF from the simulated samples. Our numerical study demonstrates that the performance of the beta kernel PDF is much better that in the second approach. Finally, we use the BTKE to estimate the CDF of the proportion of health costs to the total household cost in Iran.

This paper focuses on CDF with support $(0,1)$. Note that if the support of the data is (a,b) for any two real numbers $a < b$, then we can use the simple transform $\frac{x-a}{b-a}$ to locate the support on $(0,1)$.

Throughout the paper, the notation $a_n = o(b_n)$ means that $a_n/b_n \to 0$ as $n \to \infty$ and the notation $a_n = O(b_n)$ means that $a_n \leq Cb_n$ for some positive constant C and for all n sufficiently large. The quantity C can depend on the target CDF F, but no other variable unless explicitly written as a subscript. For example, $a_n = O_x(b_n)$ means that C depends on $x \in (0,1)$.

The simulations and plots in this paper were carried out using ***MATLAB*** and ***R*** software.

Materials and Methods

BTKE for estimating a CDF with compact support

Let $\{X_n,\ n \geq 1\}$ be a sequence of independent and identity random variables with density $f(\cdot)$ (PDF) and distribution function $F(\cdot)$ (CDF) with support (0, 1). Throughout the paper, we make the following two basic assumptions:

Assumption 1. The target CDF *F* has two continuous and bounded derivatives on (0, 1).

Assumption 2. The bandwidth parameter $b = b_n > 0$ is a function of *n* such that $b \to 0\ as\ n \to \infty$.

Chen (1999) proposed the estimation of f by beta kernel PDF estimator,

$$\hat{f}_{n,b}(x) = \frac{1}{n}\sum_{i=1}^{n} k_B\left(X_i;\ \frac{x}{b}+1, \frac{1-x}{b}+1\right), \quad (1)$$

where the kernel $k_B(u;\ \alpha,\beta)$ denotes the density function of a $Beta(\alpha,\beta)$ random variable, and the parameter b is the bandwidth chosen such that $b \to 0$ as $n \to \infty$. What motivated Chen to introduce $\hat{f}_{n,b}$ was the flexible shape of asymmetric beta kernel. In addition, the estimates produced by the beta kernel PDF estimator is always nonnegative because its support matches the supports of the probability densities to be estimated (17). Chen (17) has shown that $\hat{f}_{n,b}$ is free from boundary bias and is appropriate for estimating a PDF defined in the unit interval. Zhang and Karunamuni (23) challenged the results of Chen and showed that the variance of Chen's estimator in the boundary region is high. Also, they showed that the overall performance of $\hat{f}_{n,b}$ is not better than that of other boundary-corrected kernel estimators.

Motivated by the results of Mombeni et al. (7), we introduce the beta kernel estimator (BTKE) for estimating a CDF with support (0,1) as follows:

$$\hat{F}^B_{n,b}(x) = n^{-1}\sum_{i=1}^{n} \bar{K}_B\left(X_i; \frac{x}{b}+b^2, \frac{1-x}{\mathrm{b}}+b^2\right). \quad (2)$$

where

$$\bar{K}_B(u;\ \alpha,\beta) = 1 - K_B(u;\ \alpha,\beta) = 1 -$$

$\int_0^u k_B(t;\, \alpha, \beta)\, dt,$ (3)

denote the survival function of the $Beta(\alpha, \beta)$ distribution and $b > 0$ is the smoothing (or bandwidth) parameter. Through a numerical study, we justify our choices for the parameters of the kernel function in BTKE.

Asymptotic properties of BTKE

In this section, we find the asymptotic expressions for the bias, variance, MSE and MISE for the BTKE. Then, we discuss how to choose an appropriate bandwidth by minimizing the MISE.

Lemma 1 (Bias). Under Assumptions 1 and 2, for any given $x \in (0, 1)$, the bias of the BTKE is

$$Bias\{\hat{F}_{n,b}^B(x)\} = \frac{b}{2}x(1-x)f'(x) + O_x(b^3).$$

Proof. Using integration by part, we have:

$$E_f\big(\hat{F}_{n,b}^B(x)\big) = E_f\left(\bar{K}_{\mathrm{Bet}}\left(T; \frac{x}{b} + b^2, \frac{1-x}{\mathrm{b}} + b^2\right)\right) = E_f\big(1 - K_{\mathrm{Bet}}(T)\big) = E_k\big(F(T)\big),$$

where $E_k\big(F(T)\big)$ is the expectation of $F(T)$, when $T \sim k_{\mathrm{Bet}}\left(t; \frac{x}{b} + b^2, \frac{1-x}{\mathrm{b}} + b^2\right)$. Using Assumptions 1 and 2 and the Taylor expansion, we have:

$$E_f\big(\hat{F}_{n,b}^B(x)\big) = E_k\big(F(T)\big) = F(x) + E(T-x)f(x) + \frac{1}{2}E(T-x)^2 f'(x) + o_x(E(T-x)^2).$$

Considering that $E(T) = (xb^{-1} + b^2)(b^{-1} + 2b^2)^{-1} = x + O(b^3)$ and $E(T-x)^2 \approx \mathrm{Var}(T) = bx(1-x) + O_x(b^4)$, we have:

$$E_k\big(F(T)\big) = F(x) + \frac{b}{2}x(1-x)f'(x) + O_x(b^3),$$

and the proof is complete.

Remark 1 (Boundary bias). Note that for the case where $x = cb$, where $0 < c < 1$, (boundary region) we have:

$$Bias\{\hat{F}_{n,b}^B(x)\} = \frac{b^2}{2}cf'(x) + O_x(b^3).$$

Hence, the rate of convergence of the bias of $\hat{F}_{n,b}^B$ to zero is of order $O(b^2)$. This indicates that $\hat{F}_{n,b}^B$ has uniform bias inside the unit interval in addition to being asymptotically unbiased at the boundary. Obviously, this estimator is then free of boundary bias.

Now, we turn to derive an asymptotic expression for the variance of $\hat{F}_{n,b}^B(x)$. We follow (13) for the proof of the following Lemma.

Lemma 2 (Variance). Under Assumptions 1 and 2, for large n (b small enough) and for any given $x \in (0,1)$ the variance of the BTKE is

$$\mathrm{Var}\left(\hat{F}_{n,b}^B(x)\right) = n^{-1}F(x)\big(1-F(x)\big) - \frac{1}{2}n^{-1}\mathrm{b}^{1/2}f(x)\left(\lim_{b\to 0}\mathrm{b}^{-1/2}E(|T_1 - T_2|)\right) + O_x(n^{-1}b),$$

where $T_1, T_2 \sim \mathrm{Beta}\left(\frac{x}{b} + b^2, \frac{1-x}{\mathrm{b}} + b^2\right)$.

Proof. Since $X_i's$ are i.i.d., we have

$$\mathrm{Var}\left(\hat{F}_{n,b}^B(x)\right) = n^{-1}\left\{E\left(\bar{K}_{\mathrm{Bet}}^2\left(T; \frac{x}{b} + b^2, \frac{1-x}{\mathrm{b}} + b^2\right)\right) - \left(E\left(\hat{F}_{n,b}^B(x)\right)\right)^2\right\},$$

and by using integration by parts and a Taylor expansion we can write,

$$E\left(\bar{K}_{\mathrm{Bet}}^2\left(T; \frac{x}{b} + b^2, \frac{1-x}{\mathrm{b}} + b^2\right)\right) = \int_0^1 \bar{K}_{\mathrm{Bet}}^2\left(t; \frac{x}{b} + b^2, \frac{1-x}{b} + b^2\right) f(t)dt = \int_0^1 F(t)\, 2k_{\mathrm{Bet}}\left(t; \frac{x}{b} + \mathrm{b}^2, \frac{1-x}{b} + \mathrm{b}^2\right)\bar{K}_{\mathrm{Bet}}\left(t; \frac{x}{b} + b^2, \frac{1-x}{b} + b^2\right)dt$$

$$= F(x) + f(x)E(Z-x) + O_x(E(Z-x)^2),$$

where $Z \sim g(z)$ and

$$g(z) = 2k_{\mathrm{Bet}}\left(z; \frac{x}{b} + b^2, \frac{1-x}{b} + b^2\right)\bar{K}_{\mathrm{Bet}}\left(z; \frac{x}{b} + b^2, \frac{1-x}{b} + b^2\right).$$

Let T_1 and T_2 be two independent random variables with distribution $k_{\mathrm{Bet}}\left(t; \frac{x}{b} + b^2, \frac{1-x}{\mathrm{b}} + b^2\right)$ then $\min\{T_1, T_2\} = \frac{1}{2}(T_1 + T_2) - \frac{1}{2}|T_1 - T_2| \sim g(z)$. Integration by parts together with the fact that $E(T_1) = E(T_2) = x + O(b^3)$ yields, for any given $x \in (0,1)$,

$$E\left(\bar{K}_{\mathrm{Bet}}^2\left(T; \frac{x}{b} + b^2, \frac{1-x}{\mathrm{b}} + b^2\right)\right) = F(x) + f(x)\left(-\frac{1}{2}E(|T_1 - T_2|)\right) + O_x(E(Z-x)^2),$$

$$= F(x) - \frac{1}{2}\mathrm{b}^{1/2}f(x)\left(\lim_{b\to 0}\mathrm{b}^{-1/2}E(|T_1 - T_2|)\right) + O_x(b),$$

(Micheaux and Ouimet (13)).

Therefore**,**

$$\mathrm{Var}\left(\hat{F}^B_{n,b}(x)\right) = n^{-1}\left\{E\left(\overline{K}^2_{\mathrm{Bet}}\left(T;\frac{x}{b}+\ b^2,\frac{1-x}{\mathrm{b}}+\ b^2\right)\right)-\left(E\left(\hat{F}^B_{n,b}(x)\right)\right)^2\right\},$$

$$= n^{-1}F(x)\big(1-F(x)\big) - \frac{1}{2}n^{-1}\mathrm{b}^{1/2}f(x)\left(\lim_{b\to 0}\mathrm{b}^{-1/2}E(|T_1-T_2|)\right)+O_x(n^{-1}b).$$

Corollary 1 (**MSE**). For any given $x \in (0,1)$,

$$\begin{aligned}\mathrm{MSE}\left(\hat{F}^B_{n,b}(x)\right) &= \mathrm{E}\left[\left(\hat{F}^B_{n,b}(x)-F(x)\right)^2\right]\\ &= \mathrm{Var}\left(\hat{F}^B_{n,b}(x)\right)+\left(\mathrm{Bias}\left(\hat{F}^B_{n,b}(x)\right)\right)^2\\ &= n^{-1}F(x)\big(1-F(x)\big)\\ &\quad -\frac{1}{2}n^{-1}\mathrm{b}^{1/2}f(x)\left(\lim_{b\to 0}\mathrm{b}^{-1/2}E(|T_1-T_2|)\right)\\ &\quad +\frac{b^2}{4}\big(x(1-x)f'(x)\big)^2+O_x(n^{-1}b)+O_x(b^3),\end{aligned}$$

where $T_1, T_2 \sim \mathrm{Beta}\left(\frac{x}{b}+b^2, \frac{1-x}{\mathrm{b}}+\ b^2\right)$.

For any given $x \in (0,1)$, if $f(x).\lim_{b\to 0}\mathrm{b}^{-1/2}E(|T_1-T_2|).f'(x)\neq 0$, the asymptotically optimal choice of b, with respect to MSE, is

$$b_{opt} = \left[\frac{\frac{1}{2}f(x)\left(\lim_{b\to 0}\mathrm{b}^{-1/2}E(|T_1-\ T_2|)\right)}{\big(x(1-\ x)f'(x)\big)^2}\right]^{2/3} n^{-2/3},$$

with the optimal MSE,

$$\mathrm{MSE}\left(\hat{F}^B_{n,b_{opt}}(x)\right) = n^{-1}F(x)\big(1-F(x)\big) - \frac{3}{4}n^{-4/3}\left[\frac{\left(\frac{1}{2}f(x)\left(\lim_{b\to 0}\mathrm{b}^{-1/2}E(|T_1-T_2|)\right)\right)^4}{\big(x(1-x)f'(x)\big)^2}\right]^{1/3}+O_x\left(n^{-4/3}\right).$$

Proposition1 (MISE). Suppose that Assumptions 1 and 2 holds. Assuming that the target PDF $f = F'$ satisfies

$$\int_0^1 f(x)\left(\lim_{b\to 0}\mathrm{b}^{-\frac{1}{2}}E(|T_1-T_2|)\right)dx<\infty$$

and

$$\int_0^1 x^2(1-x)^2\big(f'(x)\big)^2dx<\infty,$$

then we have:

$$\begin{aligned}MISE\left(\hat{F}^B_{n,b}(x)\right) &= \int_0^1 \mathrm{Var}\left(\hat{F}^B_{n,b}(x)\right)dx\\ &\quad +\int_0^1\left(\mathrm{Bias}\left(\hat{F}^B_{n,b}(x)\right)\right)^2dx\\ &= n^{-1}\int_0^1 F(x)\big(1-F(x)\big)\,dx\\ &\quad -n^{-1}\mathrm{b}^{1/2}\int_0^1\frac{1}{2}f(x)\left(\lim_{b\to 0}\mathrm{b}^{-1/2}E(|T_1\\ &\quad -T_2|)\right)dx\end{aligned}$$

As Micheaux and Ouimet (13) have mentioned, the quantity $\lim_{b\to 0}\mathrm{b}^{-1/2}E(|T_1-T_2|)$ needs to be approximated numerically. However, the optimal bandwidth in Eq. (4) also depends on the unknown PDF, $f(x)$. In practice, we use a cross-validation bandwidth selector inspired by the approach introduced in (30). They define the cross-validation (CV) function as

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\int\left\{I(X_i\le x)-\hat{F}_{-i}(x)\right\},$$

In the above expression, the symbol $\hat{F}_{-i}(x)$ represents the kernel estimate based on all observations except the *i*th observation.

Results and Discussion

In this section, we illustrate the performance of the BTKE via a simulation study and compare the results with some other CDF estimators. We considered eight various distributions with support (0,1) including A:U(0, 1), B: Beta (2, 2), C: Beta (2, 4), D: Beta (4, 2), E: T.Exp. (0.5), F: T.Normal (0,0.25), G: T.Lognormal (0,1) and H: T. HalfNormal (0,1), where T. denotes truncated. In this paper, we have compared the numerical performance of BTKE against three traditional estimators, the EDF, the OKE ((5) or (6)) and the BKE (9). In both OKE and BKE, the Epanechnikov kernel, i.e. $k(u) = 3/4(1-u^2).I(u\le 1)$ was used and the optimal bandwidth proposed by (31) and (9) were used for OKE and BKE, respectively.

We used samples of sizes n =100, 500 and 1000 from eight various distributions. In order to estimate the bandwidth for the BTKE, we used the cross-validation method proposed by (30). As an error metric, we considered the integrated squared error (ISE), $ISE = \int_0^1(\hat{F}(x)-F(x))^2dx$ where $\hat{F}(x)$ denotes the above CDF estimators (BTKE, OKE, BKE and EDF). In our setting, we approximated the integral with the summation.

Table 1 shows the mean and standard deviation of ISE (in parentheses) in 1000 repetitions for each of the estimators and various distributions for three different sample sizes. To simplify the comparison, in each case,

Table 1. The mean and standard deviation of the ISE (1000 repetitions) in estimating eight distributions via four methods (seethe text for explanation) for n =100, 500 and 1000.(values are $\cdots \times 10^{-4}$).

Distribution	Sample size	BTKE	OKE	BKE	EDF
	100	11.0(11.8)	15.1(11.9)	13.0(13.1)	16.6(11.1)
U(0, 1)	500	2.4(2.6)	4.1(2.5)	2.7(2.7)	3.2(2.8)
	1000	1.3(1.3)	2.5(1.3)	1.4(1.4)	1.6(1.4)
	100	9.4(9.6)	11.5(10.4)	9.7(10.3)	12.9(11.0)
Beta (2, 2)	500	2.1(2.1)	2.8(2.3)	2.2(2.2)	2.6(2.3)
	1000	1.1(1.0)	1.5(1.1)	1.2(1.0)	1.3(1.0)
	100	8.0(7.9)	9.3(8.3)	8.2(7.9)	10.0(8.5)
Beta (2, 4)	500	1.7(1.5)	2.2(1.7)	**1.7(1.6)**	1.9(1.6)
	1000	0.91(0.88)	1.2(0.93)	0.93(0.89)	0.99(0.88)
	100	8.0(7.9)	9.5(8.0)	**7.9(8.1)**	9.9(8.2)
Beta (4, 2)	500	1.8(1.7)	2.3(1.9)	**1.8(1.7)**	2.1(1.8)
	1000	0.88(0.83)	1.2(0.92)	0.89(0.82)	0.98(0.83)
	100	10.4(11.9)	14.3(12.4)	10.91(11.22)	13.9(11.5)
T.Exp. (0.5)	500	2.2(2.2)	3.9(2.4)	2.3(2.2)	2.7(2.3)
	1000	1.2(1.2)	2.4(1.4)	1.3(1.1)	1.4(1.2)
	100	9.8(11.3)	13.0(11.7)	10.5(11.0)	13.8(11.6)
T.Normal (0,0.25)	500	2.3(2.5)	3.5(2.8)	2.4(2.4)	2.8(2.5)
	1000	1.1(1.1)	2.0(1.2)	1.2 (1.1)	1.3(1.1)
	100	9.5(10.3)	13.5(11.2)	10.9(11.5)	14.3(12.3)
T.Lognormal (0,1)	500	2.2(2.3)	3.5(2.4)	2.5(2.4)	2.9(2.5)
	1000	1.1(1. 1)	1.9(1. 2)	1.2(1. 2)	1.4(1. 2)
	100	10.4(11.7)	14.5(12.1)	12.8(12.3)	13.7(15.8)
T.Half Normal(0,1)	500	2.7(2.9)	4.2(2.8)	3.0(2.9)	3.1(3.3)
	1000	1.3(1. 1)	2.3(1. 2)	1.5(1. 3)	1.3(1. 5)

we bolded the lowest mean ISE. Note that, in all cases, the mean and standard deviation of the ISE decreased as the sample size increased. The results demonstrate the superiority of the BTKE in comparison with other competitors as the BTKE provides the lowest ISE in almost all scenarios. There are few cases where the BKE is at the par of the BTKE and there is an exception (Beta(4,2) with sample size n =100) where the BKE has the lowest ISE among others. However, even for this case, the ISE of BTKE is very close and when the sample size increases, the BTKE gains its superiority. In Figure 1, we illustrated the results given in Table 1 for the sample size n = 1000 via boxplot. The good performance of the BTKE is obvious in all scenarios. Mombeni et al. (7) showed that the performance of estimators depends on the design point where the estimation is performed. Figure 2 depicts the MSE of the three kernel type estimators (the BTKE, the OKE, and the BKE) at various points of the support of the considered distributions in 1000 repetitions for the sample size n = 1000. Poor performance of the OKE is significant in the boundary regions. The MSE of the two estimators BTKE and BKE in the boundary regions is considerably much less than the MSE of the OKE. These two estimators show almost the same performance in most places.

In order to evaluate the variance of the proposed estimator in various design points, in Figure 3, we depict the variance of estimating the considered distributions by BTKE, EDF, BKE and OKE in 1000 repetitions for the sample size n =1000. As can be expected, in all cases, EDF poses the largest variance in most design points even for a sample of size 1000. The variance of OKE in the boundary region is very large although in the interior region it has a smaller variance than the other estimators. In most cases, the variance of the proposed estimator is less than the variance of the BKE in almost all design points.

To see an example of the estimated distributions, in Figure 4, we showed 30 estimates of U(0,1) via four methods in blue with the actual CDF in bold red (n =200). As expected, the OKE in the boundary region is biased and the empirical estimates are not smooth. In fact, the proposed estimator does not have a boundary problem and provides smooth estimates.

The kernel parameters in BTKE

This section is devoted to discussing the kernel choice in our proposed estimator. At first, motivated by Chen's estimator (17), we decided to use the following beta kernel estimator for a CDF with support (0,1)

$$\hat{F}_{n,b}^{chen}(x) = n^{-1}\sum_{i=1}^{n} \overline{K}_B\left(X_i; \frac{x}{b}+1, \frac{1-x}{\mathrm{b}}+1\right),$$

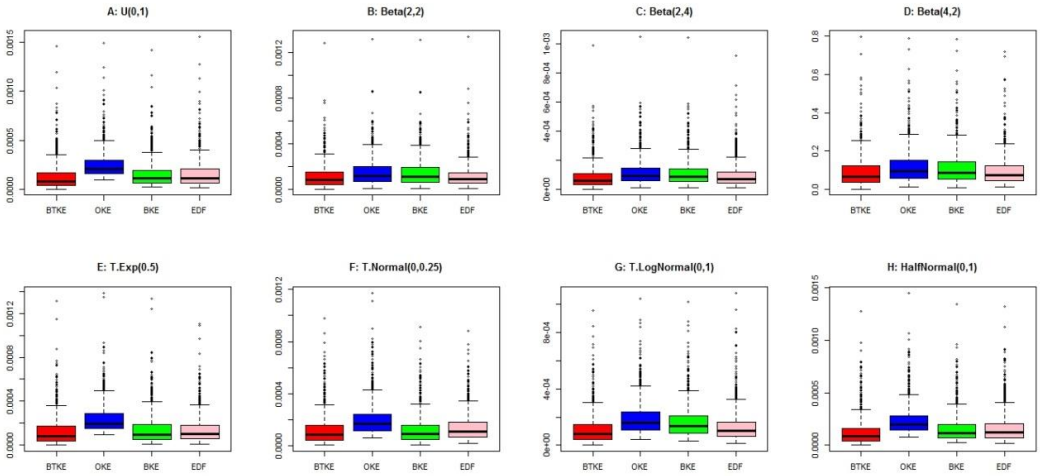


Figure 1. Boxplots of the ISE in 1000 repetitions (n = 1000) for the BTKE, the OKE, the BKE and the EDF.

where $\bar{K}_B(u;\ \alpha,\beta)$ is defined in Eq.(3), and $b > 0$ is the smoothing (or bandwidth) parameter. We investigated the performance of $\hat{F}_{n,b}^{chen}(x)$ via a simulation study and what was found was discouraging. Actually, the performance of the estimator in the boundary region is poor, though its performance in the interior region is appropriate. To clarify the discussion, consider the following example. Lets simulate 100 samples of size 100 from a truncated exponential (T. Exp.) distribution with the CDF

$$F(x) = \begin{cases} 0, & x < 0 \\ \dfrac{1-\exp\left(-\frac{x}{a}\right)}{1-\exp\left(-\frac{1}{a}\right)}, & 0 < x < 1 \\ 1, & x > 1 \end{cases}$$

with parameter $a = 0.25$ and then estimate the CDF using $\hat{F}_{n,b}^{chen}(x)$. In Figure 5a, 100 estimates (dashed curves) and the true distribution (bold solid curve) are presented. We labelled $\hat{F}_{n,b}^{chen}(x)$ as CBTKE in the figures. The severe bias of the estimates at the left boundary is obvious. Figure 6a depicts the MSE of the 100 estimates using $\hat{F}_{n,b}(x)$ in various design points x where the estimation is performed. In order to assess the performance of the $\hat{F}_{n,b}^{chen}(x)$, we added the MSE in 100 estimations by the OKE and the BKE. As can be seen, the MSE of $\hat{F}_{n,b}^{chen}(x)$ and the OKE are increased as x approaches the left bound zero. The performance of $\hat{F}_{n,b}^{chen}(x)$ is even worse than the OKE at the boundary region.

In order to remedy the bad performance of $\hat{F}_{n,b}^{chen}(x)$ at the boundary region, we introduced $\hat{F}_{n,b}^{B}(x)$ for estimating a CDF with support $(0,1)$. Figure 5b shows 100 estimates (dashed curves) by the BTKE and the true distribution of the truncated exponential distribution with parameter $a = 0.25$ (bold solid curve). The sample size is 100. The estimates converge to the true CDF in the boundary region and expose a very good performance. Figure 6b compares the MSE of the BTKE, the OKE and the BKE in 100 estimates of CDF of the truncated exponential distribution with parameter $a = 0.25$. Consider that the MSE of the proposed estimator is even lower than the MSE of the BKE at the left boundary region.

As a part of our study, we investigated an estimator of the form

$$\hat{F}_{n,b}^{B(s,t)}(x) = n^{-1}\sum_{i=1}^{n}\bar{K}_B\left(X_i;\frac{x}{b^s}+b^t,\frac{1-x}{b^s}+b^t\right),$$

to find the best combination of $t > 0$ and $s > 0$. In two Lemmas 3.1 and 3.2, we have derived the bias and variance of BTKE, where *s* is 1 and *t* is 2, respectively. Theoretically, it is difficult to conclude about the best combination of *s* and *t*. Therefore, we resorted to a numerical study to find the best *s* and *t* combination and we considered many different distributions. What we found was that there is almost the same pattern for various distributions. To explain the pattern, consider Figure 7 which shows the MISE of $\hat{F}_{n,b}^{B(s,t)}$ for *t* and *s* values from 0.1 to 0.9 and 1 to 20 in estimating four CDFs: a)T.Exp. (0.5), b)T.Lognormal(0,1), c)T.Normal (0,0.25), and d)Beta (4, 2),where T. denotes truncated, using 1000 simulated samples of size 100 from these distributions. As can be seen in Figure 7, the MISE decreases with a large slope for *s* values between 0.1 and

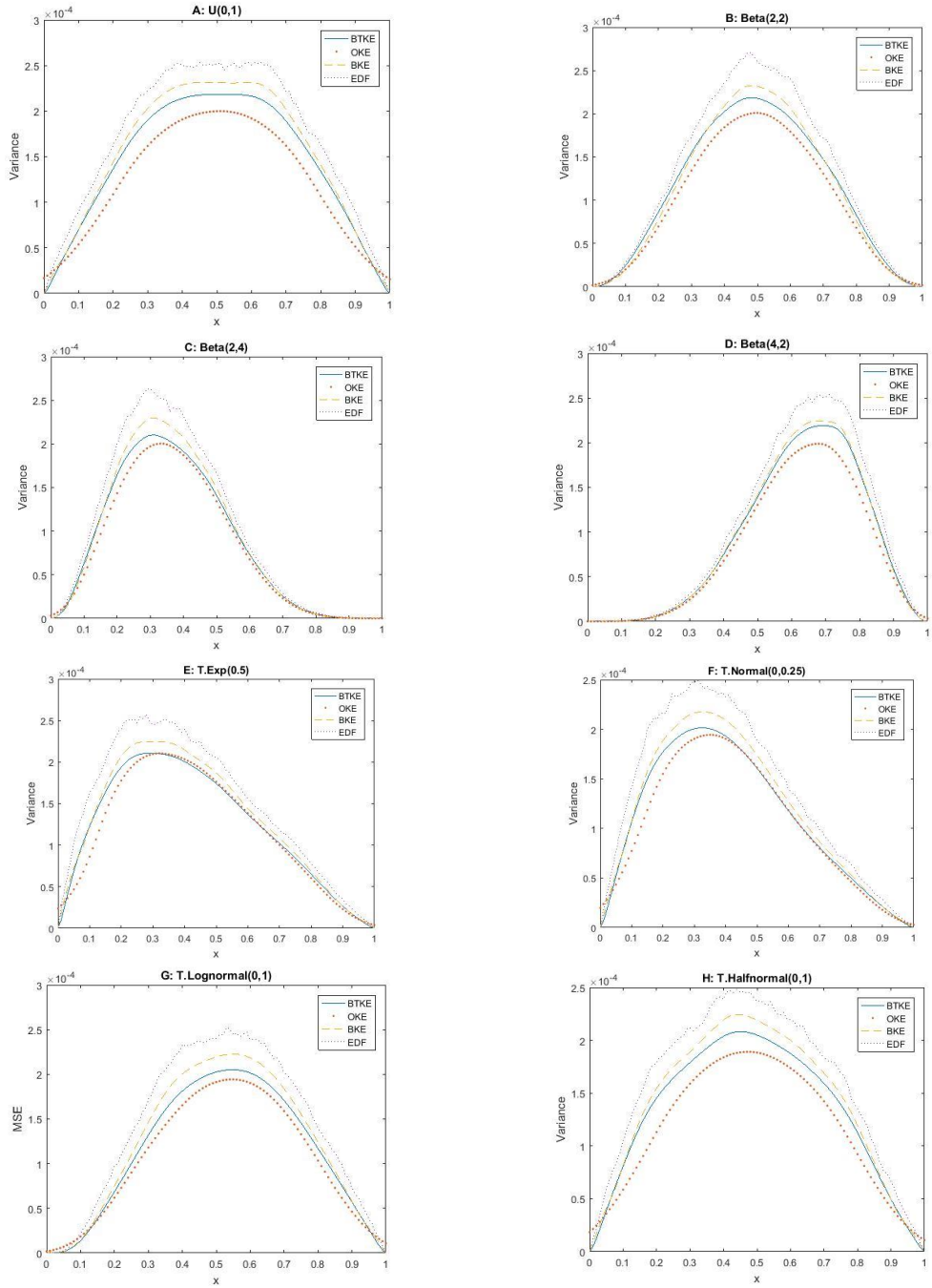


Figure 3. The plot of the variance in estimating eight distribution functions via four methods in 1000 repetitions (n = 1000) (see the text for further explanation).

1, but then for values of *s* larger than 1, the decrease in the MISE is not substantial. As seen for *t*, the results for *t* equal to 2 are acceptable. Of course, lower values of the MISE are possible for *t* larger than2 and s larger than 1, but the reduction is not very significant, and due to the simplicity of the form, we suggest the combination $s = 1$ and $t = 2$ i.e. BTKE. Our comprehensive simulations with various distributions and various sample sizes, from which only a part is presented in simulation, demonstrated that the change we use in the beta kernel is

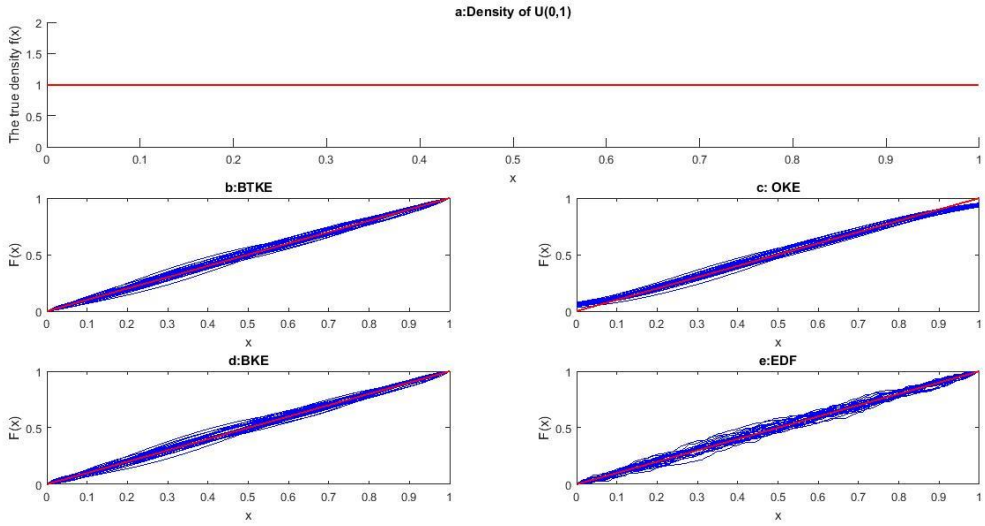
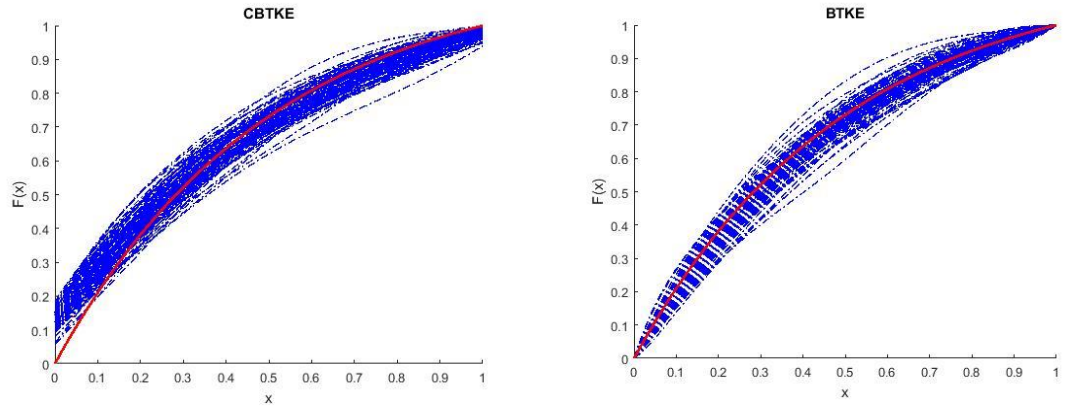


Figure 4. Thirty estimates of U(0,1) via four methods in blue with the actual CDF in bold red (n =200). The top plot (a) shows the PDF of U(0,1).



(a) (b)

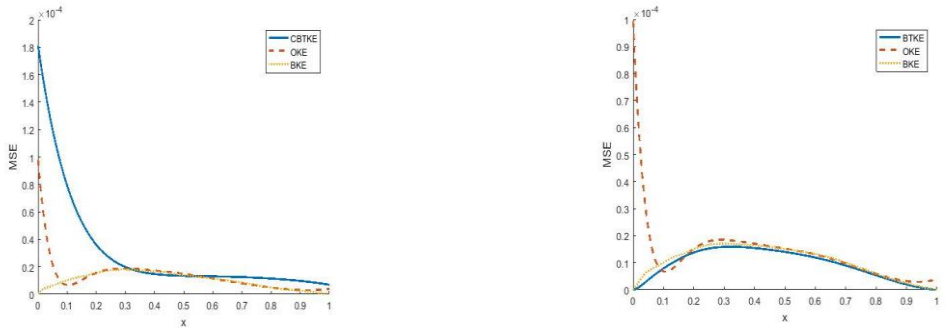
Figure 5. One hundred estimates (dashed curves) of CDF of T.Exp. distribution with parameter a=0.5 (a) using $\tilde{F}_b(x)$and (b) using $\hat{F}_{BTKK}(x)$. The true c.d.f. is shown by the bold solid curve.

very effective and that the BTKE outperforms not only the BKE but also the other competitors.

Simulation by BTKE to improve the performance of the kernel PDF estimation

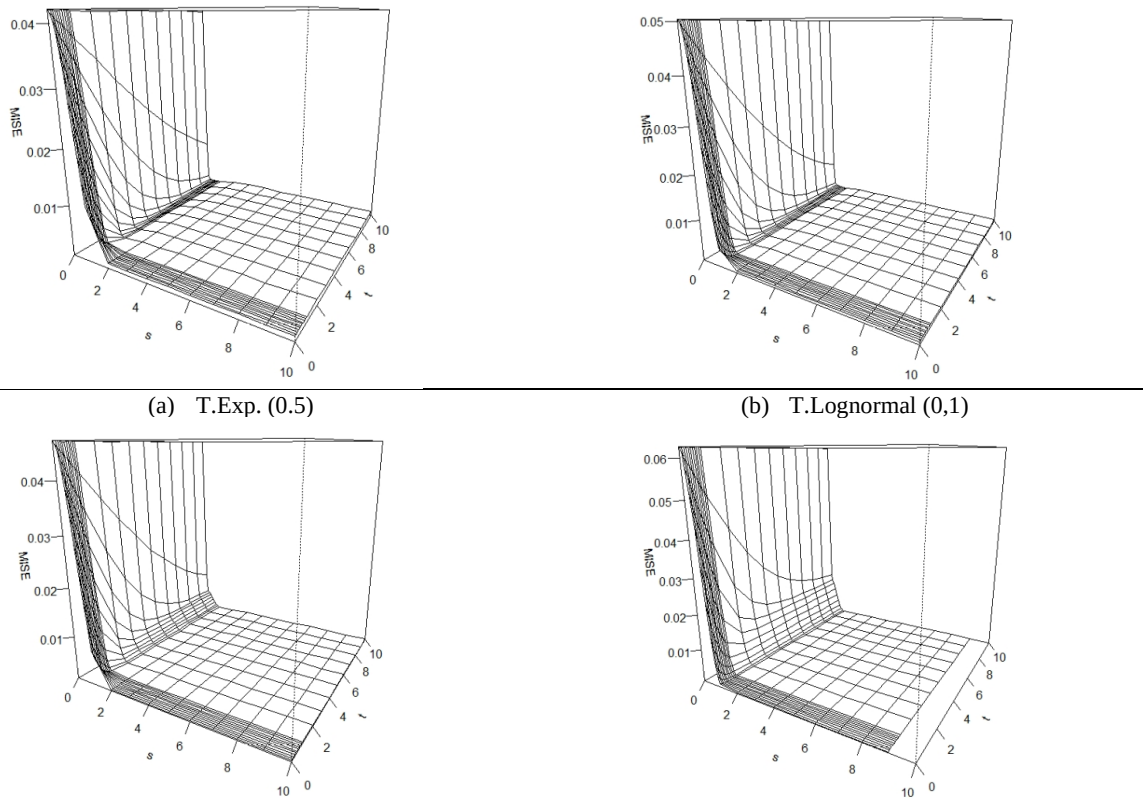
Kernel PDF estimation is a fundamental tool in nonparametric inference and has applications in many various fields. Many authors have shown the consistency of symmetric and asymmetric kernel PDF estimators ((17), (32) and (33)). Often, the MISE of the kernel PDF estimators is $O(n^{-4/5})$. Therefore, we expect lower MISE and more accurate estimates for larger sample sizes. When the sample size is small, we can use $\hat{F}^B_{n,b}$ to simulate a larger sample in order to improve the performance of the kernel PDF estimation. In fact, in statistical inferences, the quality of estimates will be improved by increasing the sample size. Although we focused on the problem of PDF estimation in this section, the method proposed here can be used to enhance the performance in any estimation problem. Note that, unlike the classic symmetric kernels where we obtain the estimate of CDF by integrating the PDF estimate, the $\hat{F}^B_{n,b}$ estimates the CDF directly from the data.

In statistics, in order to generate a random variable from a continuous CDF F, we use $X = F^{-1}(U) = \inf\{x: F(x) \geq U\}$ where $U \sim U(0,1)$. When F is unknown, given i.i.d. random sample $X_1, \ldots, X_n$ from F,



(a) (b)

Figure 6. The MSE in 100 estimations of T. Exp. distribution with parameter a=0.5 via three estimators. (a) $\tilde{F}_b(x)$ (solid curve), the OKE (dashed curve), and the BKE (dotted curve). (b) the BTKE (solid curve), the OKE (dashed curve), and the BKE (dotted curve).



(a) T.Exp. (0.5) (b) T.Lognormal (0,1)

(c) T.Normal (0,0.25) (d) Beta (4, 2)

Figure 7. The MISE in 1000 estimations of four distributions: a) T. Exp. (0.5), b) T.Lognormal (0,1), c) T.Normal (0,0.25) and d) Beta (4, 2), where T. denotes truncated, for combinations of $t, s = 0.1,0.2, \dots ,0.9,1,2, \dots ,20$.

we generate new samples $\mathrm{X}_1^*, \dots, \mathrm{X}_\mathrm{n}^*$ from the distribution F using sampling by replacement from the original samples. Another approach is to use a smooth estimate of the CDF to generate new samples (see (28) p. 266 for more details). In this section, we propose useing $\hat{F}_{n,b}^B$ estimated from original sample $X_1, \dots, X_n$ to generate new samples $X_1^*, \dots, X_m^*$ for an arbitrarily large m and then useing simulated data for PDF estimation. For this purpose, we generate $\mathrm{U}_1, \dots, \mathrm{U}_m$ from $\mathrm{U}(0,1)$ and then find $X_i^* = \left(\hat{F}_{n,b}^B\right)^{-1}(U_i)$ by interpolating in the plot of $\hat{F}_{n,b}^B(x)$ versus x for $i = 1, \dots, m$.

In order to show the effectiveness of our method, we simulated samples of sizes n =100, 500 and 1000 from eight distributions introduced. For the sake of simplicity,

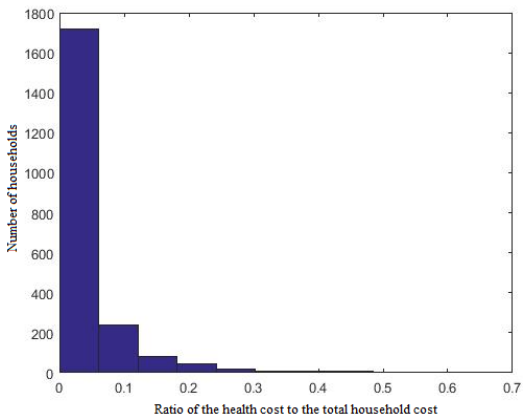
we refer to these samples as original samples. Table 2 shows the MISE in 100 repetitions in estimating PDF of eight distributions using "***kde.boundary***" command in *R* package *ks* via two approaches. In the first approach, PDFs are estimated from the original samples whereas in the second approach, the original samples are used to estimate $\hat{F}_{n,b}^{B}$ and then simulate a sample of size $m = 10000$ to estimate PDF. The command "*kde.boundary*" employs the second form of the beta boundary kernel of (17) and the bandwidth is selected by the plug-in method. In Table 2, the lower MISE is shown by bold face in each case and the last column shows the ratio of MISEs. The results indicate that our proposed method is effective, especially for small and medium sample sizes. The only exceptions are U (0,1) with $n = 500$ and Beta (4,2) with $n = 1000$. In all other cases, MISE is substantially decreased when we use simulated data generated from $\hat{F}_{n,b}^{B}$.

An application to health cost data

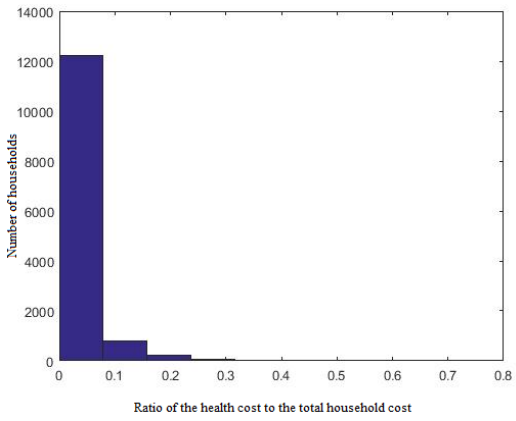
In order to obtain consumption coefficients for calculating the cost-of-living index and household budget, the Statistical Center of Iran has annually

Table 2. The MISE (100 repetitions) in estimating eight PDF using ***ks*** package from the original sample and simulated sample (seethe text for explanation) for $n = 100$, 500 and 1000.

		MISE		
Distribution	Sample size	Original sample	Simulated sample	MISE (Simulate)/ MISE (Original)
	100	0.1121	**0.0641**	0.5716
T.Normal (0,0.25)	500	0.0601	**0.0434**	0.7239
	1000	0.0453	**0.0387**	0.8529
	100	0.0436	**0.0224**	0.5133
T.Exp. (0.5)	500	0.0205	**0.0143**	0.6964
	1000	0.0145	**0.0114**	0.7904
	100	0.0391	**0.0344**	0.8781
T.Lognormal (0,1)	500	0.0149	**0.0125**	0.8427
	1000	0.0091	**0.0083**	0.9132
	100	0.0795	**0.0558**	0.702
T.weibull (2,3)	500	0.0439	**0.0376**	0.8582
	1000	0.0373	**0.0341**	0.9147
	100	0.0447	**0.0235**	0.5266
U (0, 1)	500	**0.0186**	0.0233	1.2523
	1000	0.0138	**0.0118**	0.8549
	100	0.0424	**0.0300**	0.7072
Beta (2, 4)	500	0.0134	**0.0117**	0.8761
	1000	0.0082	**0.0080**	0.9885
	100	0.0438	**0.0342**	0.7803
Beta (4, 2)	500	0.0139	**0.0135**	0.9691
	1000	**0.0078**	0.0087	1.1279
	100	0.0315	**0.0232**	0.7375
Beta (2, 2)	500	0.0104	**0.0089**	0.8541



(a)



(b)

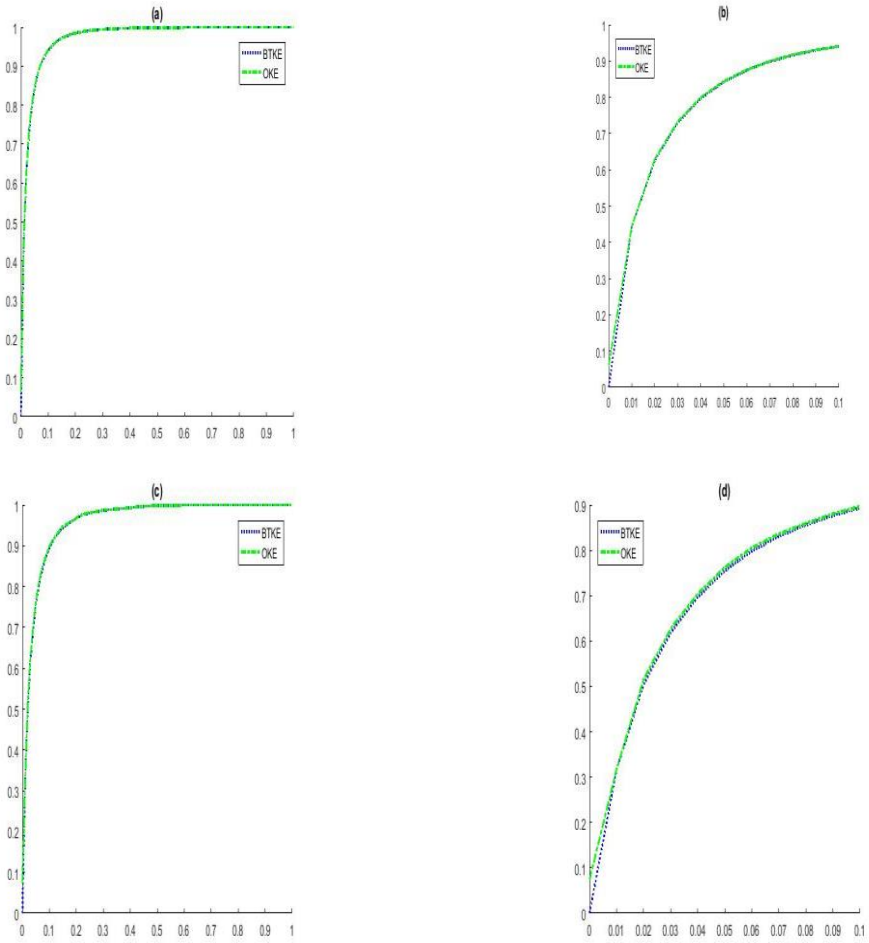


Figure 9. Estimating the CDF of the ratio of health cost to the total household cost via the BTKE (dotted–black curve) and the OKE (dashed –green curve). (a) and (c) represent urban and rural households, respectively. (b) and (d) focus on the left boundary to provide more details for urban and rural households, respectively.

collected statistics on household expenditure and income in various sectors through an extensive sampling program from all urban and rural areas of the country for more than half a century. Prior to measures taken by this center, such statistics were collected by the Central Bank of Iran and Bank Melli Iran for several decades. Household health cost and its portion in the household cost basket in urban and rural areas is one of the main concerns of the health system in any country. In this section, the CDF of the ratio of health costs to total household costs in urban and rural areas of Iran in 1398 AH (2019 AD) has been estimated using the BTKE. The data of this section are the ratio of the health cost to the total household cost and related to 19821 urban households and 18370 rural households taken from the site www.amar.org.ir. Figures 8a and 8b show the histogram of the ratio of health cost to total household cost for urban and rural households, respectively. Data distribution is positively skewed and their focus on the border area is evident.

Figures 9a and 9c show the estimates of the CDF of the ratio of the health cost to the total household cost for urban and rural households, respectively. In addition to the BTKE, the OKE is also plotted in these figures for more comparison. Figures 9b and 9d zoom on the boundary region to show more details for urban and rural households, respectively. The bias of the OKE in the boundary region is obvious.

References

1. Glivenko V. Sulla determinazione empirica delle leggi di probabilita. Gion. Ist. Ital. Attauri..1933;4:92-9.
2. Cantelli FP. Sulla determinazione empirica delle leggi di probabilita. Giorn. Ist. Ital. Attuari. 1933;4(421-424).
3. Azzalini A. A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika.

1981 Apr 1;68(1):326-8.
4. Lejeune M, Sarda P. Smooth estimators of distribution and density functions. Computational Statistics & Data Analysis. 1992 Nov 1;14(4):457-71.
5. Watson GS, Leadbetter MR. Hazard analysis II. Sankhyā: The Indian Journal of Statistics, Series A. 1964 Jul 1:101-16.
6. Nadaraya EA. Some new estimates for distribution functions. Theory of Probability & Its Applications. 1964;9(3):497-500.
7. Mombeni HA, Mansouri B, Akhoond M. Asymmetric kernels for boundary modification in distribution function estimation. REVSTAT-Statistical Journal. 2021 Dec 2;19(4):463-84.
8. Wen K, Wu X. An improved transformation-based kernel estimator of densities on the unit interval. Journal of the American Statistical Association. 2015 Apr 3;110(510):773-83.
9. Tenreiro C. Boundary kernels for distribution function estimation. REVSTAT-Statistical Journal. 2013 Jun 24;11(2):169-90.
10. Tenreiro C. A new class of boundary kernels for distribution function estimation. Communications in Statistics-Theory and Methods. 2018 Nov 2;47(21):5319-32.
11. Koláček J, Karunamuni RJ. On boundary correction in kernel estimation of ROC curves. Austrian Journal of Statistics. 2009;38(1):17-32.
12. Kolácek J, Karunamuni RJ. A generalized reflection method for kernel distribution and hazard functions estimation. Journal of Applied Probability and Statistics. 2011;6(2):73-85.
13. Lafaye de Micheaux P, Ouimet F. A study of seven asymmetric kernels for the estimation of cumulative distribution functions. Mathematics. 2021 Oct 16;9(20):2605.
14. Mansouri B, AtiyahSayyid Al-Farttosi S, Mombeni H, Chinipardaz R. Estimating cumulative distribution function using gamma kernel. Journal of Sciences, Islamic Republic of Iran. 2022 Mar 1;33(1):45-54.
15. Mansouri, B., AtiyahSayyid Al-Farttosi S., Mombeni, H., & Chinipardaz, R. Statistical analysis and estimation of the cumulative distribution function of COVID-19 cure duration in Iraq. Journal of Statistics and Management Systems. 2022; 25(8), 2101-2112.
16. Mombeni HA, Mansouri B, Akhoond M. Estimating receiver operating characteristic curve (ROC) using Birnbaum-Saunders kernel, Journal of Advanced Mathematical Modeling. 2022; (12)3:344-356.
17. Chen SX. Beta kernel estimators for density functions. Computational Statistics & Data Analysis. 1999 Aug 28;31(2):131-45.
18. Chen SX. Beta kernel smoothers for regression curves. StatisticaSinica. 2000 Jan 1:73-91.
19. Charpentier A, Fermanian JD, Scaillet O. The estimation of copulas: Theory and practice. Copulas: From theory to application in finance. 2007:35-64.
20. Bertin K, Klutchnikoff N. Minimax properties of beta kernel estimators. Journal of statistical planning and inference. 2011 Jul 1;141(7):2287-97.
21. Bertin K, Klutchnikoff N. Adaptive estimation of a density function using beta kernels. ESAIM: Probability and Statistics. 2014;18:400-17.
22. Igarashi G. Bias reductions for beta kernel estimation. Journal of Nonparametric Statistics. 2016 Jan 2;28(1):1-30.
23. Zhang S, Karunamuni RJ. Boundary performance of the beta kernel estimators. Journal of Nonparametric Statistics. 2010 Jan 1;22(1):81-104.
24. Omelka M, Gijbels I, Veraverbeke N. Improved kernel estimation of copulas: weak convergence and goodness-of-fit testing.
25. Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. Journal of the American Statistical Association. 1998 Dec 1;93(444):1356-64.
26. Pulit M. A new method of kernel-smoothing estimation of the ROC curve. Metrika. 2016 Jul;79(5):603-34.
27. Duong T. Non-parametric smoothed estimation of multivariate cumulative distribution and survival functions, and receiver operating characteristic curves. Journal of the Korean Statistical Society. 2016 Mar 1;45(1):33-50.
28. Simonoff JS. Smoothing methods in statistics. Springer Science & Business Media; 2012 Dec 6.
29. Duong T. ks: kernel smoothing R package version 1.12. 0. http://CRAN. R-project. org/package= ks. 2021.
30. Bowman A, Hall P, Prvan T. Bandwidth selection for the smoothing of distribution functions. Biometrika. 1998 Dec 1;85(4):799-808.
31. Altman N, Leger C. Bandwidth selection for kernel distribution function estimation. Journal of Statistical Planning and Inference. 1995 Aug 1;46(2):195-214.
32. Silverman BW. Monographs on statistics and applied

probability. Density estimation for statistics and data analysis. 1986;26.
33. Bouezmarni T, Scaillet O. Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data. Econometric Theory. 2005 Apr;21(2):390-412.