Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 17–30

A NEW CLASS OF BOUNDARY KERNELS FOR DISTRIBUTION FUNCTION ESTIMATION

CARLOS TENREIRO

ABSTRACT: In this work we introduce a new class of boundary kernels for distribution function estimation which shows itself to be especially performing when the classical kernel distribution function estimator suffers from severe boundary problems.

KEYWORDS: Distribution function estimation; kernel estimator; boundary kernels.
AMS SUBJECT CLASSIFICATION (2010): 62G05, 62G20.

1. Introduction

Given $X_1, \ldots, X_n$ independent copies of an absolutely continuous real random variable with unknown density and distribution functions f and F, respectively, the classical kernel estimator of F introduced by authors such as Tiago de Oliveira (1963), Nadaraya (1964) or Watson and Leadbetter (1964), is defined, for $x \in \mathbb{R}$, by

$$\bar{F}_{nh}(x) = \frac{1}{n} \sum_{i=1}^{n} \bar{K}\left(\frac{x - X_i}{h}\right), \tag{1}$$

where, for $u \in \mathbb{R}$,

$$\bar{K}(u) = \int_{-\infty}^{u} K(v)dv,$$

with K a kernel on $\mathbb{R}$, that is, a bounded and symmetric probability density function with support $[-1, 1]$ and $h = h_n$ a sequence of strictly positive real numbers converging to zero when n goes to infinity. For some recent references on this classical estimator see Giné and Nickl (2009), Chacón and Rodríguez-Casal (2010), Mason and Swanepoel (2011) and Chacón, Monfort and Tenreiro (2014).

If the support of f is known to be the finite interval $[a, b]$, that is, $a = \inf\{x : F(x) > 0\} > -\infty$ and $b = \sup\{x : F(x) < 1\} < +\infty$, the previous kernel estimator suffers from boundary problems if $F'_+(a) \neq 0$ or $F'_-(b) \neq 0$. This

Received June 12, 2017.

question is addressed in Tenreiro (2013) who extend to the distribution function estimation framework the approach followed in nonparametric regression and density function estimation by authors such as Gasser and Müller (1979), Rice (1984), Gasser et al. (1985) and Müller (1991). Especially, the author considers the boundary modified kernel distribution function estimator given by

$$\tilde{F}_{nh}(x) = \begin{cases} 0, & x \leq a \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \bar{K}_{x,h}\left(\dfrac{x - X_i}{h}\right), & a < x < b \\ 1, & x \geq b, \end{cases} \tag{2}$$

where $0 < h \leq (b-a)/2$ and

$$\bar{K}_{x,h}(u) = \begin{cases} \bar{K}^L(u;(x-a)/h), & a < x < a+h \\ \bar{K}(u), & a+h \leq x \leq b-h \\ \bar{K}^R(u;(b-x)/h), & b-h < x < b, \end{cases}$$

with

$$\bar{K}^L(u;\alpha) = \int_{-\infty}^{u} K^L(v;\alpha)dv \quad \text{and} \quad \bar{K}^R(u;\alpha) = 1 - \int_{u}^{+\infty} K^R(v;\alpha)dv,$$

where $K^L(\cdot;\alpha)$ and $K^R(\cdot;\alpha)$ are, respectively, left and right boundary kernels for $\alpha \in]0,1[$, that is, their supports are contained in the intervals $[-1,\alpha]$ and $[-\alpha,1]$, respectively, and $|\mu_{0,\ell}|(\alpha) = \int |K^\ell(u;\alpha)|\,du < \infty$ for all $\alpha \in]0,1[$ and $\ell = L, R$ (here and bellow integrals without integration limits are meant over the whole real line).

For ease of presentation, from now on we assume that the right boundary kernel K^R is given by $K^R(u;\alpha) = K^L(-u;\alpha)$, the reason why only the left boundary kernel is mentioned from now on. By assuming that $K^L(\cdot;\alpha)$ is a second order kernel, that is,

$$\mu_{0,L}(\alpha) = 1,\ \mu_{1,L}(\alpha) = 0 \text{ and } \mu_{2,L}(\alpha) \neq 0,\ \text{for all } \alpha \in]0,1[, \tag{3}$$

where we denote

$$\mu_{k,L}(\alpha) = \int u^k K^L(u;\alpha)\,du,\ \text{for } k \in \mathbb{N},$$

Tenreiro (2013) shows that the previous estimator is free of boundary problems and that the theoretical advantage of using boundary kernels is compatible with the natural property of getting a proper distribution function estimate. In fact, it is easy to see that the kernel distribution function estimator based on each one of the second order left boundary kernels

$$K_1^L(u;\alpha) = (2\bar{K}(\alpha)-1)^{-1}K(u)I(-\alpha \le u \le \alpha), \tag{4}$$

where we assume that K is such that $\int_0^\alpha K(u)du > 0$ for all $\alpha > 0$, and

$$K_2^L(u;\alpha) = K(u/\alpha)/\alpha, \tag{5}$$

is, with probability one, a continuous probability distribution function (see Tenreiro, 2013, Examples 2.2 and 2.3). Additionally, it is shown that the Chung-Smirnov law of iterated logarithm is valid for the new estimator, and an asymptotic expansion for its mean integrated squared error is presented, from which the choice of h is discussed (see Tenreiro, 2013, Theorems 3.2, 4.1 and 4.2).

A careful analysis of the asymptotic expansions presented in Tenreiro (2013, p. 171, 178) for the local bias and the integrated squared bias of estimator (1), suggests that the previous properties may still be valid for all the boundary kernels satisfying the less restricted condition

$$\alpha\,(1-\mu_{0,L}(\alpha)) + \mu_{1,L}(\alpha) = 0, \text{ for all } \alpha \in]0,1[, \tag{6}$$

which is in particular fulfilled by the left boundary kernel

$$K_3^L(u;\alpha) = \alpha K(u)I(-1 \le u \le \alpha)\big/(\alpha\mu_{0,\alpha}(K) - \mu_{1,\alpha}(K)), \tag{7}$$

where we denote $\mu_{k,\alpha}(K) = \int_{-1}^\alpha u^k K(u)\,du$, for $k \in \mathbb{N}$ (see Figure 1). This observation motivated the present note, which is organized as follows. In Section 2 we describe the global and boundary behaviour of $\tilde{F}_{nh}$ to the broad class of boundary kernels satisfying assumption (6). In Section 3 we refine the previous analysis by describing the asymptotic behaviour of the bias and variance of $\tilde{F}_{nh}(x)$ at the extreme boundary region, that is, for x taking the form $x = a + \alpha h$, where $\alpha = \alpha_n$ converges to zero as n tends to infinity. This local analysis enables us to identify different orders of convergence for the mean square error of the estimators associates to boundary kernels K_1^L and K_2^L and to boundary kernel K_3^L, which indicate that this latter boundary kernel can be especially performing when the classical kernel estimator suffers from severe boundary problems. In Section 4 we present some exact finite sample comparisons between the estimators based on the previous boundary kernels. The proofs of all results are deferred to Section 5.

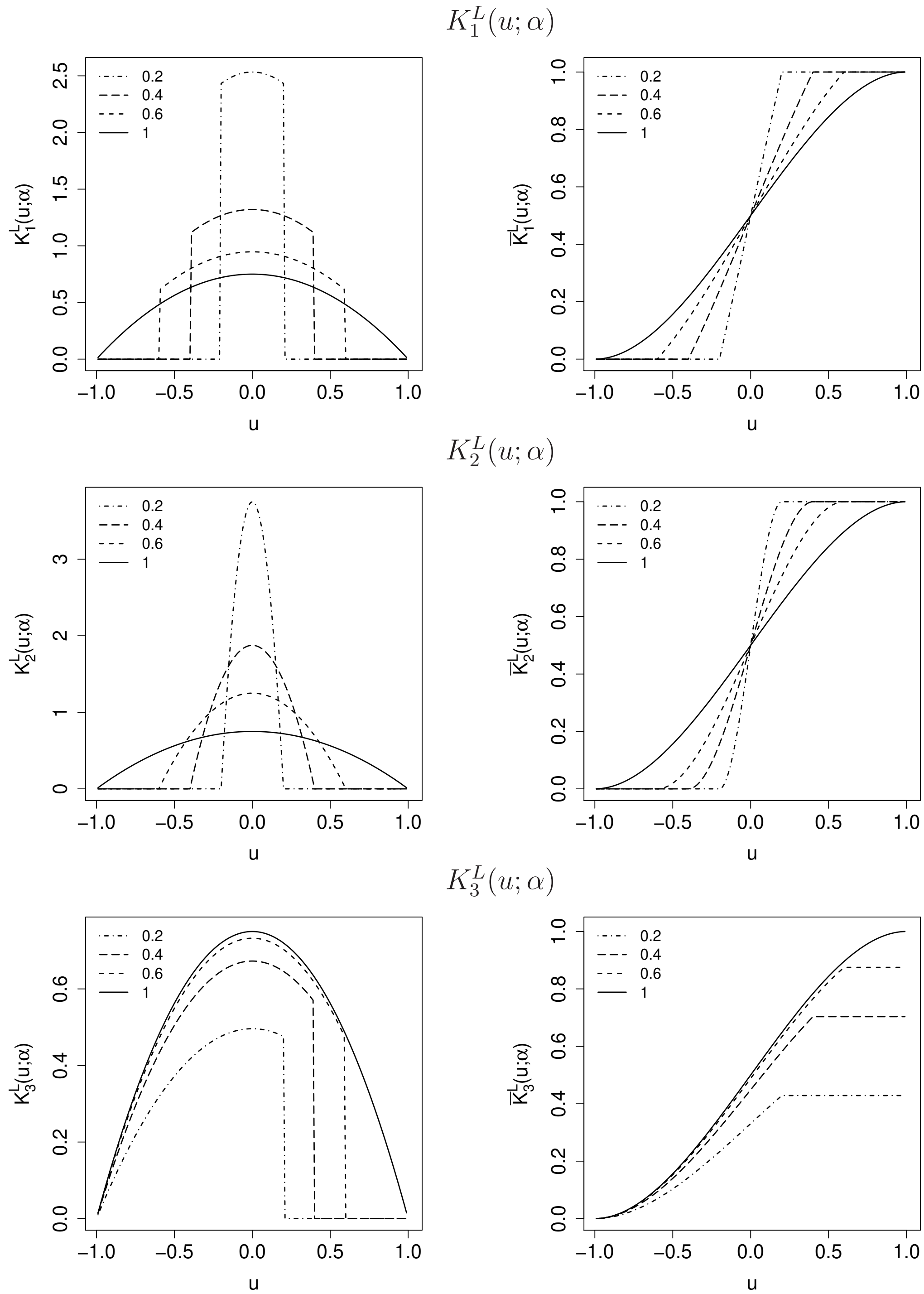

FIGURE 1. *Left boundary kernels* $K_q^L(u;\alpha)$ *(left column) and* $\bar{K}_q^L(u;\alpha)$ *(right column) for* $q = 1, 2, 3$ *and* $\alpha = 0.2, 0.4, 0.6, 1$*, where* K *is the Epanechnikov kernel* $K(t) = \frac{3}{4}(1 - t^2)I(|t| \leq 1)$.

2. Global and boundary behaviour

In this section we describe the global and boundary behaviour of the boundary modified kernel distribution function estimator $\tilde{F}_{nh}$ defined by (2). As mentioned before, for each one of the families of boundary kernels (4) and (5), $\tilde{F}_{nh}$ is, under general conditions on K, a continuous probability distribution function (with probability one). It is not hard to see that this is also true for the new family of boundary kernels (7) whenever K is continuous on $]-1,1[$.

2.1. Global behaviour. A classical measure of a distribution function estimator performance is the supremum distance between such an estimator and the underlying distribution function F. Next we extend Theorems 3.1 and 3.2 of Tenreiro (2013) by establishing the almost complete uniform convergence and the Chung-Smirnov law of iterated logarithm for kernel estimator (2). These properties have been first obtained for estimator (1) by Nadaraya (1964), Winter (1973, 1979) and Yamato (1973). We denote by $||\cdot||$ the supremum norm.

Theorem 1. *If $K^L(u;\alpha)$ satisfies*

$$\sup_{\alpha\in\,]0,1[} |\mu_{0,L}|(\alpha) < \infty, \tag{8}$$

we have

$$||\tilde{F}_{nh} - F|| \to 0 \quad \textit{almost completely.}$$

Additionally, if F is Lipschitz on $[a,b]$ and

$$(n/\log\log n)^{1/2} h \to 0, \tag{9}$$

then $\tilde{F}_{nh}$ has the Chung-Smirnov property, i.e.,

$$\limsup_{n\to\infty}\, (2n/\log\log n)^{1/2} ||\tilde{F}_{nh} - F|| \leq 1 \;\; \textit{almost surely.}$$

The same is true under the less restrictive condition

$$(n/\log\log n)^{1/2} h^2 \to 0, \tag{10}$$

whenever K^L satisfies (6) and F' is Lipschitz on $[a,b]$.

If the restriction of F to the interval $[a,b]$ is twice continuously differentiable, it can be proved that the expansion of the mean integrated squared error of the estimator $\tilde{F}_{nh}$ given in Theorem 2.4 of Tenreiro (2013) is also valid for the boundary modified kernel estimator (2) when the left boundary kernel satisfies condition (6) with $\int_0^1 |\mu_{0,L}|(\alpha)^2 d\alpha < \infty$. The asymptotically optimal bandwidth, in the sense of minimising the main terms of that expansion, is given by

$$h_0 = \delta(K) \left(\int F''(x)^2 dx \right)^{-1/3} n^{-1/3}, \tag{11}$$

where $\delta(K) = \left(\int uB(u)\, du\right)^{1/3} \left(\int u^2 K(u) du\right)^{-2/3}$ and $B(u) = 2\bar{K}(u)K(u)$. This optimal bandwidth satisfies condition (10) but not condition (9).

2.2. Boundary behaviour. In the next result we present asymptotic expansions for the bias and variance of $\tilde{F}_{nh}(x)$ with x in the boundary support region. They extend the corresponding expansions presented in Tenreiro (2013, p. 174) for second order boundary kernels. We will restrict our attention to the left boundary region $]a, a + h[$, but similar results are valid for the right boundary region $]b - h, b[$.

Theorem 2. *If $K^L(u; \alpha)$ satisfies conditions (6) and (8), and the restriction of F to the interval $[a, b]$ is twice continuously differentiable, we have:*

a)

$$\sup_{x \in]a,a+h[} \left| \mathrm{E}\tilde{F}_{nh}(x) - F(x) - \frac{h^2}{2} F''(x)\mu_L\big((x - a)/h\big) \right| = o(h^2),$$

where

$$\mu_L(\alpha) = \mu_{2,L}(\alpha) - \alpha\mu_{1,L}(\alpha),\ \alpha \in]0, 1[.$$

b)

$$\sup_{x \in]a,a+h[} \Bigg| \mathrm{Var}\tilde{F}_{nh}(x) - \frac{F(x)\big(1 - F(x)\big)}{n} + \frac{h}{n} F'(x)\nu_{1,L}\big((x - a)/h\big) - \frac{h^2}{2n} F''(x)\nu_{2,L}\big((x - a)/h\big) \Bigg| = o(n^{-1}h^2),$$

where

$$\nu_{1,L}(\alpha) = m_{1,L}(\alpha) + \alpha(1 - \mu_{0,L}(\alpha)^2)$$

and

$$\nu_{2,L}(\alpha) = m_{2,L}(\alpha) + \alpha^2(1 - \mu_{0,L}(\alpha)^2),$$

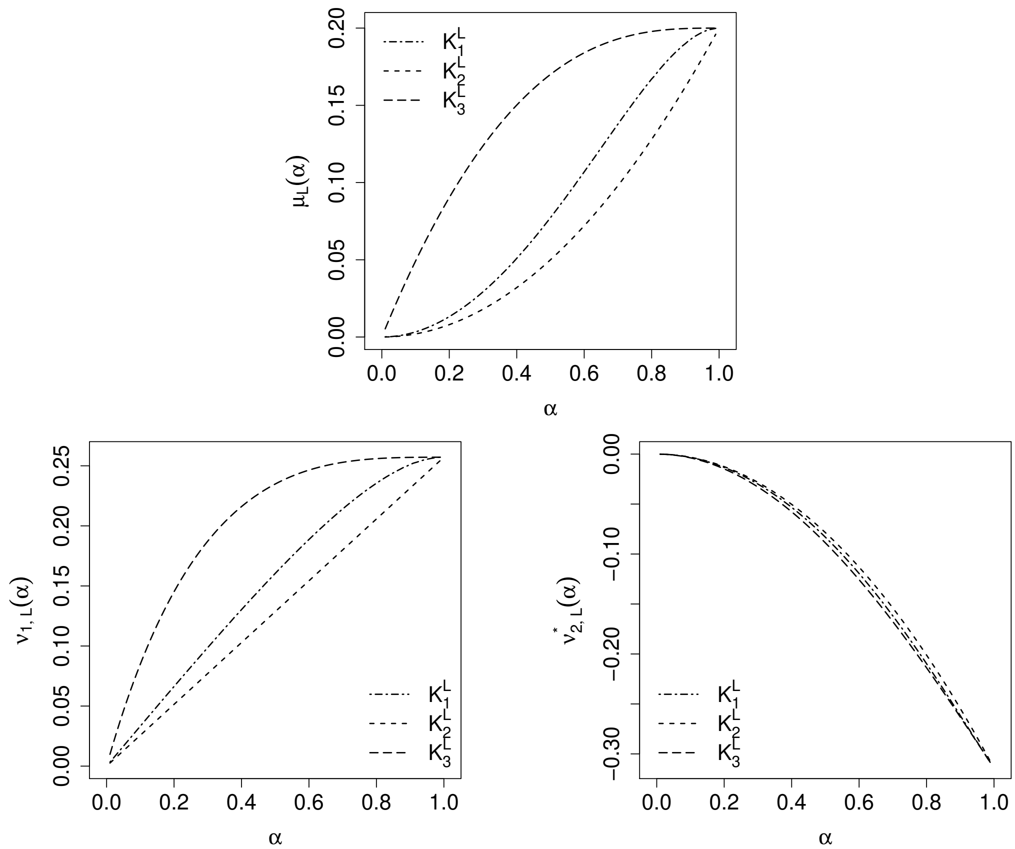

FIGURE 2. *Functions μ_L (top), $\nu_{1,L}$ (left bottom) and $\nu^*_{2,L}$ (right bottom) for the left boundary kernels K_q^L, with $q = 1, 2, 3$, where K is the Epanechnikov kernel.*

with $m_{k,L}(\alpha) = \int u^k B^L(u;\alpha)\,du$, *for* $k = 1, 2$, *and*

$$B^L(u;\alpha) = 2\bar{K}^L(u;\alpha)K^L(u;\alpha),$$

for $\alpha \in]0, 1[$. *Additionally, if* $F'_+(a) = 0$ *the previous expansion takes the form*

$$\sup_{x\in]a,a+h[}\left|\mathrm{Var}\tilde{F}_{nh}(x) - \frac{F(x)\big(1-F(x)\big)}{n} - \frac{h^2}{2n}F''(x)\nu^*_{2,L}\big((x-a)/h\big)\right| = o(n^{-1}h^2),$$

where

$$\nu^*_{2,L}(\alpha) = \nu_{2,L}(\alpha) - 2\alpha\nu_{1,L}(\alpha).$$

For all boundary kernels satisfying (6) it can be shown that

$$\nu_{1,L}(\alpha) = \int_{-1}^{\alpha} \bar{K}^L(u;\alpha)\big(1 - \bar{K}^L(u;\alpha)\big)du,$$

from which we deduce that $\nu_{1,L}(\alpha) > 0$ for all $\alpha \in]0,1[$, whenever the boundary kernel family satisfies $0 \le \bar{K}^L(u;\alpha) \le 1$, for all $u \in \mathbb{R}$ and $\alpha \in]0,1[$. Therefore, and similarly to what has been pointed out by other authors (see Azzalini, 1981, Tenreiro, 2013), we conclude that the kernel estimator $\tilde{F}_{nh}$ presents a local variance smaller than the variance of the empirical distribution function estimator whenever $F'_+(a) > 0$. The same conclusion is valid in the case $F'_+(a) = 0$ whenever the boundary kernel family satisfies

$$\nu^*_{2,L}(\alpha) = 2\int (\alpha - u)\bar{K}^L(u;\alpha)^2 du - \alpha^2 < 0, \text{ for all } \alpha \in]0,1[.$$

In order to undertake a first asymptotic comparison between the boundary kernels K^L_q given by (4), (5) and (7), we plot in Figure 2 the functions μ_L, $\nu_{1,L}$ and $\nu^*_{2,L}$ which are the coefficients of the most significant terms depending on the kernel in the expansions of the local variance and bias of estimator $\tilde{F}_{nh}(x)$ for x in the left boundary region. We take for K the Bartlett or Epanechnikov kernel $K(t) = \frac{3}{4}(1-t^2)I(|t| \le 1)$, but similar conclusions are valid for other polynomial kernels such as the uniform (in this case $K^L_1 = K^L_2$), the biweight or the triweight kernels (for the definition of these kernels see Wand and Jones, 1995, p. 31). In particular, the left boundary kernels K^L_q associated to all these kernels satisfy $\nu_{1,L}(\alpha) > 0$ and $\nu^*_{2,L}(\alpha) < 0$ for all $\alpha \in]0,1[$.

From the plots we also conclude that the boundary kernel K^L_2 has, uniformly over the boundary region, the lowest asymptotic bias but also the biggest asymptotic variance among the considered boundary kernels. In the case $F'_+(a) > 0$, the lowest asymptotic variance is obtained by the boundary kernel K^L_3, which also has the biggest asymptotic bias among the considered boundary kernels. In the case $F'_+(a) = 0$ we see that the three considered kernels present similar asymptotic variances with a small advantage for kernel K^L_3. Taking into account the bias behaviour, we conclude that the estimator based on kernel K^L_2 can be specially performing when $F'_+(a) = 0$. We postpone to Section 4 the analysis of the combined effect of bias and variance which depends on the underlying distribution F, especially throughout $F'(x)$ and $F''(x)$.

3. Extreme boundary behaviour

As we have seen in the previous section, although the estimators based on the considered classes of boundary kernels present different behaviours in the boundary region, the mean square error order of convergence does not reflect those differences. In fact, under the conditions of Theorem 2, for $h = Cn^{-1/3}$, with $C > 0$, we always have

$$\mathrm{MSE}\tilde{F}_{nh}(x) = O\big(n^{-4/3}\big).$$

Next we extend the previous analysis to the extreme boundary region. More precisely, we describe the asymptotic behaviour of the bias and variance of $\tilde{F}_{nh}(x)$ when x takes the form $x = a+\alpha h$, where $\alpha = \alpha_n$ converges to zero as n tends to infinity. As a consequence of this analysis, we will able to identify different orders of convergence for the mean square error of the estimators associate to boundary kernels K_1^L and K_2^L, and to K_3^L. For $k = 0, 1, \ldots$ and $\alpha \in]0,1[$, we will write $|\mu_{k,L}|(\alpha) := \int |u|^k |K^L(u;\alpha)| du \leq |\mu_{0,L}|(\alpha)$.

Theorem 3. *Under the conditions of Theorem 2, for $x = a + \alpha_n h$, with $\alpha_n \to 0$ as $n \to \infty$, we have:*

a)

$$\mathrm{E}\tilde{F}_{nh}(x) - F(x) = \frac{h^2}{2} F''(x)\mu_L(\alpha_n) + o\big(h^2|\mu_{2,L}|(\alpha_n)\big) + o\big(h^2\alpha_n^2\big).$$

b)

$$\begin{aligned}\mathrm{Var}\tilde{F}_{nh}(x) = {} & \frac{F(x)\big(1-F(x)\big)}{n} - \frac{h}{n} F'(x)\nu_{1,L}(\alpha_n) + \frac{h^2}{2n} F''(x)\nu_{2,L}(\alpha_n) \\ & + o\big(n^{-1}h^2|\mu_{2,L}|(\alpha_n)\big) + o\big(n^{-1}h^2\alpha_n^2\big).\end{aligned}$$

Additionally, if $F'_+(a) = 0$ the previous expansion takes the form

$$\begin{aligned}\mathrm{Var}\tilde{F}_{nh}(x) = {} & \frac{F(x)\big(1-F(x)\big)}{n} + \frac{h^2}{2n} F''(x)\nu^*_{2,L}(\alpha_n) \\ & + o\big(n^{-1}h^2|\mu_{1,L}|(\alpha_n)\big) + o\big(n^{-1}h^2|\mu_{2,L}|(\alpha_n)\big) + o\big(n^{-1}h^2\alpha_n^2\big).\end{aligned}$$

From the previous expansions we see that the mean square error order of convergence of $\tilde{F}_{nh}(x)$, for x is in the extreme boundary region, depends on the behaviour of $\mu_L(\alpha)$, $\nu_{1,L}(\alpha)$ and $\nu^*_{2,L}(\alpha)$ for α close to zero. For each one of the considered boundary kernel families K_q^L, for $q = 1, 2, 3$, we can obtain the expansions

$$\mu_L(\alpha) = \begin{cases} \frac{1}{3}\alpha^2 + o\big(\alpha^2\big), & \text{for } K_1^L \\ \int u^2 K(u)du\,\alpha^2, & \text{for } K_2^L \\ C\alpha + o(\alpha), & \text{for } K_3^L, \end{cases}$$

$$\nu_{1,L}(\alpha) = \begin{cases} \frac{1}{3}\alpha + o(\alpha), & \text{for } K_1^L \\ \int uB(u)du\,\alpha, & \text{for } K_2^L \\ \alpha - C_1\alpha^2 + o(\alpha^2), & \text{for } K_3^L \end{cases}$$

and

$$\nu_{2,L}^*(\alpha) = \begin{cases} -\frac{1}{3}\alpha^2 + o(\alpha^2), & \text{for } K_1^L \\ -\int (2u - u^2)B(u)du\,\alpha^2, & \text{for } K_2^L \\ -(1 - C_2)\alpha^2 + o(\alpha^2), & \text{for } K_3^L, \end{cases}$$

where $0 < \int uB(u)du < 1$ for a general kernel K, $C = \int_0^1 u^2 K(u)du \big/ \int_0^1 uK(u)du$, $C_k = \int_0^1 u^k B(-u)du \big/ \big(\int_0^1 uK(u)du\big)^2$, for $k = 1, 2$, and K is assumed to be differentiable on a right neighbourhood of the origin with $K(0) \neq 0$ (these additional assumptions on K are exclusively used to derive the previous expansions for the boundary kernel family K_1^L). In particular, taking for K the Epanechnikov kernel we get $\int u^2 K(u)du = 1/5$, $\int uB(u)du = 9/35$, $\int (2u - u^2)B(u)du = 11/35$, $C = 8/15$, $C_1 = 176/105$, and $C_2 = 17/45$.

From Theorem 3 we conclude that different orders of convergence for the bias are obtained for kernels K_1^L and K_2^L and for kernel K_3^L. In fact, the bias order of convergence to zero is bigger for kernels K_1^L and K_2^L than for kernel K_3^L. More precisely, we have

$$\begin{aligned} \mathrm{E}\tilde{F}_{nh}(x) - F(x) &= \frac{h^2}{2} F_+''(a)\mu_L(\alpha_n)(1 + o(1)) \\ &= \begin{cases} O\big(h^2\alpha_n^2\big), & \text{for } K_1^L \text{ and } K_2^L \\ O\big(h^2\alpha_n\big), & \text{for } K_3^L. \end{cases} \end{aligned}$$

In relation to the variance of the estimator, its order of convergence to zero is bigger for kernel K_3^L than for kernels K_1^L and K_2^L whenever $F_+'(a) > 0$. In fact, in this case we have

$$\begin{aligned} \mathrm{Var}\tilde{F}_{nh}(x) &= \frac{h}{n} F_+'(a)\big(\alpha_n - \nu_{1,L}(\alpha_n)\big)(1 + o(1)) \\ &= \begin{cases} O\big(n^{-1}h\alpha_n\big), & \text{for } K_1^L \text{ and } K_2^L \\ O\big(n^{-1}h\alpha_n^2\big), & \text{for } K_3^L. \end{cases} \end{aligned}$$

Finally, if $F'_+(a) = 0$ the order of convergence of the variance is the same for the three families of estimators

$$\mathrm{Var}\tilde{F}_{nh}(x) = \frac{h^2}{2n} F''_+(a)\big(\alpha_n^2 + \nu^*_{2,L}(\alpha_n)\big)(1 + o(1)) = O\big(n^{-1}h^2\alpha_n^2\big).$$

As a consequence of the previous expansions, we summarize in the following theorem the different orders of convergence we can observe for the mean square error of $\tilde{F}_{nh}(x)$ when x is in the extreme boundary region and h has the order of convergence of the asymptotically optimal bandwidth (11). We conclude that the mean square error order of convergence is bigger for kernel K_3^L than for kernels K_1^L and K_2^L whenever $F'_+(a) > 0$, and the inverse situation occurs whenever $F'_+(a) = 0$. This result suggests that the new class of boundary kernels K_3^L can be especially performing when the classical kernel distribution function estimator suffers from boundary problems.

Theorem 4. *Under the conditions of Theorem 3, let x be such that $x = a + \alpha_n h$, with $\alpha_n \to 0$ as $n \to \infty$, and take $h = Cn^{-1/3}$, with $C > 0$.*

a) If $F'_+(a) > 0$ we have

$$\mathrm{MSE}\tilde{F}_{nh}(x) = \begin{cases} O\big(n^{-4/3}\alpha_n\big), & \textit{for } K_1^L \textit{ and } K_2^L \\ O\big(n^{-4/3}\alpha_n^2\big), & \textit{for } K_3^L. \end{cases}$$

b) If $F'_+(a) = 0$ we have

$$\mathrm{MSE}\tilde{F}_{nh}(x) = \begin{cases} O\big(n^{-4/3}\alpha_n^2(n^{-1/3} + \alpha_n^2)\big), & \textit{for } K_1^L \textit{ and } K_2^L \\ O\big(n^{-4/3}\alpha_n^2\big), & \textit{for } K_3^L. \end{cases}$$

4. Exact finite sample comparisons

In this section we compare the boundary performance of the kernel estimator $\tilde{F}_{nh}$ when we take for K^L one of the left boundary kernels given by (4), (5) and (7), respectively. For that, we use as test distributions some beta mixtures of the form $wB(1,2) + (1-w)B(2,b)$, where $w \in [0,1]$ and the shape parameter b is such that $b \geq 2$. Four values of $w = 0, 0.25, 0.5, 0.75$ are considered, which lead to distributions with $F'_+(0) = 0, 0.5, 1, 1.5$, respectively. For each one of the previous weights w, two values for the shape parameter b are taken in order to get a second order derivative $F''_+(0)$ equal to 6 and 30. As the results observed for the test distributions with $F''_+(0)$ equal to 6 or 30 were quite similar, we will focus our comments on the results obtained for the test distributions with $F''_+(0) = 6$ whose probability densities are shown in Figure 3.

For each one of these test distributions we present in Figure 4 the exact variance, $\mathrm{V}(x)$, square bias, $\mathrm{B}(x)^2$, and mean square error, $\mathrm{MSE}(x) = \mathrm{V}(x) + \mathrm{B}(x)^2$, of $\tilde{F}_{nh}(x)$, for $x = \alpha h$ and $\alpha \in]0, 1[$, where

$$n\mathrm{V}(x) := n\mathrm{Var}\tilde{F}_{nh}(\alpha h) = \int F((\alpha - u)h)B^L(u;\alpha)du - \big(\mathrm{E}\tilde{F}_{nh}(\alpha h)\big)^2$$

and

$$\mathrm{B}(x) := \mathrm{E}\tilde{F}_{nh}(\alpha h) - F(\alpha h) = \int F((\alpha - u)h)K^L(u;\alpha)\,du - F(\alpha h)$$

(on these expressions see Section 5 below). For comparative purposes the mean square error of the sample distribution function estimator is also included in the graphics. We have considered the sample size $n = 100$. Similar pictures were generated for other sample sizes but they are not included here to save space. As before, we have taken for K the Epanechnikov kernel. The global bandwidth h that determines the boundary region was always taken equal to the asymptotically optimal bandwidth h_0 given by (11).

From the graphics we conclude that the boundary behaviour of the kernel estimator based on the boundary kernels K_q^L, for $q = 1, 2, 3$, is dominated by the magnitude of the underlying density $f = F'$ over the boundary region. As predicted by the asymptotic theory previously exposed, the kernel estimator based on the boundary kernel K_3^L presents the lowest variance among the considered boundary kernel estimators for all the test distributions. The reduced bias shown by this estimator for distributions with large values of

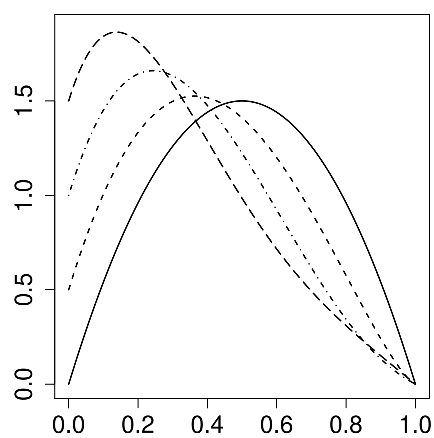

FIGURE 3. *Beta mixture test densities* $wB(1,2) + (1-w)B(2,b)$ *with* $F'_+(0) = 0, 0.5, 1, 1.5$ *and* $F''_+(0) = 6$.

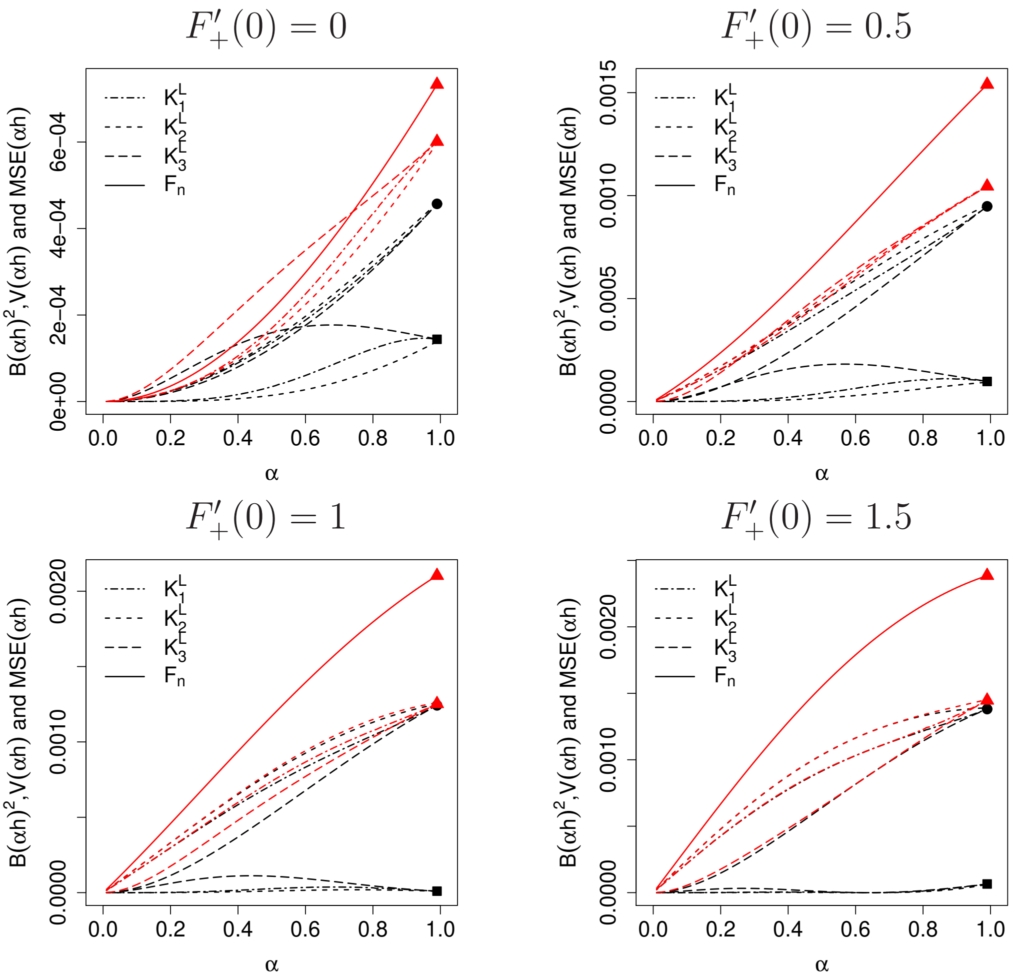

FIGURE 4: $\mathrm{B}(\alpha h)^2$ (■), $\mathrm{V}(\alpha h)$ (●) *and* $\mathrm{MSE}(\alpha h)$ (▲), *for* K_q^L, $q = 1, 2, 3$, *with* K *the Epanechnikov kernel, and for the sample distribution function* F_n, *where* F *is the beta mixture distributions shown in Figure 3. The sample size is* $n = 100$.

$F'_+(0)$ explains its superior mean square error performance in relation to both boundary kernels K_1^L and K_2^L . The graphics obtained for the test distributions with $F''_+(0) = 30$ (but not shown here) also reveal that this advantage over the second order boundary kernels K_1^L and K_2^L is bigger for small than for large values of $F''_+(0)^2$, which is in accordance with the asymptotic expansion for the bias presented in Theorem 2. When the underlying density is such that $F'_+(0) = 0$, in which case the classical kernel estimator does not suffer from boundary problems, we see that the boundary kernels K_1^L and K_2^L perform similarly being both better than K_3^L. The large bias presented by the kernel estimator based on the boundary kernel K_3^L explains the poor mean square error results obtained for this estimator. Finally, for intermediate values of $F'_+(0)$ the three considered left boundary kernels have shown a similar performance.

Based on this evidence, we conclude that none of the considered boundary kernels is the best over the considered set of test distributions. The kernel estimator based on the new boundary kernel K_3^L has shown to be especially performing when the classical kernel estimator suffers from severe boundary problems. However, it may present a large bias otherwise, being outperformed by the estimators based on the boundary kernels K_1^L and K_2^L. These findings agree with the asymptotic based conclusions gathered in Theorem 4.

5. Proofs

Proof of Theorem 1: This proof follows closely the lines of the proofs of Theorems 3.1 and 3.2 of Tenreiro (2013) the reason way the details are omitted. However, in order to deal with boundary kernels for which $\mu_{0,L}(\alpha) = 0$ for some $\alpha \in]0,1[$, the following integration by parts formula, that generalizes Lemma 6.1 of Tenreiro (2013), is needed.

Lemma 1. *If Φ is a probability distribution function and $\Psi(u) = \int_{-\infty}^{u} \psi(v)dv$ where ψ is a Lebesgue integrable function, then $\int \Phi d\Psi + \int \Psi d\Phi = \int \psi(v)dv$.*

Proof: If $\int \psi(v)dv \neq 0$, define $\psi_0 = \psi / \int \psi(v)dv$ and use Lemma 6.1 of Tenreiro (2013) with $\psi = \psi_0$. If $\int \psi(v)dv = 0$, consider ψ^+ and ψ^-, the positive and the negative parts of ψ, that satisfy $\int \psi^+(v)dv = \int \psi^-(v)dv = I$ (say). If $I = 0$, the stated result is obvious because in this case $\psi = 0$ a.e.. If $I > 0$, the stated result follows from the first part of the proof by taking $\psi = \psi^+$ and $\psi = \psi^-$. ∎

Proof of Theorem 2: For $x \in]a, a+h[$, the expectation of $\tilde{F}_{nh}(x)$ is given by

$$\begin{aligned} \mathrm{E}\tilde{F}_{nh}(x) &= \int \bar{K}^L((x-y)/h; (x-a)/h) f(y)\, dy \\ &= \int F(x-uh) K^L(u; (x-a)/h)\, du, \end{aligned}$$

(see Tenreiro, 2013, p. 186). By the continuity of the second derivative of F on $[a,b]$ and Taylor's formula, we have

$$F(x-uh) = F(x) - uhF'(x) + u^2h^2 \int_0^1 (1-t)F''(x-tuh)\,dt, \tag{12}$$

for $-1 \le u \le (x-a)/h$, from which we deduce that

$$\mathrm{E}\tilde{F}_{nh}(x) - F(x) - \frac{h^2}{2}F''(x)\mu_L((x-a)/h) = A(x,h) + B(x,h), \tag{13}$$

where

$$\begin{aligned} A(x,h) = F(x)\big(\mu_{0,L}((x-a)/h) - 1\big) - hF'(x)\mu_{1,L}((x-a)/h) \\ + \frac{h^2}{2}F''(x)((x-a)/h)\mu_{1,L}((x-a)/h), \end{aligned}$$

and

$$B(x,h) = h^2 \iint_0^1 (1-t)\big(F''(x-tuh) - F''(x)\big)dt\, u^2K^L(u;(x-a)/h)\,du,$$

is such that

$$\sup_{x\in\,]a,a+h[} |B(x,h)| \le \frac{h^2}{2} \sup_{\alpha\in\,]0,1[} |\mu_{0,L}|(\alpha) \sup_{y,z\in[a,b]:\,|y-z|\le h} |F''(y) - F''(z)|. \tag{14}$$

On the other hand, taking into account that $F(a) = 0$ and using condition (6) and the Taylor's expansion

$$\begin{aligned} F(x) = (x-a)F'(x) - \frac{1}{2}(x-a)^2F''(x) \\ - (x-a)^2 \int_0^1 (1-t)\big(F''(x-(x-a)t) - F''(x)\big)dt, \end{aligned} \tag{15}$$

we get

$$A(x,h) = -\big(\mu_{0,L}((x-a)/h) - 1\big)(x-a)^2 \int_0^1 (1-t)\big(F''(x-(x-a)t) - F''(x)\big)dt,$$

where

$$\sup_{x\in\,]a,a+h[} |A(x,h)| \le h^2 \sup_{\alpha\in\,]0,1[} |\mu_{0,L}(\alpha) - 1| \sup_{y,z\in[a,b]:\,|y-z|\le h} |F''(y) - F''(z)|. \tag{16}$$

Part a) of Theorem 2 follows now from (13), (14) and (16), and the fact that

$$\sup_{y,z\in[a,b]:\,|y-z|\leq h} |F''(y) - F''(z)| = o(1).$$

From Part a), the variance of $\tilde{F}_{nh}(x)$ is given by

$$\begin{aligned} n\mathrm{Var}\tilde{F}_{nh}(x) &= \int \bar{K}^L(u;(x-a)/h)^2 h f(x-uh)du - \big(\mathrm{E}\tilde{F}_{nh}(x)\big)^2 \\ &= F(x)(1-F(x)) + C(x,h) + o\big(h^2\big), \end{aligned} \tag{17}$$

uniformly in $x \in\,]a, a+h[$, where

$$\begin{aligned} C(x,h) &= \int \bar{K}^L(u;(x-a)/h)^2 h f(x-uh)du - F(x) \\ &= \int F(x-zh)B^L(z;(x-a)/h)dz - F(x). \end{aligned}$$

Moreover, using (12) and the fact that $\int B^L(z;\alpha)dz = \mu_{0,L}(\alpha)^2$, we deduce that

$$\begin{aligned} C(x,h) &= F(x)\big(\mu_{0,L}((x-a)/h)^2 - 1\big) - hF'(x)m_{1,L}((x-a)/h) \\ &\quad + h^2 \iint_0^1 (1-t)F''(x-tuh)dt u^2 B^L(u;(x-a)/h)du \qquad (18) \\ &= F(x)\big(\mu_{0,L}((x-a)/h)^2 - 1\big) - hF'(x)m_{1,L}((x-a)/h) \\ &\quad + \frac{h^2}{2}F''(x)m_{2,L}((x-a)/h) + o(h^2), \end{aligned}$$

uniformly in $x \in\,]a, a+h[$, as $\sup_{\alpha\in\,]0,1[} \int |u^2 B^L(u;\alpha)|du < \infty$.

Finally, from Taylor's expansion (15) we get

$$\sup_{x\in\,]a,a+h[} \left| C(x,h) + h\,F'(x)\nu_{1,L}\big((x-a)/h\big) - \frac{h^2}{2}\,F''(x)\nu_{2,L}\big((x-a)/h\big)\right| = o(h^2),$$

which concludes the proof. ∎

Proof of Theorem 3: Part a) follows from (13) where for $x = a + \alpha_n h$

$$A(x,h) = -\big(\mu_{0,L}(\alpha_n) - 1\big)(x-a)^2 \int_0^1 (1-t)\big(F''(x-(x-a)t) - F''(x)\big)dt,$$

and

$$B(x,h) = h^2 \iint_0^1 (1-t)\big(F''(x-tuh) - F''(x)\big) dt\, u^2 K^L(u;\alpha_n)\, du,$$

with

$$|A(x,h)| \le |\mu_{0,L}(\alpha_n) - 1| h^2 \alpha_n^2 \sup_{y,z\in[a,b]:\, |y-z|\le h} |F''(y) - F''(z)|/2 = o\big(h^2\alpha_n^2\big),$$

and

$$|B(x,h)| \le h^2 |\mu_{2,L}|(\alpha_n) \sup_{y,z\in[a,b]:\, |y-z|\le h} |F''(y) - F''(z)|/2 = o\big(h^2 |\mu_{2,L}|(\alpha_n)\big).$$

In order to establish Part b), we start by using (17) and (18) to write

$$\begin{aligned} n\mathrm{Var}\tilde{F}_{nh}(x) &= F(x)(1-F(x)) + C(x,h) \\ &\quad - \big(\mathrm{E}\tilde{F}_{nh}(x) - F(x)\big)^2 + 2\big(\mathrm{E}\tilde{F}_{nh}(x) - F(x)\big)F(x), \end{aligned} \tag{19}$$

with

$$\begin{aligned} C(x,h) &= F(x)\big(\mu_{0,L}(\alpha_n)^2 - 1\big) - hF'(x)m_{1,L}(\alpha_n) + \frac{h^2}{2}F''(x)m_{2,L}(\alpha_n) \\ &\quad + h^2 \iint_0^1 (1-t)\big(F''(x-tzh) - F''(x)\big) dt z^2 B^L(z;\alpha_n) dz, \end{aligned}$$

and

$$\mathrm{E}\tilde{F}_{nh}(x) - F(x) = O\big(h^2(|\mu_{2,L}|(\alpha_n) + \alpha_n^2)\big), \tag{20}$$

where the latter equality follows from Part a) and conditions (6) and (8).

But

$$F(x) = h\alpha_n F'(x) - \frac{h^2}{2}\alpha_n^2 F''(x) - h^2\alpha_n^2 \int_0^1 (1-t)\big(F''(x-h\alpha_n t) - F''(x)\big) dt,$$

which leads to

$$\begin{aligned} C(x,h) &= -hF'(x)\nu_{1,L}(\alpha_n) + \frac{h^2}{2}F''(x)\nu_{2,L}(\alpha_n) \\ &\quad - h^2\alpha_n^2\big(\mu_{0,L}(\alpha_n)^2 - 1\big) \int_0^1 (1-t)\big(F''(x-h\alpha_n t) - F''(x)\big) dt \\ &\quad + h^2 \iint_0^1 (1-t)\big(F''(x-tzh) - F''(x)\big) dt z^2 B^L(z;\alpha_n) dz \end{aligned}$$

$$= -hF'(x)\nu_{1,L}(\alpha_n) + \frac{h^2}{2}F''(x)\nu_{2,L}(\alpha_n) + o(h^2\alpha_n^2) + o\big(h^2|\mu_{2,L}|(\alpha_n)\big).$$

Additionally, if $F'(a) = 0$, we have

$$F'(x) = h\alpha_n F''(x) + h\alpha_n \int_0^1 \big(F''(x + h\alpha_n t) - F''(x)\big)dt,$$

and in this case

$$C(x,h) = \frac{h^2}{2}F''(x)\nu^*_{2,L}(\alpha_n) + o\big(h^2\alpha_n|\mu_{1,L}|(\alpha_n)\big) + o(h^2\alpha_n^2) + o\big(h^2|\mu_{2,L}|(\alpha_n)\big).$$

Part b) of Theorem 3 follows now from (19), (20) and the previous expressions for $C(x,h)$. ■

Acknowledgments

Research partially supported by the Centre for Mathematics of the University of Coimbra – UID/MAT/00324/2013, funded by the Portuguese Government through FCT/MEC and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

References

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika* 68, 326–328.

Chacón, J.E., Monfort, P., Tenreiro, C. (2014). Fourier methods for smooth distribution function estimation. *Statist. Probab. Lett.* 84, 223–230.

Chacón, J.E., Rodríguez-Casal, A. (2010). A note on the universal consistency of the kernel distribution function estimator. *Statist. Probab. Lett.* 80, 1414–1419.

Gasser, T., Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation* (T. Gasser and M. Rosenblatt, Eds.), Lecture Notes in Mathematics 757, 23–68.

Gasser, T., Müller, H.-G., Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 47, 238–252.

Giné, E., Nickl, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields* 143, 569–596.

Jones, M.C. (1990). The performance of kernel density functions in kernel distribution function estimation. *Statis. Probab. Lett.* 9, 129–132.

Mason, D.M., Swanepoel, J.W.H. (2011). A general result on the uniform in bandwidth consistency of kernel-type function estimators. *TEST* 20, 72–94.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* 78, 521–530.

Nadaraya, E.A. (1964). Some new estimates for distribution functions. *Theory Probab. Appl.* 9, 497–500.

Parzen, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* 74, 105–121.

Rice, J. (1984). Boundary modification for kernel regression. *Comm. Statist. Theory Methods* 13, 893–900.

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

Tiago de Oliveira, J. (1963). Estatística de densidades: resultados assintóticos. *Rev. Fac. Ciên. Lisboa* 9, 111–206.

Wand, M.P., Jones, M.C. (1995). *Kernel smoothing*. London: Chapman & Hall.

Watson, G.S., Leadbetter, M.R. (1964). Hazard analysis II. *Sankhyā Ser. A* 26, 101–116.

Winter, B.B. (1973). Strong uniform consistency of integrals of density estimators. *Canad. J. Statist.* 1, 247–253.

Winter, B.B. (1979). Convergence rate of perturbed empirical distribution functions. *J. Appl. Probab.* 16, 163–173.

Yamato, H. (1973). Uniform convergence of an estimator of a distribution function. *Bull. Math. Statist.* 15, 69–78.

Carlos Tenreiro
CMUC, Department of Mathematics, University of Coimbra, 3001–501 Coimbra, Portugal
E-mail address: tenreiro@mat.uc.pt