Estimating a Distribution Function at the Boundary

Shunpu Zhang
University of Central Florida

Zhong Li
Zhejiang Sci-Tech University

Zhiying Zhang
Zhejiang Sci-Tech University

Abstract

Estimation of distribution functions has many real-world applications. We study kernel estimation of a distribution function when the density function has compact support. We show that, for densities taking value zero at the endpoints of the support, the kernel distribution estimator does not need boundary correction. Otherwise, boundary correction is necessary. In this paper, we propose a boundary distribution kernel estimator which is free of boundary problem and provides non-negative and non-decreasing distribution estimates between zero and one. Extensive simulation results show that boundary distribution kernel estimator provides better distribution estimates than the existing boundary correction methods. For practical application of the proposed methods, a data-dependent method for choosing the bandwidth is also proposed.

Keywords: boundary distribution kernel, distribution function estimation, integrated mean squared error, kernel estimator, mean squared error.

1. Introduction

As an effect of global warming, the insurance industry is increasingly exposed to extreme events such as hurricanes, hail storms and tornados, etc. Such events cause catastrophic losses. It is necessary to estimate the probability of such events and the probability of the payout exceeding certain amounts (such as \$1,000,000) in order for the insurance companies to determine the appropriate premiums. Denote by X the amount of the payout from an accident, the quantity of interest is $P(X > x)$, where x is a pre-specified amount of payout.

In this paper, we assume that X is a random variable from a population with density f and cumulative distribution function (CDF) $F = P(X \leq x)$ (hence $P(X > x) = 1 - F(x)$). Further assume that we have available the past data $X_1, \cdots, X_n$ and assume that $X_1, \cdots, X_n$ are independent and identically distributed (*i.i.d.*) random variables, a commonly used method to estimate F is the kernel method

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

where K is defined by

$$K(t) = \int_{-\infty}^{t} k(y)\,dy, \tag{2}$$

and k is a kernel function satisfying

$$\int_{-\infty}^{+\infty} k(y)\,dy = 1, \int_{-\infty}^{+\infty} yk(y)\,dy = 0, \int_{-\infty}^{+\infty} y^2 k(y)\,dy \neq 0. \tag{3}$$

$\hat{F}_n(x)$ is called a kernel distribution estimator. In the literature of density estimation, a kernel satisfying (3) is called an order (0, 2) kernel, where "0" means that the purpose is to estimate the density function and "2" means that such kernel yields bias of order $O(h^2)$, see Gasser and Műller (1979), Gasser, Műller, and Mammitzsch (1985), Műller (1991) and Zhang and Karunamuni (1998, 2000) for more references on this topic. Since the purpose of this paper is to discuss the estimation of the distribution function, to distinguish K from k, we will call K defined by (2) a distribution kernel and k a density kernel.

For the kernel distribution estimator (1), it has been shown (Azzalini 1981; Jones 1990) that

$$Bias(\hat{F}_n(x)) = E\hat{F}(x) - F(x) = \frac{1}{2}h^2 \mu_2(k) f^{(1)}(x) + o(h^2) \tag{4}$$

and

$$Var(\hat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n} - \frac{h}{n} f(x)\psi(K) + o(\frac{h}{n}), \tag{5}$$

where $f^{(1)}$ is the second derivative of F (or the first derivative of f), $\mu_2(K) = \int y^2 k(y)\,dy$ and $\psi(K) = 2\int yk(y)K(y)\,dy$, respectively.

Combining (4) and (5), we have

$$MSE(\hat{F}_n(x)) \cong \frac{F(x)(1 - F(x))}{n} + \frac{1}{4}h^4 [\mu_2(k)]^2 [f^{(1)}(x)]^2 - \frac{h}{n} f(x)\psi(K) \tag{6}$$

and

$$IMSE(\hat{F}_n(x)) \cong \frac{1}{n}\int F(x)(1 - F(x))dx + \frac{1}{4}h^4 [\mu_2(k)]^2 \int [f^{(1)}(x)]^2 dx - \frac{h}{n}\psi(K), \tag{7}$$

where MSE and IMSE are the abbreviations of mean squared error and integrated mean squared error.

It can be seen that the optimal bandwidths for minimizing (6) and (7) are both of order $O(n^{-1/3})$ and have the form

$$h_L^{opt}(x) = \left\{ \frac{\psi(K)}{[\mu_2(k)]^2} \right\}^{1/3} \left\{ \frac{f(x)}{[f^{(1)}(x)]^2} \right\}^{1/3} n^{-1/3} \tag{8}$$

and

$$h_G^{opt} = \left\{ \frac{\psi(K)}{[\mu_2(k)]^2} \right\}^{1/3} \frac{1}{\left\{ \int [f^{(1)}(x)]^2\,dx \right\}^{1/3}} n^{-1/3}, \tag{9}$$

respectively.

The bandwidths $h_L^{opt}(x)$ and h_G^{opt} are called the optimal local and optimal global bandwidths. With their respective optimal bandwidths, the MSE and IMSE of $\hat{F}_n(x)$ are

$$MSE(\hat{F}_n(x)) \cong \frac{F(x)(1 - F(x))}{n} - \frac{3}{4}\left\{ \frac{[f(x)]^4}{[f^{(1)}(x)]^2} \right\}^{1/3} \left\{ \frac{[\psi(K)]^4}{[\mu_2(k)]^2} \right\}^{1/3} n^{-4/3} \tag{10}$$

and

$$IMSE(\hat{F}_n(x)) \cong \int \frac{F(x)(1-F(x))}{n}dx - \frac{3}{4}\frac{1}{\left\{\int [f^{(1)}(x)]^2 dx\right\}^{1/3}}\left\{\frac{[\psi(K)]^4}{[\mu_2(k)]^2}\right\}^{1/3}n^{-4/3}. \quad (11)$$

Hence, the optimal rate of convergence of MSE and IMSE is of order $O(n^{-1})$, the same as that of the empirical cumulative distribution function (CDF) since the last two terms in (10) and (11) converge to zero at the rate $O(n^{-4/3})$. Note that the first term on the right hand side (RHS) of (10) and (11) are the MSE and IMSE of the empirical CDF. Further note that the second term on the RHS of (10) and (11) are both negative, the kernel distribution estimator has smaller MSE and IMSE than the empirical CDF, which is the motivation of developing the kernel distribution function estimators.

It is obvious that minimization of (6) and (7) also concerns the choice of the kernel function. This problem was solved in Swanepoel (1988) and Jones (1990) in which it is shown that the optimal density kernel k which minimizes the IMSE (or equivalently, maximizing $\frac{[\psi(K)]^4}{[\mu_2(k)]^2}$) for all values of h is the uniform density on $[-\sqrt{3}, \sqrt{3}]$. This, without loss of generality, can be changed to the uniform density on $[-1, 1]$, i.e.,

$$k(y) = 1/2, \quad -1 \le y \le 1, \quad (12)$$

due to the invariance of the minimized IMSE to the change of scale. Consequently, the resulting optimal distribution kernel is

$$K(y) = \begin{cases} 0, & y < -1 \\ \dfrac{1+y}{2}, & -1 \le y < 1 \\ 1, & y \ge 1 \end{cases} . \quad (13)$$

2. The boundary problem of the kernel distribution estimator

The asymptotic results discussed in the previous section are obtained under the assumption that the support of the density is $(-\infty, \infty)$. In practice, the data are often obtained from a population whose probability density function has finite support. Examples of such data include the truncated, survival or censored data which often appear in financial and clinical studies. In such cases, the asymptotic results (4)-(7) do not hold anymore for the points near the end of the support. Hence, the kernel distribution estimator (1) may not provide appropriate estimates of the distribution function at such points. The purpose of this section is to provide a detailed study of the performance of the kernel distribution estimator at the points near or at the end of the support and propose a method to correct the boundary problem.

Without loss of generality, assume that the density function has compact support $[a, b]$ (hence $F(a) = 0, F(b) = 1$), and assume that the density kernel k vanishes outside $[-1, 1]$. Then, it can be shown that, for $x = a + ch, 0 \le c \le 1$, we have

$$\begin{aligned} Bias(\hat{F}_n(x)) =& hf(a+)\left\{c[\int_{-1}^{c}k(y)dy - 1] - \int_{-1}^{c}yk(y)dy\right\} \\ &+ \frac{h^2}{2}f^{(1)}(a+)\left\{c^2[\int_{-1}^{c}k(y)dy - 1] - 2c\int_{-1}^{c}yk(y)dy + \mu_{2,c}(k)\right\} + o(h^2), \end{aligned} \quad (14)$$

where $\mu_{2,c}(k) = \int_{-1}^{c} y^2 k(y) dy$. Similarly, for $x = b - ch$, $0 \le c \le 1$, we have

$$
\begin{aligned}
Bias(\hat{F}_n(x)) = & -hf(b-)\left\{c[\int_{-c}^{1} k(y)dy - 1] + \int_{-c}^{1} yk(y)dy\right\} \\
& + \frac{h^2}{2}f^{(1)}(b-)\left\{c^2[\int_{-c}^{1} k(y)dy - 1] + 2c\int_{-c}^{1} yk(y)dy + \mu_{2,c}^{*}(k)\right\} + o(h^2),
\end{aligned}
\tag{15}
$$

where $\mu_{2,c}^{*}(k) = \int_{-c}^{1} y^2 k(y) dy$.

The proofs of (14) and (15) are provided in the Appendix. When $f(a+) \ne 0$ and $f(b-) \ne 0$, (14) and (15) show that, for the points in the intervals $[a, a+h)$ and $(b-h, b]$, the bias of the kernel distribution estimator $\hat{F}_n(x)$ converges to zero at the rate $O(h)$, which is slower than the rate $O(h^2)$ observed in (4). This is the boundary problem of the kernel distribution estimator. Note that the boundary problem in kernel density estimation has the non-consistency problem, in addition to the slower convergence problem of the bias (Gasser and Műller 1979; Gasser *et al.* 1985; Műller 1991; Zhang and Karunamuni 1998). The boundary problem in kernel distribution estimation is less severe than in kernel density estimation. This is due to the extra information $F(a) = 0, F(b) = 1$. As in density estimation, the intervals $[a, a+h)$ and $(b-h, b]$ will be called the boundary region and the interval $a + ch \le x \le b - ch$ is called the interior region. However, if we know that $f(a+) = 0$ or $f(b-) = 0$, the first order term in (14) and (15) disappears and the bias converges to zero at the usual rate $O(h^2)$. Hence, the distribution kernel estimator $\hat{F}_n(x)$ is free of boundary problem in such a case.

Koláček and Karunamuni (2009) considered the boundary problem in distribution function estimation in estimating ROC curves using the transformation method discussed in Zhang et al. (1999). Tenreiro (2013) proposed a boundary kernel method for correcting the boundary problem. However, Tenreiro (2013) did not reveal the fact that there is no boundary problem in distribution function estimation if the density has value zero at the endpoints of the support. In his method, the boundary kernel k_c is constructed by truncating a density kernel at $[-c, c]$, and then normalizing it so that it integrates to 1 on $[-c, c]$. Realizing the fact that such boundary kernel corrects the boundary problem by shrinking the bandwidth to zero when $x \to a+$ or $b-$, the resulting distribution estimates may have high variability at such points.

The purpose of this paper is to develop a boundary distribution kernel method for correcting the boundary problem of $\hat{F}_n(x)$, which is continuous, non-decreasing and does not have the aforementioned high variability problem of the estimator proposed in Tenreiro (2013).

2.1. The boundary distribution kernel estimator

It can be seen from (14) that the boundary problem can be removed if, at the boundary point $x = a + ch$ ($0 \le c \le 1$), the density kernel function k satisfies

$$
c[\int_{-1}^{c} k(y)dy - 1] - \int_{-1}^{c} yk(y)dy = 0.
\tag{16}
$$

Since such k depends on c, to distinguish it from the interior kernel k we will denote it by k_c. It can be easily seen that (16) is satisfied if k_c is a left hand side (LHS) boundary kernel in density estimation (we will call it boundary density kernel) since a LHS boundary density kernel satisfies

$$
\int_{-1}^{c} k_c(y)dy = 1 \; and \; \int_{-1}^{c} yk_c(y)dy = 0.
$$

However, as will be seen in Section 3, the drawback of using the boundary density kernel is that it may produce distribution estimates which are negative or larger than 1 for the distribution due to the negative part on its support. In distribution estimation case, the

derivation of (14) shows that the kernel is only required to satisfy (16) in order to solve the boundary problem of $\hat{F}_n(x)$ at the left boundary. Re-write (16) as

$$\int_{-1}^{c} \frac{c-y}{c} k_c(y) dy = 1. \tag{17}$$

Given an interior symmetric density kernel k which satisfies (3), k_c satisfying (17) can be easily constructed as follows

$$k_c(y) = \begin{cases} ck(y)/\int_{-1}^{c}(c-s)k(s)ds, & -1 \le y \le c \\ \dfrac{1}{c-1}\int_{-1}^{c}sk(s)ds/\int_{-1}^{c}(c-s)k(s)ds, & c < y \le 1 \end{cases}. \tag{18}$$

Define $K_c(t) = \int_{-1}^{t} k_c(y) dy$. $K_c(t)$ will be called a LHS boundary distribution kernel. Noticing that $\int_{-1}^{c} \frac{c-y}{c} k(y) dy \to 1$ as $c \to 1$, we have $k_c \to k$ as $c \to 1$. Hence, k_c defined in (18) is a natural continuation of the interior kernel to the left boundary region.

Similarly, a RHS boundary distribution kernel $K_c^*(t) = \int_{-1}^{t} k_c^*(y) dy$ can be constructed by finding a function k_c^* satisfing

$$c[\int_{-c}^{1} k_c^*(y) dy - 1] + \int_{-c}^{1} y k_c^*(y) dy = 0. \tag{19}$$

or equivalently

$$\int_{-c}^{1} \frac{c+y}{c} k_c^*(y) dy = 1. \tag{20}$$

Given an interior symmetric density kernel k, k_c^* satisfying (20) can be constructed as follows

$$k_c^*(y) = \begin{cases} \dfrac{1}{c+1}\int_{-c}^{1}sk(s)ds/\int_{-c}^{1}(c+s)k(s)ds, & -1 \le y < -c \\ ck(y)/\int_{-c}^{1}(c+s)k(s)ds, & -c \le y \le 1 \end{cases}. \tag{21}$$

It is obvious that $k_c^* \to k$ as $c \to 1$. Hence, k_c^* defined in (21) provides a natural continuation of the interior kernel to the right boundary region. The following result shows that the above defined $k_c(y)$ and $k_c^*(y)$ are always non-negative.

Proposition. *With a symmetric density kernel k satisfying (3), $k_c(y)$ and $k_c^*(y)$ defined in (18) and (21) are non-negative.*

Proof. The proof of the proposition is straight forward. Taking (18) for example, it is obvious that $ck(y)/\int_{-1}^{c}(c-s)k(s)ds$ is non-negative on $-1 \le y \le c$. For the part defined on $c < y \le 1$, the non-negativity follows from the fact that $\int_{-1}^{c} sk(s)ds \le 0$ for $0 \le c \le 1$. In the following we provide examples of $k_c(y)$ and $k_c^*(y)$ and their resulting boundary distribution kernels. $\square$

Examples of boundary distribution kernel

As mentioned at the end of Section 1, the optimal interior distribution kernel is the uniform kernel (12). Hence, it makes more sense to construct a boundary distribution kernel based on

the uniform kernel. Plugging (12) in (18) and (21), we have

$$
k_c(y) = \begin{cases} \dfrac{2c}{(c+1)^2}, & -1 \le y \le c \\ \dfrac{1}{c+1}, & c < y \le 1 \end{cases}
\tag{22}
$$

and

$$
k_c^*(y) = \begin{cases} \dfrac{1}{(c+1)}, & -1 \le y < c \\ \dfrac{2c}{(c+1)^2}, & c \le y \le 1 \end{cases}.
\tag{23}
$$

Hence, the resulting boundary distribution kernels are

$$
k_c(t) = \begin{cases} 0, & t < -1 \\ \dfrac{2c}{(c+1)^2}(t+1) & -1 \le t < c \\ \dfrac{t+c}{c+1}, & c \le t < 1 \\ 1, & t \ge 1 \end{cases}
\tag{24}
$$

for the left boundary, and

$$
k_c^*(t) = \begin{cases} 0, & t < -1 \\ \dfrac{t+1}{c+1} & -1 \le t < -c \\ \dfrac{c^2 + 2ct + 1}{(c+1)^2}, & -c \le t < 1 \\ 1, & t \ge 1 \end{cases}
\tag{25}
$$

for the right boundary, respectively.

With $K_c(t)$ and $K_c^*(t)$ defined above, the boundary distribution kernel estimator is defined as

$$
\hat{F}_B(x) = \begin{cases} \dfrac{1}{n}\sum_{i=1}^{n} K_c(\dfrac{x-X_i}{h}), & a \le x < a+h, c = (x-a)/h \\ \dfrac{1}{n}\sum_{i=1}^{n} K(\dfrac{x-X_i}{h}), & a+h \le x \le b-h \\ p\dfrac{1}{n}\sum_{i=1}^{n} K_c^*(\dfrac{x-X_i}{h}), & b-h \le x \le b, c = (b-x)/h \end{cases}.
\tag{26}
$$

Theorem 1. *Assume that the second derivative of $F(x)$ (or the first derivative of f) exists on $[a,b]$ and is continuous in a neighborhood of x, we have*

i). $\hat{F}_B(a) = 0$ *and* $\hat{F}_B(b) = 1$.
ii). $\hat{F}_B(x)$ *is non-decreasing.*
iii). *For* $x = a + ch, 0 \le c \le 1$,

$$
Bias(\hat{F}_B(x)) = \frac{h^2}{2} f^{(1)}(a+)\left[\int_{-1}^{c}(c-y)^2 k_c(y)dy - c^2\right] + o(h^2)
\tag{27}
$$

and

$$
Var(\hat{F}_B(x)) = \frac{2h}{n} f(a+)\int_{-1}^{c}(c-y)k_c(y)K_c(y)dy + o(\frac{h}{n}).
\tag{28}
$$

iv). *For* $x = b - ch, 0 \le c \le 1$,

$$
Bias(\hat{F}_B(x)) = \frac{h^2}{2} f^{(1)}(b-)\left[\int_{-c}^{1}(c+y)^2 k_c^*(y)dy - c^2\right] + o(h^2)
\tag{29}
$$

and

$$Var(\hat{F}_B(x)) \cong \frac{2h}{n} f(b-) \int_{-c}^{1} (c+y) k_c^*(y)[1 - K_c^*(y)] dy + o(\frac{h}{n}). \tag{30}$$

The proof of Theorem 1 is deferred in the Appendix. Theorem 1 shows that the bias and variance of the boundary distribution kernel estimator $\hat{F}_B(x)$ converges to zero at the interior rate $O(h^2)$ and $O(\frac{h}{n})$ at the boundary points, respectively. Hence the boundary distribution kernel estimator is free of the boundary problem.

Combining (27)-(30) and the fact that the boundary distribution kernel estimator is the same as $\hat{F}_n(x)$ of (1) for the interior points in the interval $[a+h, b-h]$, we have

$$
\begin{aligned}
MISE(\hat{F}_B) \cong &\frac{1}{4} h^4 [\mu_2(k)]^2 \int_{a+h}^{b-h} [f^{(1)}(x)]^2 dx \\
&+ \frac{1}{n} \int_{a+h}^{b-h} F(x)(1 - F(x)) dx - \frac{h}{n} \psi(K) \int_{a+h}^{b-h} f(x) dx \\
&+ \frac{1}{4} h^4 [f^{(1)}(a+)]^2 \int_{a}^{a+h} \left[\int_{-1}^{(x-a)/h} (\frac{x-a}{h} - y)^2 k_{(x-a)/h}(y) dy - (\frac{x-a}{h})^2 \right]^2 dx \\
&+ \frac{2h}{n} f(a+) \int_{a}^{a+h} \int_{-1}^{(x-a)/h} (\frac{x-a}{h} - y) k_{(x-a)/h}(y) K_{(x-a)/h}(y) dy dx \\
&+ \frac{1}{4} h^4 [f^{(1)}(b-)]^2 \int_{b-h}^{b} \left[\int_{-(b-x)/h}^{1} (\frac{b-x}{h} - y) k_{(b-x)/h}^*(y) dy - (\frac{b-x}{h})^2 \right]^2 dx \\
&+ \frac{2h}{n} f(b-) \int_{b-h}^{b} \int_{-(b-x)/h}^{1} (\frac{b-x}{h} + y) k_{(b-x)/h}^*(y)[1 - K_{(b-x)/h}^*(y)] dy dx
\end{aligned}
\tag{31}
$$

Unlike (7), minimization of (31) with respect to (w.r.t.) h does not yield explicit solution and hence has to be solved numerically.

3. Numerical results

In this section, we numerically compare the performances of the empirical CDF (denote it by $\hat{F}(x)$), the kernel distribution estimator $\hat{F}_n(x)$ with distribution kernel (13), the boundary distribution kernel estimator $\hat{F}_B(x)$ with distribution boundary kernels (24)-(25) and the Tenreiro estimator (denote it by $\tilde{F}_n(x)$) with distribution kernel (13).

As mentioned in the discussion circa (16), boundary correction can also be achieved by directly using the density boundary kernels as k_c and k_c^* and their corresponding K_c and K_c^* in the boundary, with the risk that such estimators may produce negative or larger than 1 distribution estimates. To see the extent of the negativity problem, in the comparison we also included the density boundary distribution estimator which is basically (26) with $K_c(t) = \int_{-1}^{t} k_c(y) dy$ and $K_c^*(t) = \int_{-1}^{t} k_c^*(y) dy$, where

$$k_c^*(t) = \left[\frac{4(1 - c + c^2)}{(1+c)^3} - \frac{6(-1+c)}{(1+c)^3} t \right] I[-1, c] \tag{32}$$

and

$$k_c^*(t) = \left[\frac{4(1 - c + c^2)}{(1+c)^3} + \frac{6(-1+c)}{(1+c)^3} t \right] I[-c, 1], \tag{33}$$

respectively. The resulting distribution is denoted by $\hat{F}_B^*(x)$ and will be called the density boundary kernel distribution estimator. It is easy to see that k_c and k_c^* defined in (32) and (33) become the uniform kernel when $c = 1$.

It is obvious that the negativity problem of the density boundary kernel distribution estimator can be easily corrected by truncating the estimates at zero if it takes a negative value or at 1 if it takes a value large than 1. For the completeness of comparison, we also included the truncated version of $\hat{F}_B^*(x)$ in the simulations.

3.1. The performances of the kernel distribution estimators based on their respective optimal bandwidths

To account for different situations, we use the following four distributions in the simulations:

1. The truncated exponential distribution on $[0, 1]$:

$$f(x) = exp(-2x)/\int_0^1 exp(-2x)dx, 0 < x \leq 1.$$

2. The truncated normal distribution on $[0, 1]$:

$$f(x) = \frac{1}{\sqrt{2\pi}}exp(-x^2/2)/\int_0^1 \frac{1}{\sqrt{2\pi}}exp(-x^2/2)dx, p0 < x < 1.$$

3. The Beta(α, β) distribution on $[0, 1]$ with $\alpha = 2, \beta = 2$:

$$f(x) = \frac{x(1-x)}{B(2, 2)}, 0 < x \leq 1,$$

where B is the beta function. Note that this density function satisfies $f(0) = f(1) = 0$.

4. The Beta mixture distribution $\sum_i^k w_i\ Beta(\alpha_i, \beta_i)$ with $k = 2, w_1 = 1/4, \alpha_1 = 1, \beta_1 = 6; w_2 = 3/4, \alpha_2 = 6, \beta_2 = 1$.

Figure 1 below shows the form of the density function and the CDF of these distributions.

In the simulations the sample size was chosen as $n = 100$. Tables 1-4 report the IMSE values of the six estimators. The IMSE values are calculated by integrating the MSE value of each estimator in $[0, 1]$ using their optimal global bandwidths. The use of optimal bandwidth is necessary in order to have a fair comparison since the performance of a kernel estimator is greatly affected by the bandwidth used. The optimal global bandwidth used for each estimator is obtained by numerically minimizing the IMSE of each estimator from 1,000 samples on a grid of possible h values between $[0, 1/2]$. The obtained optimal global bandwidth h_G^{opt} for each estimator is shown in Column 2 of the tables. The IMSE values (multiplied by 1000 for easy comparison) corresponding to h_{opt} (calculated as the average of the integrated squared error (ISE) from the 1,000 samples) are reported in Column 3 of the tables.

It can be seen from Tables 1-4 that all the kernel distribution estimators have smaller IMSE than the empirical CDF. Comparing among the kernel distribution estimators, we see that the boundary distribution kernel estimator $\hat{F}_B(x)$ has the smallest IMSE values for all distributions except Distribution 3, followed by the Tenreiro estimator, the truncated density boundary kernel distribution estimator, the density boundary kernel distribution estimator, and the kernel distribution estimator. Note that for Distribution 3 (the $Beta(2, 2)$), the kernel distribution estimator $\hat{F}_n(x)$ is free of boundary problem since the density satisfies $f(a) = f(b) = 0$. In this case, $\hat{F}_n(x)$ and the boundary distribution kernel estimator $\hat{F}_B(x)$ have similar IMSE values, while the density boundary kernel distribution estimator $\hat{F}_B^*(x)$, its truncated version and the Tenreiro estimator $\tilde{F}_n(x)$ have similar IMSE values which are smaller than those of $\hat{F}_n(x)$ and $\hat{F}_B(x)$.
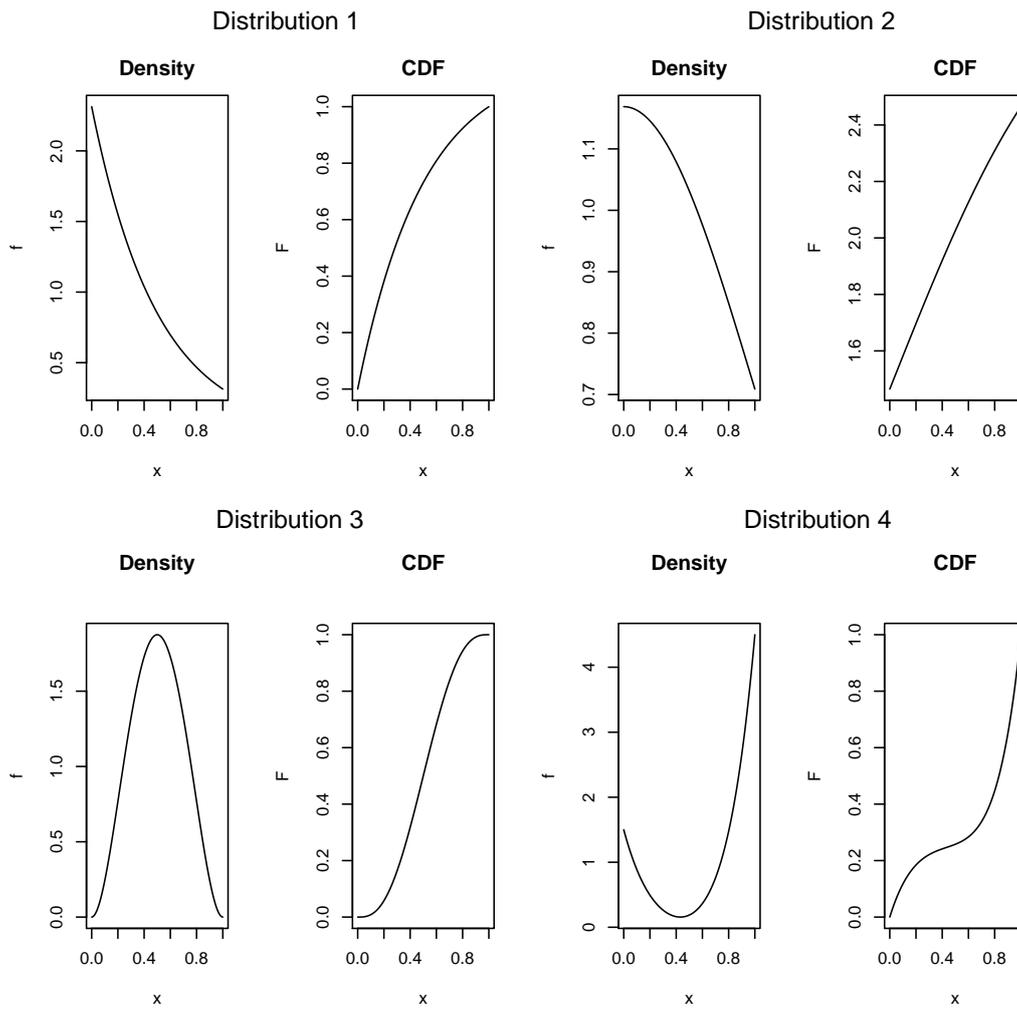
Figure 1: Plots of distribution function and density function

Table 1: The IMSE values of $\hat{F}(x)$, $\hat{F}_n(x)$, $\hat{F}_B^*(x)$, the truncated $\hat{F}_B^*(x)$, $\widetilde{F}_n(x)$ and $\hat{F}_B(x)$ for the truncated exponential distribution on $[0, 1]$

Method	h_G^{opt}	IMSE$\times 10^3$ based on h_G^{opt}	95% C.I. of $\hat{h}_{CV}$	Average ISE$\times 10^3$ based on $\hat{h}_{CV}$
$\hat{F}(x)$	N/A	1.443	N/A	N/A
$\hat{F}_n(x)$	0.11	1.2194	0.12 (.04, .23)	1.2853
$\hat{F}_B^*(x)$	0.18	1.081	0.19 (.06, .40)	1.2599
Truncated $\hat{F}_B^*(x)$	0.18	1.078	0.19 (.06, .44)	1.2622
$\widetilde{F}_n(x)$	0.19	1.0693	0.20 (.07, .50)	1.2644
$\hat{F}_B(x)$	0.19	1.0324	0.20 (.10, .29)	1.2635

Table 2: The IMSE values of $\hat{F}(x)$, $\hat{F}_n(x)$, $\hat{F}_B^*(x)$, the truncated $\hat{F}_B^*(x)$, $\widetilde{F}_n(x)$ and $\hat{F}_B(x)$ for the truncated normal distribution on $[0, 1]$

Method	h_G^{opt}	IMSE$\times 10^3$ based on h_G^{opt}	95% C.I. of $\hat{h}_{CV}$	Average ISE$\times 10^3$ based on $\hat{h}_{CV}$
$\hat{F}(x)$	N/A	1.6148	N/A	N/A
$\hat{F}_n(x)$	0.16	1.29	0.17 (.06, .37)	1.3509
$\hat{F}_B^*(x)$	0.42	0.9218	0.42 (.06, .50)	1.0473
Truncated $\hat{F}_B^*(x)$	0.43	0.9021	0.42 (.06, .50)	1.0320
$\widetilde{F}_n(x)$	0.49	0.8990	0.43 (.06, .50)	1.0029
$\hat{F}_B(x)$	0.5	0.6844	0.44 (0.2, 0.5)	0.89

The biggest advantage of the boundary distribution kernel estimator $\hat{F}_B(x)$ over the other methods is the quality of estimation at the boundary. To see this, for each method we also plotted 10 estimates of the CDF based on their respective optimal bandwidths. The results are shown in Figures 2-5. For clarity, we only plotted the estimated CDF of each method at the left and right boundaries separately. These figures clearly show the boundary problem of the kernel distribution estimator: the estimates are systematically biased except Figure 4 of Distribution 3. They also show that it is quite common that the density boundary kernel distribution estimator may produce estimates which are either negative or larger than 1. The effect of truncation of the density boundary kernel distribution estimator can be clearly seen in the figures, especially in Figure 4. It can also been seen that, although the Tenreiro estimator does not have the boundary problem, its estimates at the boundary are rugged (as mentioned in Section 2), not as smooth as those from the kernel distribution estimator and the boundary distribution kernel estimator. As a matter of fact, we see that the Tenreiro estimator behaves very similarly to the empirical CDF in the boundary region, due to the fact that the bandwidth used in the boundary region ($h = x - a$) decreases to zero as $x \to a$. On the other hand, throughout all the plots we see that the boundary distribution kernel estimator provides smooth estimates for the distribution function in the boundary region while remaining to be free of boundary problem.

Recall that we observed for Distribution 3 that the IMSE of the boundary distribution kernel estimator is higher than that of the truncated density boundary kernel distribution estimator and the Tenreiro estimator. Figure 4 reveals that this is caused by the fact that the latter estimators take zero values for the points near zero. The zero estimates are the results of truncation for the truncated density boundary kernel distribution estimator and the result of the shrinking bandwidth ($h = x - a$) for the Tenreiro estimator. It is obvious that an estimator
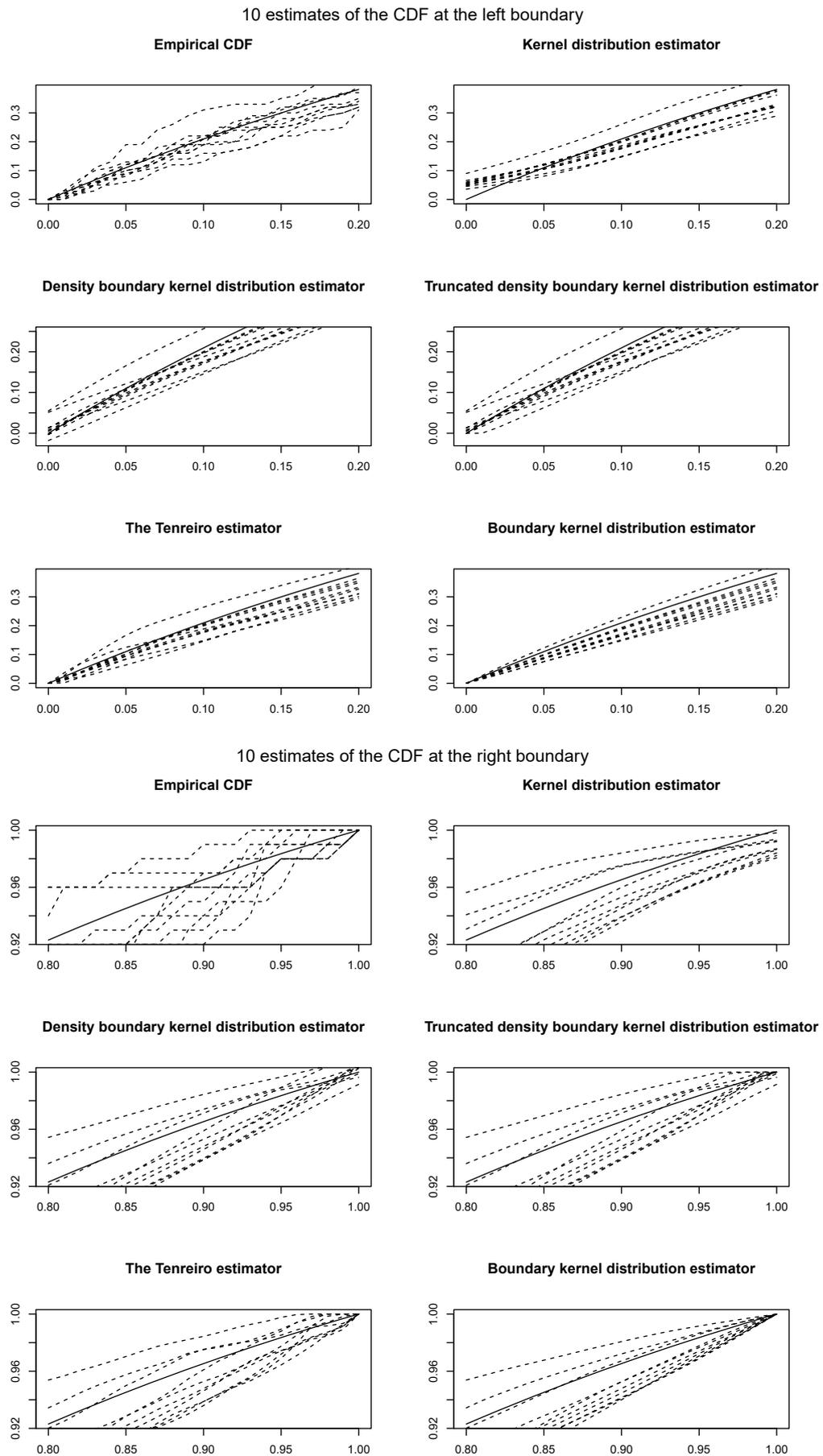
Figure 2: Performance of the different distribution estimators for Distribution 1

10 estimates of the CDF at the left boundary

Empirical CDF

Kernel distribution estimator

Density boundary kernel distribution estimator

Truncated density boundary kernel distribution estimator

The Tenreiro estimator

Boundary kernel distribution estimator

10 estimates of the CDF at the right boundary

Empirical CDF

Kernel distribution estimator

Density boundary kernel distribution estimator

Truncated density boundary kernel distribution estimator

The Tenreiro estimator

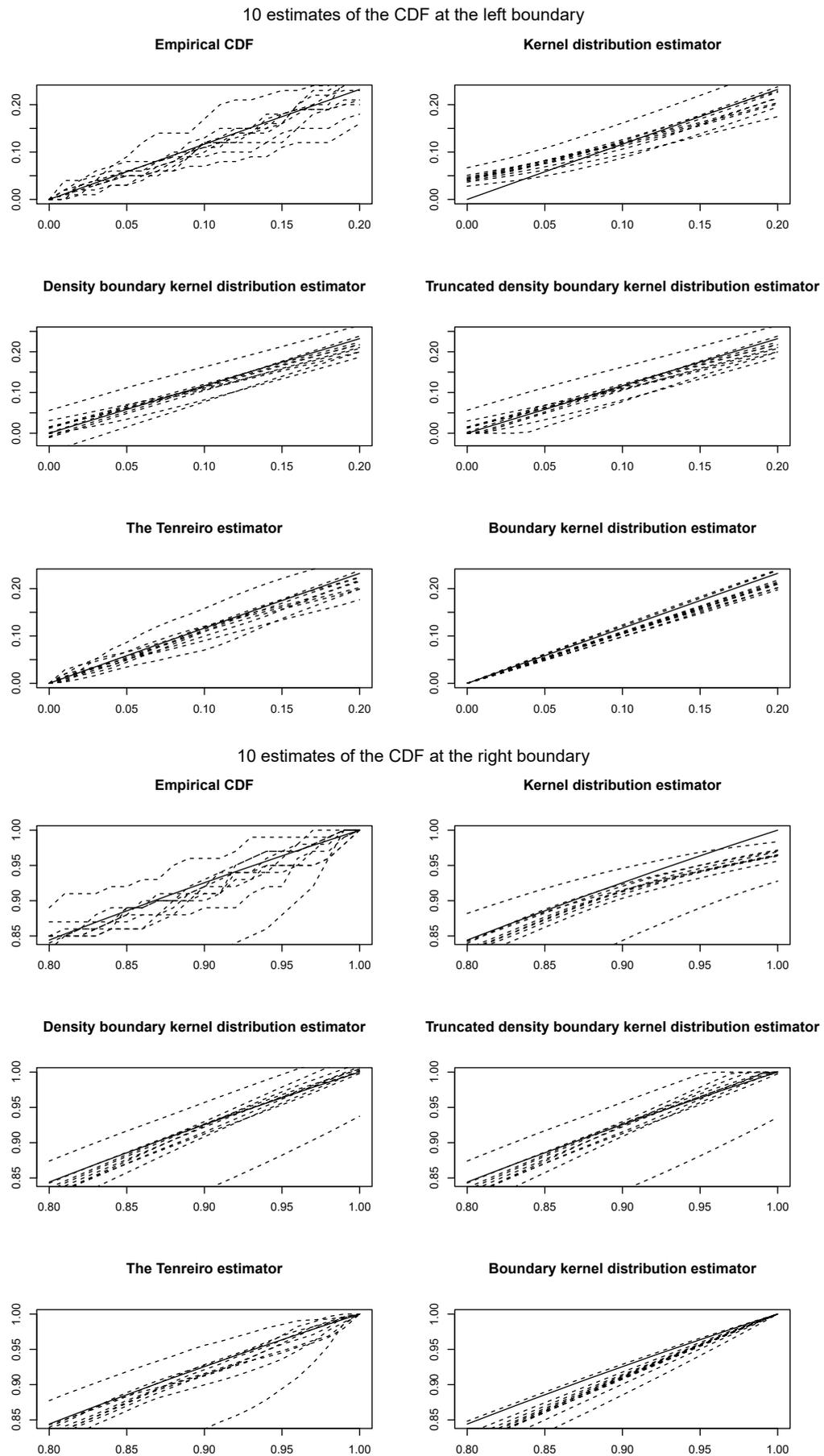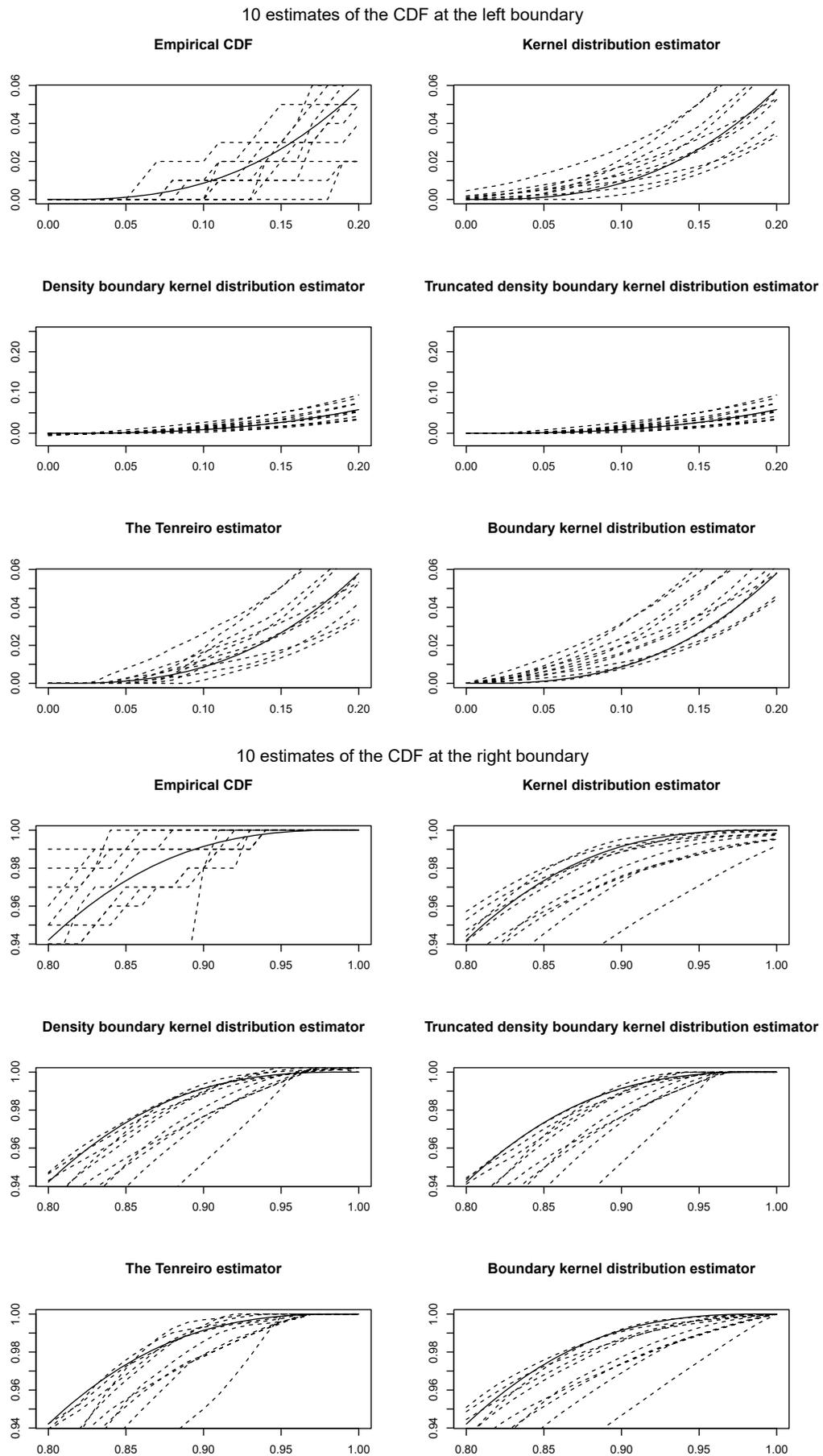**Boundary kernel distribution estimator**

Figure 3: Performance of the different distribution estimators for Distribution 2

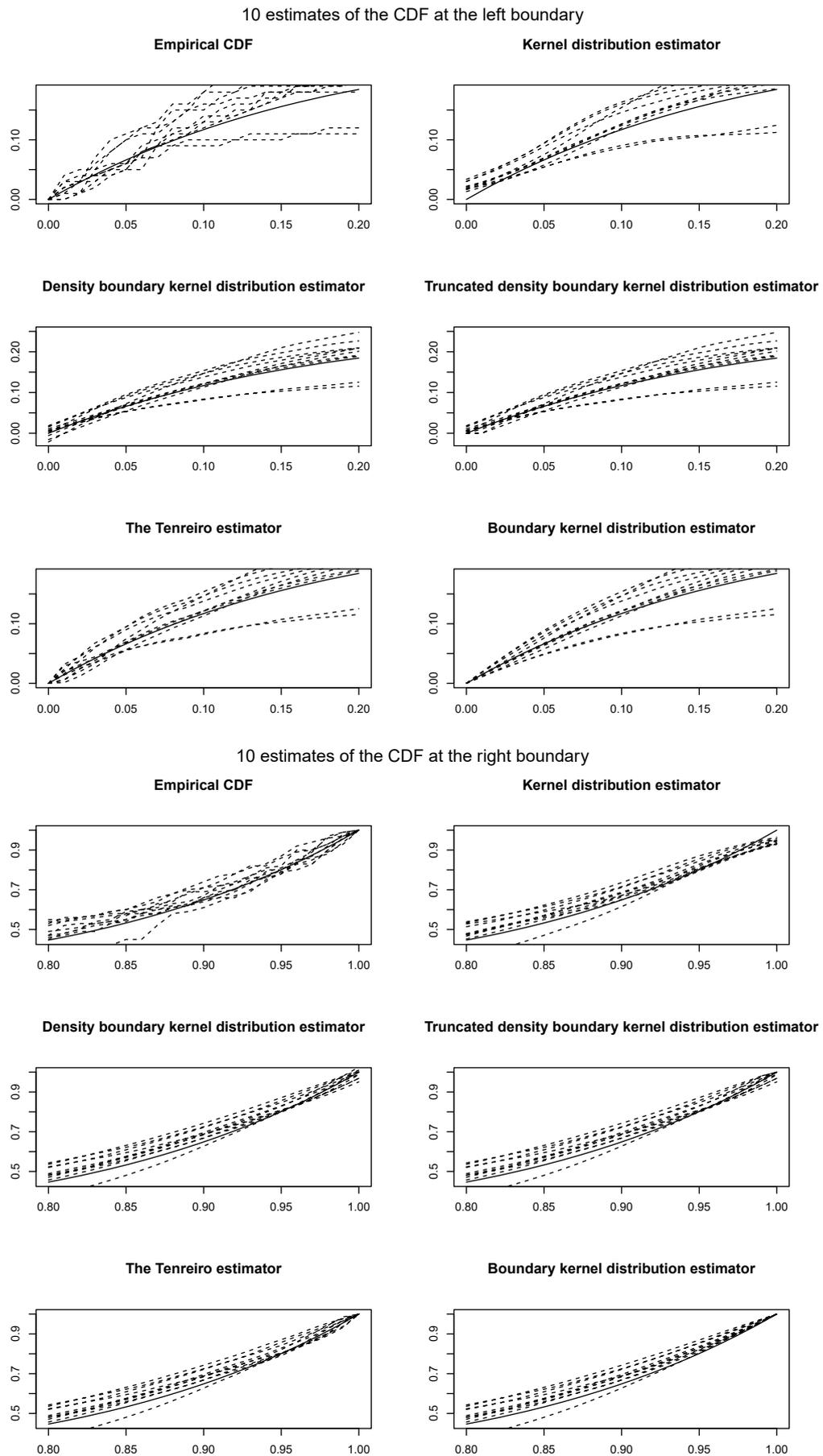Figure 4: Performance of the different distribution estimators for Distribution 3

10 estimates of the CDF at the left boundary

Empirical CDF

Kernel distribution estimator

Density boundary kernel distribution estimator

Truncated density boundary kernel distribution estimator

The Tenreiro estimator

Boundary kernel distribution estimator

10 estimates of the CDF at the right boundary

Empirical CDF

Kernel distribution estimator

Density boundary kernel distribution estimator

Truncated density boundary kernel distribution estimator

The Tenreiro estimator

Boundary kernel distribution estimator

Figure 5: Performance of the different distribution estimators for Distribution 4

Table 3: The IMSE values of $\hat{F}(x)$, $\hat{F}_n(x)$, $\hat{F}_B^*(x)$, the truncated $\hat{F}_B^*(x)$, $\widetilde{F}_n(x)$ and $\hat{F}_B(x)$ for the beta(α, β) distribution with $\alpha = 2, \beta = 2$

Method	h_G^{opt}	IMSE$\times 10^3$ based on h_G^{opt}	95% C.I. of $\hat{h}_{CV}$	Average ISE$\times 10^3$ based on $\hat{h}_{CV}$
$\hat{F}(x)$	N/A	1.2501	N/A	N/A
$\hat{F}_n(x)$	0.13	0.9655	0.13 (.08, .17)	1.0111
$\hat{F}_B^*(x)$	0.16	0.9175	0.21(.09, .49)	0.9940
Truncated $\hat{F}_B^*(x)$	0.17	0.9087	0.28(.09, .50)	0.9568
$\widetilde{F}_n(x)$	0.18	0.8996	0.29(.09, .50)	0.9934
$\hat{F}_B(x)$	0.15	0.9406	0.16 (.08, .26)	1.0477

Table 4: The IMSE values of $\hat{F}(x)$, $\hat{F}_n(x)$, $\hat{F}_B^*(x)$, the truncated $\hat{F}_B^*(x)$, $\widetilde{F}_n(x)$ and $\hat{F}_B(x)$ for the Beta mixture distribution (4)

Method	h_G^{opt}	IMSE$\times 10^3$ based on h_G^{opt}	95% C.I. of $\hat{h}_{CV}$	Average ISE$\times 10^3$ based on $\hat{h}_{CV}$
$\hat{F}(x)$	N/A	1.8068	N/A	N/A
$\hat{F}_n(x)$	0.06	1.6981	0.06 (.01, .13)	1.7350
$\hat{F}_B^*(x)$	0.09	1.6200	0.09 (.03, .14)	1.6826
Truncated $\hat{F}_B^*(x)$	0.09	1.6186	0.09 (.03, .14)	1.6814
$\widetilde{F}_n(x)$	0.09	1.6098	0.09 (.03, .18)	1.6778
$\hat{F}_B(x)$	0.09	1.5927	0.09 (.05, .12)	1.6760

which takes zero value in the neighborhood of $x = a$ may have an advantage on the value of IMSE since $F(a) = 0$. However, such an estimator does not provide useful information for the points near the end of the support since the true distribution function is not zero at these points.

3.2. Estimation of the optimal global bandwidth

In order to apply the proposed kernel distribution estimators in practical situations, we need to develop a method to estimate the optimal global bandwidth. A number of methods have been proposed in the literature to choose the bandwidth for the kernel distribution estimator in the infinite support case. The simplest one is the reference distribution method in which the bandwidth is estimated by replacing the true distribution by a reference distribution such as the normal distribution. Other methods for choosing the bandwidth include the "leave-one-out" method of Sarda (1993), the "plug-in" method of Altman and Leger (1994), and the cross-validation (CV) method of Bowman, Hall, and Prvan (1998).

In the following, we discuss how to use the CV method to select the bandwidth in kernel distribution estimation for densities with finite support. Although we have discussed both local and global bandwidths for kernel distribution estimation, we will only focus on the estimation of the optimal global bandwidth due to the fact that the bandwidth selector developed based on the global optimal bandwidth is more stable than that based on the local optimal bandwidth. We follow the idea of Bowman *et al.* (1998) to use the CV method for estimating the bandwidth for the kernel distribution estimators when the density has compact support $[a, b]$. Define

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \int_{a}^{b} \{I[x_i \leq x] - \tilde{F}_{-i}(x)\}^2 dx \qquad (34)$$

where $\tilde{F}_{-i}(x)$ is a kernel distribution estimator (boundary corrected or not) using the data without observation x_i. Then, following the same lines as those (page 801-803) of Bowman *et al.* (1998), one can show that, by ignoring an unknown constant factor, $CV(h)$ is an unbiased estimator of the true IMSE for sample size $n - 1$.

To evaluate the performance of the bandwidth selector $CV(h)$, for each method and each sample we obtained $\hat{h}_{CV}$, the estimated optimal global bandwidth, by minimizing (49) on a grid of possible h values between $[0, 1/2]$. The mean, the 95% confidence interval of these bandwidths are also reported in Column 4 of Tables 1-4. These results show that the CV method based on (34) provides satisfactory estimates of the optimal global bandwidth for the kernel distribution estimator $\hat{F}(x)$. For the boundary distribution kernel estimator $\hat{F}_B(x)$ and the Tenreiro estimator $\tilde{F}_n(x)$, the bottom two row of Table 2 shows that the estimated optimal global bandwidth $\hat{h}_{CV}$ are smaller than the optimal global bandwidths. This is caused by the fact that the true optimal bandwidth for $\hat{F}_B(x)$ are too close to 0.50 (.49 and .50 respectively), the right endpoint of the interval $[0, 1/2]$, in which we search for the optimal bandwidth. We also conducted the simulations (not reported here) for truncated normal densities with support $[0, 2]$, and found that this problem had disappeared for both methods. For the density boundary kernel distribution estimator, its truncated version and the Tenreiro estimator, Table 3 shows that the estimated optimal global bandwidths are significantly larger than the true optimal global bandwidths. This indicates the instability of (34) as a bandwidth selector for these three estimators.

In order to see the performance of the estimators w.r.t. their respective estimated optimal bandwidths, we also calculated their average ISE values. The results (multiplied by 1000) are reported in Column 5 of Tables 1-4. As expected, the IMSE values reported in Column 5 are greater than those in Column 3 due to the use of estimated optimal bandwidths. Column 5 also confirms the findings from Column 3 that the boundary kernel estimator in general has the best performance among all the estimators we have discussed in this paper.

3.3. A real data example

To examine the performance of different methods in real applications, we applied the methods we have discussed in this paper to the Massachusetts auto bodily injury liability data, which are provided in Rempala and Derrig (2005). The data consist of outpatient medical provider's total billings on a sample of 348 auto bodily injury liability claims closed in Massachusetts during 2001. The data range from 0.045 to 50.00 in thousands of dollars. We randomly divided the data into the training sample (50% of the data) to fit the kernel estimators and the test sample (the remaining 50% of the data). The histogram in Figure 6 shows that the majority of claims are small claims (less than 5 thousand dollars), indicating that the value of the density function may be larger than zero at zero-the left endpoint of the support. Hence, boundary correction around zero is necessary. In our analysis, we first obtained the bandwidths from the training sample using the CV method (49), and they are 0.40 for the kernel distribution estimator, 0.48 for the density boundary kernel distribution estimator, the

truncated density boundary kernel distribution estimator and the Tenreiro estimator, and .49 for the boundary distribution kernel estimator, respectively.
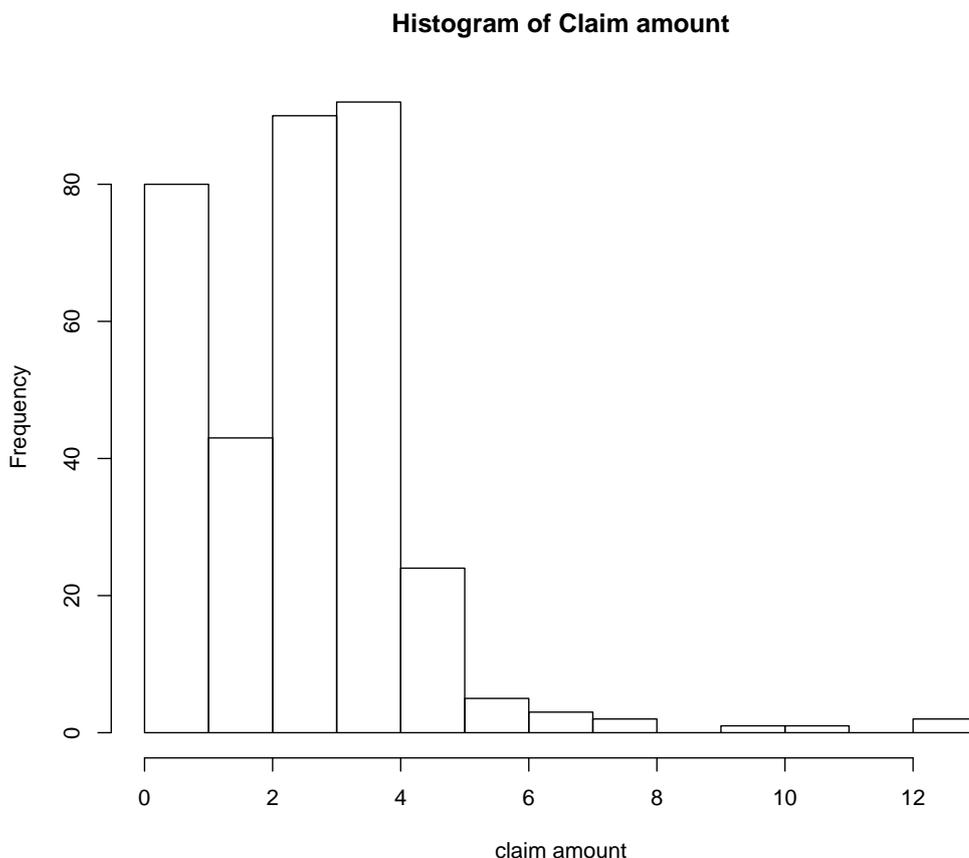
Histogram of Claim amount

Figure 6: Histogram of claim amounts from the training sample

In Figure 7, we plotted the estimated exceedance probabilities for each method for claim amounts $0 \leq x \leq .5$ in the test sample, and these probabilities were compared with the proportions of claims exceeding x from the test sample (the solid line). The exceedance probability can be useful in premium calculation in the insurance industry. Figure 7 clearly shows the boundary problem of the kernel distribution estimator $\hat{F}(x)$ (the short dashed line)-it down-estimates the exceedance probability for claim amounts less than \$200. The density boundary kernel distribution estimator and its truncated version (the dot-dash line) are the same line (i.e., no need for truncation). Since the value of these two estimator at $x = 0$ is less than 1, this shows that they both down-estimate the exceedance probability for small claims. On the other hand, both the Tenreiro estimator (the long dashed line) and the boundary distribution kernel estimator (the dotted line) take the value 1 at $x = 0$. It can also been seen the Tenreiro estimator is less stable in the boundary region than the boundary distribution kernel estimator. The flat region of the Tenreiro estimator ($\tilde{F}_n(x) = 1$) for the points near at $x = 0$ does not provide useful information for insurance claims falling into this region.

4. Conclusion and discussion

Estimation of distribution function has found numerous applications in econometrics, climatology, hydraulics, among others. One of the major methods in estimating the distribution function is the kernel method. In this paper, we have provided an in-depth study of the boundary problem of distribution kernel estimator when the density function has compact support. In order to eliminate the problem of the kernel distribution estimator, we have
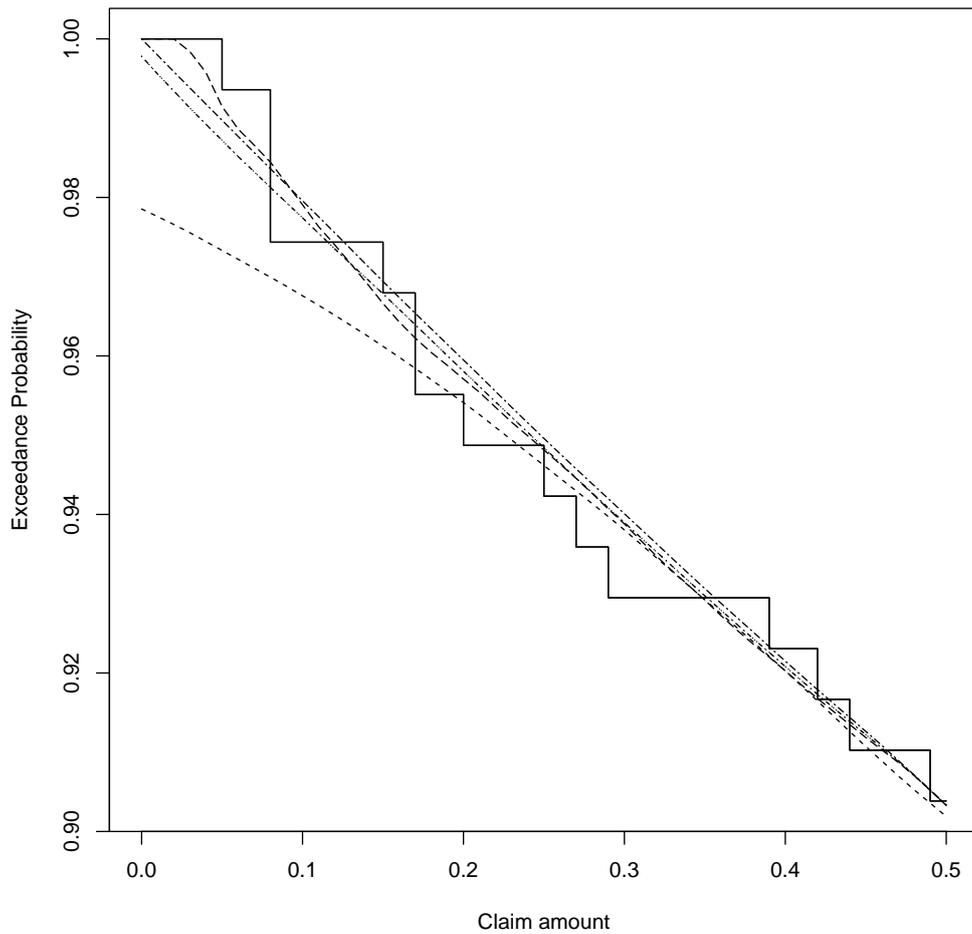
Figure 7: Exceedance probability plot from the test sample.The proportions of claims exceeding x(claim amount)-the solid line, $\hat{F}_n(x)$-the short dashed line, $\hat{F}_B^*(x)$ and its truncated version-the dot-dash line, $\tilde{F}_n(x)$-the long dashed line, and $\hat{F}_B(x)$-the dotted line.

proposed a boundary distribution kernel estimator. We have shown that the boundary distribution kernel estimator uses the available information more efficiently for estimating the CDF at the boundary point $x = a + ch$ (or $x = b - ch$) by using the data in the interval $[a, a + h)$ for the left boundary region (or $(b - h, b]$ for the right boundary region), compared with $[a, x + ch]$ (or $[b - ch, b]$) for the Tenreiro estimator. Numerical comparisons also show that the boundary distribution kernel estimator, as well as the optimal global bandwidth based on it, in general has the best performance in the boundary region among all the estimators we have considered in this paper.

Acknowledgements

References

Altman N, Leger C (1994). "Bandwidth Selection for Kernel Distribution Function Estimation." *Journal of Statistical Planning and Inference*, **46**(2), 195–214.

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

Bowman A, Hall P, Prvan T (1998). "Bandwidth Selection for the Smoothing of Distribution functions." *Biometrika*, **85**(4), 799–808.

Gasser T, Műller HG (1979). *Kernel Estimation of Regression Functions in Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics 757, eds. T. Gasser and M. Rosenblatt.* Springer Berlin Heidelberg.

Gasser T, Műller HG, Mammitzsch V (1985). "Kernels for Nonparametric Curve Estimation." *Journal of the Royal Statistical Society, Series B*, **47**(2), 238–256.

Jones MC (1990). "The Performance of Kernel Density Functions in Kernel Distribution Estimation." *Statistics and Probability Letters*, **9**(2), 129–132.

Koláček J, Karunamuni RJ (2009). "On Boundary Correction in Kernel Estimation of ROC Curves." *Austrian Journal of Statistics*, **38**(1), 17–32.

Műller HG (1991). "Smooth Optimum Kernel Estimators near Endpoints." *Biometrika*, **78**(3), 521–530.

Rempala GA, Derrig RA (2005). "Modeling Hidden Exposures in Claim Severity via the EM Algorithm." *North American Actuarial Journal*, **9**(2), 108–128.

Sarda P (1993). "Smoothing Parameter Selection for Smooth Distribution Functions." *Journal of Statistical Planning and Inference*, **35**(1), 65–75.

Swanepoel JWH (1988). "Mean Integrated Squared Error Properties and Optimal Kernels when Estimating a Distribution Function." *Communications in Statistics*, **17**(11), 3785–3799.

Tenreiro C (2013). "Boundary Kernels for Distribution Estimation." *REVSTAT-Statistical Journal*, **11**(2), 169–190.

Zhang S, Karunamuni RJ (1998). "On Kernel Density Estimation near Endpoints." *Journal of Statistical Planning Inference*, **70**(2), 301–316.

Zhang S, Karunamuni RJ (2000). "On Nonparametric Density Estimation at the Boundary."
    Journal of Nonparametric Statistics, **12**(2), 197–221.

Appendix

Proof of (14, 15). We will only prove (15). The proof of (14) can be similarly obtained. Since $X_1, \ldots, X_n$ are *i.i.d.*, at $x = b - ch$, we have

$$
\begin{aligned}
Bias(\hat{F}_n(x)) &= E\hat{F}_n(x) - F(x) \\
&= E[\frac{1}{n}\sum_{i=1}^{n} K(\frac{x - X_i}{h})] - F(x) \\
&= EK(\frac{x - X_i}{h}) - F(x) \\
&= \int_a^b K(\frac{x - t}{h})dF(t) - F(x) \\
&= K(\frac{x - t}{h})F(t)|_a^b + \frac{1}{n}\int_a^b F(t)k(\frac{x - t}{h})dt - F(x) \\
&= K(-c) + \frac{1}{n}\int_a^b F(t)k(\frac{x - t}{h})dt - F(x).
\end{aligned}
\tag{A1}
$$

Let $y = (x - t)/h$, and expand $F(x - hy)$ around x, we have

$$
\begin{aligned}
\frac{1}{n}\int_a^b F(t)k(\frac{x - t}{h})dt &= \int_{-c}^1 k(y)F(x - hy)dy \\
&= \int_{-c}^1 k(y)[F(x) - hyf(x) + \frac{1}{2}h^2y^2f^{(1)}(x)]dy + o(h^2) \\
&= F(x)\int_{-c}^1 k(y)dy - hf(x)\int_{-c}^1 yk(y)dy \\
&\quad + \frac{1}{2}h^2f^{(1)}(x)\int_{-c}^1 y^2k(y)dydy + o(h^2).
\end{aligned}
\tag{A2}
$$

Plugging (A2) into (A1), we have

$$
\begin{aligned}
Bias(\hat{F}_n(x)) =&K(-c) + F(x)(\int_{-c}^1 k(y)dy - 1) - hf(x)\int_{-c}^1 yk(y)dy \\
&+ \frac{1}{2}h^2f^{(1)}(x)\int_{-1}^c y^2k(y)dy + o(h^2).
\end{aligned}
\tag{A3}
$$

Note that for $x = b - ch$

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

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

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

Plugging these into (A3), we have

$$\begin{aligned}
Bias(\hat{F}_n(x)) =& K(-c) + (\int_{-c}^{1} k(y)d(y) - 1) - chf(b-)(\int_{-c}^{1} k(y)d(y) - 1) \\
& + \frac{1}{2}c^2h^2 f^{(1)}(b-)(\int_{-c}^{1} k(y)d(y) - 1) \\
& - h(f(b-) - chf^{(1)}(b-)) \int_{-c}^{1} yk(y)d(y) \\
& + \frac{h^2}{2} f^{(1)}(b-) \int_{-c}^{1} y^2 k(y)d(y) \\
=& - hf(b-)[c(\int_{-c}^{1} k(y)dy - 1) + \int_{-c}^{1} yk(y)dy] \\
& + \frac{h^2}{2} f^{(1)}(b-)[c^2(\int_{-c}^{1} k(y)dy - 1) + 2c \int_{-c}^{1} yk(y)dy + \int_{-c}^{1} y^2 k(y)dy] + o(h^2).
\end{aligned}$$

This completes the proof of (15). $\qquad\qquad\square$

Proof of Theorem 1. We will only prove (i), (ii) and (28). (27) and (29) are direct results of (14) and (15). (30) can be proved along the same lines as those used in proving (28)

Note that for $x = a$, $\dfrac{x - X_i}{h} < 0$. The construction of $K_c(t)$ (see discussion circa (18)) shows that $K_0(t) = 0$ for $t \leq 0$. Hence, $\hat{F}_B(a) = \dfrac{1}{n} \sum_{i=1}^{n} K_c(\dfrac{x - X_i}{h}) = 0$. $\hat{F}_B(b) = 1$ can be similarly proved.

To show $\hat{F}_B(x)$ is non-decreasing; it is enough to show that it is non-decreasing in the boundary regions. We will only prove this for the left boundary region.

For any $a \leq x_1 \leq x_2 \leq a + h$, note that

$$\begin{aligned}
\hat{F}_B(x_2) - \hat{F}_B(x_1) =& \frac{1}{n} \sum_{i=1}^{n} K_{c_2}(\frac{x_2 - X_i}{h}) - \frac{1}{n} \sum_{i=1}^{n} K_{c_1}(\frac{x_1 - X_i}{h}) \\
=& \frac{1}{n} \sum_{i=1}^{n} [K_{c_2}(\frac{x_2 - X_i}{h}) - K_{c_1}(\frac{x_1 - X_i}{h})],
\end{aligned} \tag{A4}$$

where $c_i = (x_i - X_i)/h, i = 1, 2$. (A4) shows that it is enough to show

$$\sum_{i=1}^{n} K_{c_2}(\frac{x_2 - X_i}{h}) - \sum_{i=1}^{n} K_{c_1}(\frac{x_1 - X_i}{h}) \geq 0 \text{ for } i = 1, 2, \ldots, n.$$

Without loss of generality, it is enough to show

$$K_{c_2}(\frac{x_2 - X_i}{h}) - K_{c_1}(\frac{x_1 - X_i}{h}) \geq 0.$$

Note that

$$\begin{aligned}
& K_{c_2}(\frac{x_2 - X_i}{h}) - K_{c_1}(\frac{x_1 - X_i}{h}) \\
=& \int_{-1}^{\frac{x_2 - X_i}{h}} k_{\frac{x_2 - X_i}{h}}(y)dy - \int_{-1}^{\frac{x_1 - X_i}{h}} k_{\frac{x_1 - X_i}{h}}(y)dy \\
=& \int_{\frac{x_1 - X_i}{h}}^{\frac{x_2 - X_i}{h}} k_{\frac{x_2 - X_i}{h}}(y)dy + \int_{-1}^{\frac{x_1 - X_i}{h}} k_{\frac{x_2 - X_i}{h}}(y)dy - \int_{-1}^{\frac{x_1 - X_i}{h}} k_{\frac{x_1 - X_i}{h}}(y)dy \\
\geq& \int_{-1}^{\frac{x_1 - X_i}{h}} \left(k_{\frac{x_2 - X_i}{h}}(y) - k_{\frac{x_1 - X_i}{h}}(y) \right) dy,
\end{aligned} \tag{A5}$$

since $k_{\frac{x_2-X_i}{h}}(y) \geq 0$. The last inequality in (A5) show that the proof is complete if we can show that $k_{\frac{x_2-X_i}{h}}(y) - k_{\frac{x_1-X_i}{h}}(y)$ for $-1 \leq y \leq \frac{x_1-X_i}{h}$, i.e., if we can show that $k_c(y)$ is non-decreasing as a function of c, for each $-1 \leq y \leq c$. From (18), we see that we only need to show $ck(y)/\int_{-1}^{c}(c-s)k(s)ds$ is non-decreasing as a function of c, for $-1 \leq y \leq c$, which directly follows from the fact that

$$\frac{d}{dc}[ck(y)/\int_{-1}^{c}(c-s)k(s)ds] = -\int_{-1}^{c}sk(s)ds \geq 0.$$

To show (28), first we realize that

$$Var(\hat{F}_n(x)) = \frac{1}{n}Var\left(K_c(\frac{x-X_i}{h})\right)$$
$$= \frac{1}{n}\left[EK_c^2(\frac{x-X_i}{h}) - (EK_c(\frac{x-X_i}{h}))^2\right].$$

Similar to the proof of the bias term, we can show that

$$EK_c^2(\frac{x-X_i}{h}) = 2\{F(x)\int_{-1}^{c}k_c(y)K_c(y)dy - hf(x)\int_{-1}^{c}yk_c(y)K_c(y)dy\} + o(h) \qquad \text{(A6)}$$

and

$$EK_c(\frac{x-X_i}{h}) = F(x)\int_{-1}^{c}k_c(y)dy - hf(x)\int_{-1}^{c}yk_c(y)dy + o(h) \qquad \text{(A7)}$$

Plugging the facts $F(x) = F(a+) + chf(a+) + o(h)$, $f(x) = f(a+) + o(1)$ and $F(a+) = 0$ into (A6) and (A7), we have

$$EK_c(\frac{x-X_i}{h}) = O(h)$$

and

$$EK_c^2\left(\frac{x-X_i}{h}\right) = 2hf(a+)\{c\int_{-1}^{c}k_c(y)K_c(y)dy - \int_{-1}^{c}yk_c(y)K_c(y)dy\}$$
$$= 2hf(a+)\int_{-1}^{c}(c-y)k_c(y)K_c(y)dy + o(h).$$

Hence,

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

This completes the proof of (28). $\square$

Affiliation:

Shunpu Zhang
Department of Statistics
University of Central Florida
Orlando,FL 32816-2370, USA
Telephone: 407-823-1566
Fax: 407-823-3930
E-mail: Shunpu.Zhang@ucf.edu

Zhong Li & Zhiying Zhang
College of Science
Zhejiang Sci-Tech University
Hangzhou, China
E-mail: lizhong@zstu.edu.cn, zzyhz@126.com