AUSTRIAN JOURNAL OF STATISTICS
Volume 38 (2009), Number 1, 17–32

On Boundary Correction in Kernel Estimation of ROC Curves

Jan Koláček[1] and Rohana J. Karunamuni[2]

[1]Dept. of Mathematics and Statistics, Brno
[2]Dept. of Mathematical and Statistical Sciences, University of Alberta

Abstract: The Receiver Operating Characteristic (ROC) curve is a statistical tool for evaluating the accuracy of diagnostics tests. The empirical ROC curve (which is a step function) is the most commonly used non-parametric estimator for the ROC curve. On the other hand, kernel smoothing methods have been used to obtain smooth ROC curves. The preceding process is based on kernel estimates of the distribution functions. It has been observed that kernel distribution estimators are not consistent when estimating a distribution function near the boundary of its support. This problem is due to "boundary effects" that occur in nonparametric functional estimation. To avoid these difficulties, we propose a generalized reflection method of boundary correction in the estimation problem of ROC curves. The proposed method generates a class of boundary corrected estimators.

Zusammenfassung: Die Receiver Operating Characteristic (ROC) Kurve ist ein statistisches Werkzeug zur Bewertung der Präzision diagnostischer Tests. Die empirische ROC Kurve (sie ist eine Treppenfunktion) ist der am weitesten verbreitete nicht-parametrische Schätzer der ROC Kurve. Andererseits wurden Kerngättungsmethoden verwendet, um glatte ROC Kurven zu erhalten. Der vorangehende Prozess basiert dabei auf Kernschätzungen der Verteilungsfunktionen. Es wurde beobachtet, dass Kernschätzer der Verteilung nicht konsistent sind falls die Verteilungsfunktion in der Nähe des Randes ihres Trägers geschätzt wird. Dieses Problem beruht auf dem "Randeffekt" der in der nicht-parametrischen funktionalen Schätzung auftritt. Um derartige Schwierigkeiten zu vermeiden, empfehlen wir eine verallgemeinerte Reflexionsmethode der Randkorrektur im Schätzproblem von ROC Kurven. Die vorgeschlagene Methode generiert eine Klasse von randkorrigierten Schätzern.

Keywords: Reflection, Distribution Estimation.

1 Introduction

The Receiver Operating Characteristic (ROC) describes the performance of a diagnostic test which classifies subjects into either group without condition $\mathcal{G}_0$ or group with condition $\mathcal{G}_1$ by means of a continuous discriminant score X, i.e., a subject is classified as $\mathcal{G}_1$ if $X \geq d$ and $\mathcal{G}_0$ otherwise for a given cutoff point $d \in \mathbb{R}$. The ROC is defined as a plot of probability of false classification of subjects from $\mathcal{G}_1$ versus the probability of true classification of subjects from $\mathcal{G}_0$ across all possible cutoff point values of X. Specifically, let

F_0 and F_1 denote the distribution functions of X in the groups $\mathcal{G}_0$ and $\mathcal{G}_1$, respectively. Then, the ROC curve can be written as

$$R(p) = 1 - F_1\left(F_0^{-1}(1-p)\right) , \qquad 0 < p < 1 ,$$

where p is the false positive rate in $(0, 1)$ as the corresponding cut-off point ranges from $-\infty$ to $+\infty$ and F_0^{-1} denotes the inverse function of F_0.

A simple non-parametric estimator for $R(p)$ is to use the empirical distribution functions for F_0 and F_1. The resulting ROC curve is a step function and it is called the empirical ROC curve. Another type of non-parametric estimator for $R(p)$ is derived from kernel smoothing methods. Kernel smoothing is most widely used mainly because it is easy to derive and has good asymptotic and small sample properties. Kernel smoothing has received a considerable attention in density estimation context; see, for example the monographs of Silverman (1986) and Wand and Jones (1995). However, applications of kernel smoothing in distribution function estimation are relatively few. Some theoretical properties of a kernel distribution function estimator have been investigated by Nadaraya (1964), Reiss (1981), and Azzalini (1981). Lloyd (1998) proposed a nonparametric estimator of ROC by using kernel estimators for the distribution functions F_0 and F_1.

Lloyd and Yong (1999) showed that Lloyd's estimator has better mean squared error properties than the empirical ROC curve estimator. However, his estimator has some drawbacks. For example, Lloyd's estimator is unreliable near the end points of the support of the ROC curve due to so-called "boundary effects" that occur in nonparametric functional estimation. Although there is a vast literature on boundary correction in density estimation context, boundary effects problem in distribution function context has been less studied.

In this paper, we develop a new kernel type estimator of the ROC curve that removes boundary effects near the end points of the support. Our estimator is based on a new boundary corrected kernel estimator of distribution functions and it is based on ideas of Karunamuni and Alberts (2005a, 2005b, 2006), Zhang and Karunamuni (1998, 2000), (Karunamuni and Zhang, 2008), and Zhang, Karunamuni, and Jones (1999) developed for boundary correction in kernel density estimation. The basic technique of construction of the proposed estimator is kind of a generalized reflection method involving reflecting a transformation of the observed data. In fact, the proposed method generates a class of boundary corrected estimators. We derive expressions for the bias and variance of the proposed estimator. Furthermore, the proposed estimator is compared with the "classical estimator" using simulation studies. We observe that the proposed estimator successfully remove boundary effects and performs considerably better than the "classical estimator".

Kernel smoothing in distribution function and ROC curve estimation is discussed in the next section. The proposed estimator is given in Section 3. Simulation results are given in Section 4. A real data example is analyzed in Section 5. Finally, some concluding remarks are given in Section 6.

2 Kernel Smoothing

2.1 Kernel ROC Estimator

Suppose that independent samples $X_{01},\dots,X_{0n_0}$ and $X_{11},\dots,X_{1n_1}$ are available from some two unknown distributions F_0 and F_1, respectively, where $F_0 \in \mathcal{G}_0$ and $F_1 \in \mathcal{G}_1$ and $\mathcal{G}_0$ and $\mathcal{G}_1$ denote two groups of continuous distribution functions. Then a simple nonparametric estimator of the ROC curve $R(p) = 1 - F_1\left(F_0^{-1}(1-p)\right)$, $0 < p < 1$, is known as the *empirical ROC* curve given by

$$\widetilde{R}_E(p) = 1 - \widetilde{F}_1\left(\widetilde{F}_0^{-1}(1-p)\right)\,, \qquad 0 \le p \le 1\,,$$

where $\widetilde{F}_0$ and $\widetilde{F}_1$ denote the empirical distribution functions of F_0 and F_1 based on the data $X_{01},\dots,X_{0n_0}$ and $X_{11},\dots,X_{1n_1}$, respectively; that is

$$\widetilde{F}_0(x) = \frac{1}{n_0}\sum_{i=1}^{n_0} I(X_{0i} \le x)\,, \qquad \widetilde{F}_1(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} I(X_{1i} \le x)\,.$$

Note that $\widetilde{R}$ is not a continuous function. In fact, it is a step function on the interval $[0,1]$. This is a notable weakness of the empirical ROC curve $\widetilde{R}(p)$. Since the ROC curve is a smooth function of p, we would expect to have an estimator that is smooth as well. Lloyd (1998) proposed a smooth estimator using kernel smoothing techniques. His idea is to replace unknown distribution F_0 and F_1 by two smooth kernel estimators. Specifically, he employed following kernel estimators of F_0 and F_1:

$$\widehat{F}_0(x) = \frac{1}{n_0}\sum_{i=1}^{n_0} W\left(\frac{x - X_{0i}}{h_0}\right)\,, \qquad \widehat{F}_1(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} W\left(\frac{x - X_{1i}}{h_1}\right)\,,$$

where $W(x) = \int_{-1}^{x} K(t)dt$, h_0 and h_1 denote bandwidths ($h_0 \to 0$ and $h_1 \to 0$ as $n_0 \to \infty$ and $n_1 \to \infty$, respectively), and K is a unimodal symmetric density function with support $[-1,1]$. The corresponding estimator of the ROC curve $R(p)$ is then given by

$$\widehat{R}(p) = 1 - \widehat{F}_1\left(\widehat{F}_0^{-1}(1-p)\right)\,, \qquad 0 \le p \le 1\,.$$

An example of a smooth estimate of $R(p)$ using $\widehat{R}(p)$ is illustrated in Figure 1.

When $\mathcal{G}_0$ and $\mathcal{G}_1$ contain distributions with finite support then the estimator $\widehat{R}$ exhibits boundary effects near the endpoints of the support due to the same boundary effects that occur in the uncorrected kernel estimators $\widehat{F}_0$ and $\widehat{F}_1$. The main purpose of this article is to improve the kernel distribution estimators and thereby to avoid boundary effects of smooth kernel ROC estimators. Details of the boundary problem with $\widehat{F}_0$ and $\widehat{F}_1$ are described in the next section.

2.2 Kernel Distribution Estimator and Boundary Effects

Let f denote a continuous density function with support $[0,a]$, $0 < a \le \infty$, and consider nonparametric estimation of the cumulative distribution function F of f based on a random sample $X_1,\dots,X_n$ from f. Suppose that $F^{(j)}$, the j-th derivative of F, exists and is

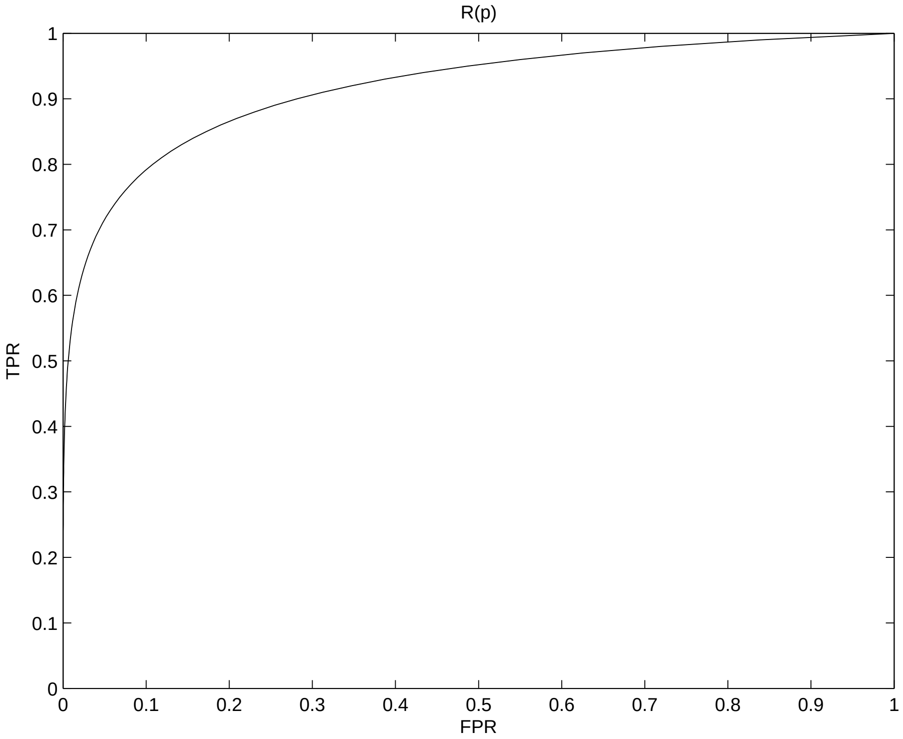


Figure 1: Smooth estimate of $R(p)$.

continuous on $[0, a]$, $j = 0, 1, 2$, with $F^{(0)} = F$ and $F^{(1)} = f$. Then the traditional kernel estimator of F is given by

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \qquad W(x) = \int_{-1}^{x} K(t)dt\,,$$

where K is a symmetric density function with support $[-1, 1]$ and h is the bandwidth ($h \to 0$ as $n \to \infty$). The basic properties of $\widehat{F}_{h,K}(x)$ at interior points are well-known (e.g. Lejeune and Sarda, 1992), and under some smoothness assumptions these include, for $h \leq x \leq a - h$,

$$\begin{aligned} \mathrm{E}\left(\widehat{F}_{h,K}(x)\right) - F(x) &= \frac{1}{2}\beta_2 f^{(1)}(x)h^2 + o(h^2) \\ n\mathrm{var}\left(\widehat{F}_{h,K}(x)\right) &= F(x)\left(1 - F(x)\right) + hf(x)\int_{-1}^{1} W(t)\left(W(t) - 1\right)dt + o(h)\,. \end{aligned}$$

The performance of $\widehat{F}_{h,K}(x)$ at boundary points, i.e., for $x \in [0, h) \cup (a - h, a]$, however, differs from the interior points due to so-called "boundary effects" that occur in nonparametric curve estimation problems. More specifically, the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ instead of $O(h^2)$ at boundary points, while the variance of $\widehat{F}_{h,K}(x)$ is of the same order. This fact can be clearly seen by examining the behavior of $\widehat{F}_{h,K}$ inside the left boundary region $[0, h]$. Let x be a point in the left boundary, i.e., $x \in [0, h]$. Then we can write

$x = ch$, $0 \le c \le 1$. The bias and variance of $\widehat{F}_{h,K}(x)$ at $x = ch$ are of the form

$$\begin{aligned} \mathrm{E}\left(\widehat{F}_{h,K}(x)\right) - F(x) &= hf(0)\int_{-1}^{-c} W(t)dt \\ &\quad + h^2 f^{(1)}(0)\left\{\frac{c^2}{2}+c\int_{-1}^{-c} W(t)dt-\int_{-1}^{c} tW(t)dt\right\} + o(h^2) \qquad (1) \\ n\mathrm{var}\left(\widehat{F}_{h,K}(x)\right) &= F(x)(1-F(x)) + hf(0)\left\{\int_{-1}^{c} W^2(t)dt - c\right\} + o(h)\,. \qquad (2) \end{aligned}$$

From expression (1) it is now clear that the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ instead of $O(h^2)$. To remove this boundary effect in kernel distribution estimation we investigate a new class of estimators in the next section.

3 The Proposed Estimator

In this section we propose a class of estimators of the distribution function F of the form

$$\widetilde{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right\}, \qquad (3)$$

where h is the bandwidth, K is a symmetric density function with support $[-1,1]$, and g_1 and g_2 are two transformations that need to be determined. The same type of estimator in density estimation case has been discussed in Zhang et al. (1999). As in the preceding paper, we assume that g_i, $i = 1,2$, are nonnegative, continuous and monotonically increasing functions defined on $[0,\infty)$. Further assume that g_i^{-1} exists, $g_i(0) = 0$, $g_i^{(1)}(0) = 1$, and that $g_i^{(2)}$ exists and is continuous on $[0,\infty)$, where $g_i^{(j)}$ denotes the j-th derivative of g_i, with $g_i^{(0)} = g_i$ and g_i^{-1} denoting the inverse function of g_i, $i = 1,2$. We will choose g_1 and g_2 such that $\widetilde{F}_{h,K}(x) \ge 0$ everywhere. Note that the i-th term of the sum in (3) can be expressed as

$$W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right) = \int_{\frac{-x+g_1(X_i)}{h}}^{\frac{x+g_2(X_i)}{h}} K(t)dt\,.$$

The preceding integral is non-negative provided the inequality $-x+g_1(X_i) \le x+g_2(X_i)$ holds. Since $x \ge 0$, the preceding inequality will be satisfied if g_1 and g_2 are such that $g_1(X_i) \le g_2(X_i)$ for $i = 1,\dots,n$. Thus we will assume that g_1 and g_2 are chosen such that $g_1(x) \le g_2(x)$ for $x \in [0,\infty)$ for our proposed estimator. Now, we can obtain the bias and variance of (3) at $x = ch$, $0 \leq c \leq 1$, as

$$
\begin{aligned}
\mathrm{E}\left(\widetilde{F}_{h,K}(x)\right) - F(x) &= h^2 \left\{ f^{(1)}(0) \left(\frac{c^2}{2} + 2c \int_{-1}^{-c} W(t)dt - \int_{-c}^{c} tW(t)dt \right) \right. \\
&\quad - f(0) g_1^{(2)}(0) \int_{-1}^{c} (c-t) W(t) dt \qquad (4) \\
&\quad \left. - f(0) g_2^{(2)}(0) \int_{-1}^{-c} (c+t) W(t) dt \right\} + o(h^2) \\
n\mathrm{var}\left(\widetilde{F}_{h,K}(x)\right) &= F(x)(1-F(x)) + hf(0) \left\{ \int_{-1}^{c} W^2(t)dt \right. \\
&\quad \left. -2 \int_{-1}^{c} W(t)W(t-2c)dt + \int_{-1}^{-c} W^2(t)dt \right\} + o(h) \, . \qquad (5)
\end{aligned}
$$

The proofs of (4) and (5) are given in the Appendix. Note that the contribution of g_2 on the bias vanishes as $c \to 1$. By comparing expressions (1), (4), (2), and (5) at boundary points we can see that the variances are of the same order and the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ whereas the bias of $\widetilde{F}_{h,K}(x)$ is of order $O(h^2)$. So our proposed estimator removes boundary effects in kernel distribution estimation since the bias at boundary points is of the same order as the bias at interior points.

It is clear that there are various possible choices available for the pair (g_1, g_2). However, we will choose g_1 and g_2 so that the condition $\widetilde{F}_{h,K}(0) = 0$ will be satisfied because of the fact that $F(0) = 0$. A sufficient (but not necessary) condition for the preceding condition to be satisfied is that g_1 and g_2 must be equal. Thus we need to construct a single transformation function g such that $g = g_1 = g_2$. Other important properties that are desirable in the estimator $\widehat{F}_{h,K}$ are the local adaptivity (i.e., the transformation function g depends on c) and that $\widetilde{F}_{h,K}(x)$ being equal to the usual kernel estimator $\widehat{F}_{h,K}(x)$ at interior points. For the latter, g must satisfy that $g(y) \to y$ as $c \to 1$. In order to display the dependance of g on c, $0 \leq c \leq 1$, we shall denote g by g_c in what follows.

Summarizing all the assumptions, it is clear now that g_c should satisfy the conditions

(i) $g_c : [0, \infty) \to [0, \infty)$, g_c is continuous, monotonically increasing and $g_c^{(i)}$ exists, $i = 1, 2$.

(ii) $g_c^{-1}(0) = 0$ and $g_c^{(1)}(0) = 1$.

(iii) $g_c(y) \to y$ for $c \to 1$.

Functions satisfying conditions (i) to (iii) are easy to construct. The trivial choice is $g_c(y) = y$, which represents the "classical" reflection method estimator. Based on extensive simulations, we observed that the following transformation adapts well to various shapes of distributions:

$$
g_c(y) = y + \frac{1}{2} I_c y^2 \, , \qquad (6)
$$

for $y \geq 0$ and $0 \leq c \leq 1$, where $I_c = \int_{-1}^{-c} W(t)dt$.

Remark: Some discussion on the above choice of g_c and other various improvements that can be made would be appropriate here. It is possible to construct functions g_c that improve the bias further under some additional conditions. For instance, if one examines the right hand side of bias expansion (4) then it is not difficult to see that the terms inside bracket (i.e., the coefficient of h^2) can be made equal to zero if g_c is appropriately chosen. Indeed, if g_c is chosen such that

$$f(0)g_c^{(2)}(0)\left\{\int_{-1}^{c}(c-t)W(t)dt+\int_{-1}^{-c}(c+t)W(t)dt\right\}$$
$$=f^{(1)}(0)\left(\frac{c^2}{2}+2c\int_{-1}^{-c}W(t)dt-\int_{-c}^{c}tW(t)dt\right),$$

then the bias of $\widetilde{F}_{h,K}(x)$ would be theoretically of order $O(h^3)$. For such a function g_c, the second derivative at zero, $g_c^{(2)}(0)$, will depend on the ratio $d_1 = f^{(1)}(0)/f(0)$. In this case, the function g_c would probably be some cubic polynomial; see e.g. Karunamuni and Alberts (2005a, 2005b, 2006). Then the problem of estimation of d_1 naturally arises as in the preceding paper. Another problem that one would face is that the second derivative $g_c^{(2)}(0)$ may not go to 0 when $c \to 1$ as in the case of density estimation context. Thus one may not be able to find any function g_c which satisfies condition (iii) and hence the estimator $\widetilde{F}_{h,K}$ loses the property of "natural extension" to the classical estimator outside the boundary points. These are basically the main reasons why we decided to implement a quadratic function defined in (6) as our choice of transformation.

4 Simulation

To test the effectiveness of our estimator, we simulated its performance against the reflection method. The simulation is based on 1000 replications. In each replication, the random variables $X_0 \sim \text{Exp}(2)$ and $X_1 \sim \text{Gamma}(3,2)$ were generated and the estimate of the ROC curve was computed. The probability distributions of both groups $\mathcal{G}_0$ and $\mathcal{G}_1$ are illustrated in Figure 2.

In all replications sample sizes of $n_0 = n_1 = 50$ were used. In this case, the actual global optimal bandwidths (see Azzalini, 1981) for F_0 and F_1 are $h_{F_0} = 2.9149$ and $h_{F_1} = 5.8298$, respectively. For the kernel estimation of the cumulative distributions we used the quartic kernel $K(x) = \frac{15}{16}(1-x^2)^2 I_{[-1,1]}$, where I_A is the indicator function on the set A. In our experience, the quality of estimated curve by using this kernel is not too sensitive to an optimal bandwidth choice. Hence we used this kernel also in the next section.

For each ROC curve we have calculated the mean integrated squared error (MISE) on the interval $[0,1]$ over all 1000 replications and have displayed the results in a boxplot in Figure 3. The variance of each estimator can be accurately gauged by the whiskers of the plot. The values of means and standard deviations for MISE of each method are given in Table 1.

We also obtained 10 typical realizations of each estimator and displayed these in Figure 4 for comparison purposes with the theoretical ROC curve. The solid line represents the theoretical ROC curve and the dotted lines illustrate the 10 realizations.

The final estimate of the ROC curve depends on estimates of the cumulative distribution functions F_0 and F_1. While boundary effects cause problems by estimating F_0 and

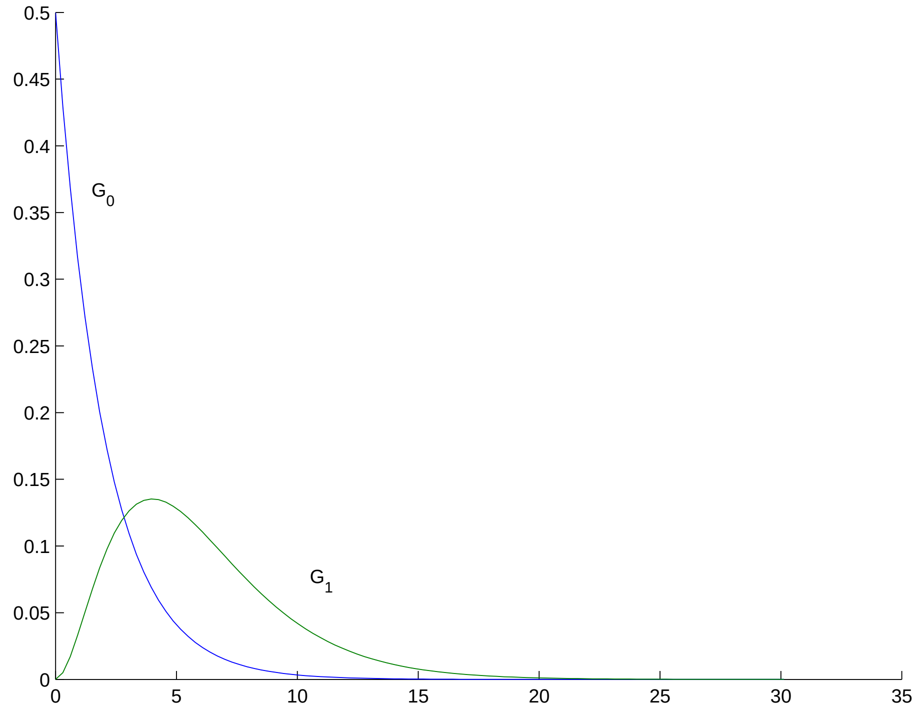


Figure 2: The probability distribution of groups $\mathcal{G}_0$ and $\mathcal{G}_1$.

Table 1: Means and standard deviations of the MISE.

Method	Mean	STD
Proposed	0.0053	0.0047
Reflection	0.0065	0.0050
Classical	0.0084	0.0054

F_1 inside the left boundary region, the quality of the final estimate of the ROC can also be influenced by these effects near the right boundary of the interval $[0, 1]$ as well. As we can see in Figure 4, the biggest difference between the above mentioned methods is in the second half part of the interval $[0, 1]$. Table 1 describes the performance of our proposed method with respect to the MISE. The values of the mean and the standard deviation for the MISE were smallest in case of our proposed estimator. Although the theoretical bias of our estimator is of the same order as in the case of the reflection method, the numerical results of estimators of the ROC curves were better for our estimator in the simulation. In our opinion, this is due to the fact that our estimator is locally adaptive.

5 Consumer Loans Data

In this example we used some (unspecified) scoring function to predict the solidity of a client. The goal here is to determine which clients are able to pay their loans. We considered a test set of 332 clients; 309 paid their loans (group $\mathcal{G}_0$) and 22 had problems with

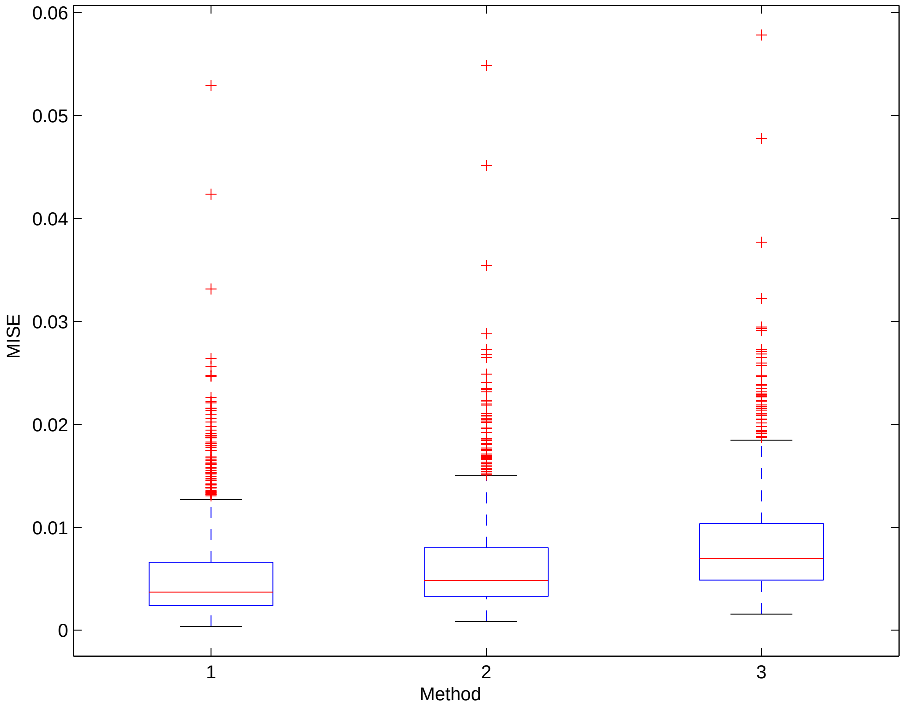
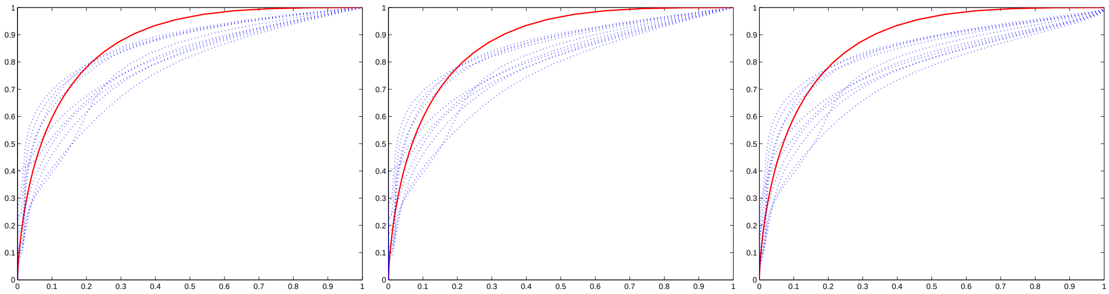


Figure 3: Boxplots of the MISE over $[0, 1]$ for our proposed method (1), the reflection method (2), and the classical estimator with boundary effects (3).



(1) (2) (3)

Figure 4: Estimates of the ROC for our proposed method (1), the reflection method (2), and the classical estimator with boundary effects (3).

payments or did not pay (group $\mathcal{G}_1$). We used the ROC curve to assess the discrimination between clients with and without a good solidity. It is of interest for us to know here if our scoring function is a good predictor of the solidity.

Estimates of ROC are illustrated in Figure 5. The dashed line represents the estimate obtained by our proposed method and the solid line is for the kernel ROC with boundary effects. When choosing the optimal bandwidths for distribution function estimation, we used the method described in Horová, Koláček, Zelinka, and El-Shaarawi (2008). A somewhat similar method for density estimation is given in Sheather and Jones (1991). The optimal bandwidths for distribution functions F_0 and F_1 were estimated as $\hat{h}_{F_0} =$

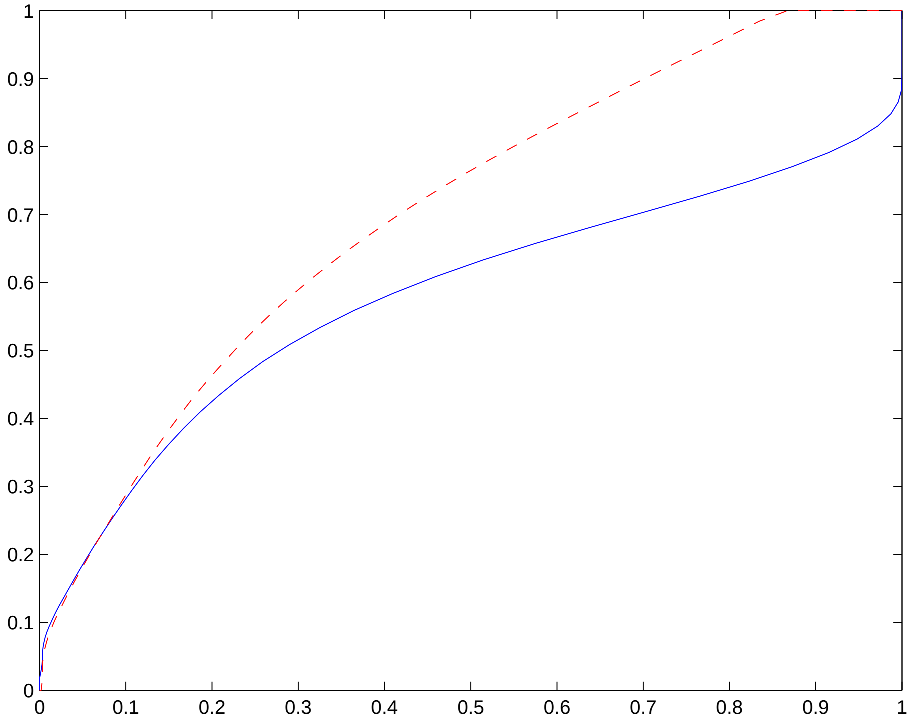


Figure 5: The estimate of the ROC for consumer the loans data.

0.0068 and $\hat{h}_{F_1} = 0.0286$, respectively.

From the estimates of the ROC one can see that the scoring function is not a good predictor of the solidity of a client. This fact could be also affected by the different sizes of both groups. When group $\mathcal{G}_1$ is too small it causes larger boundary effects. It is clearly visible that the estimate of the ROC obtained by the classical estimator (solid line) has some values under the diagonal of the unit square. However, this situation does not show up theoretically. Thus there is a larger influence of boundary effects to the quality of final estimates of the ROC.

6 Conclusion

In this paper we proposed a new kernel-type distribution estimator to avoid the difficulties near the boundary. The technique implemented is a kind of generalized reflection method involving reflecting a transformation of the data. The proposed method generates a class of boundary corrected estimators and it is based on ideas of boundary corrections for kernel density estimators presented in Karunamuni and Alberts (2005a, 2005b, 2006). We showed some good properties of our proposed method (e.g., local adaptivity). Furthermore, it is shown that bias of the proposed estimator is smaller than that of the “classical” case.

Acknowledgements

The research was supported by the Jaroslav Hájek center for theoretical and applied statistics (grant No. LC 06024). The second author's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix

Proof of (4). For $x = ch$, $0 \le c \le 1$, using the property $W(t) = 1 - W(-t)$ we obtain

$$\begin{aligned}
\mathrm{E}(\widetilde{F}_{h,K}(x)) &= \mathrm{E}\left(W\left(\frac{x - g_1(X_i)}{h}\right)\right) - \mathrm{E}\left(W\left(-\frac{x + g_2(X_i)}{h}\right)\right)\\
&= \int_0^\infty W\left(\frac{x - g_1(y)}{h}\right) f(y)dy - \int_0^\infty W\left(-\frac{x + g_2(y)}{h}\right) f(y)dy\\
&= h\int_{-1}^{c} W(t)\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt - h\int_{-1}^{-c} W(t)\frac{f\left(g_2^{-1}((-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}dt\\
&= h\int_{-1}^{-c} W(t)\left\{\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} - \frac{f\left(g_2^{-1}((-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}\right\}dt\\
&\quad + h\int_{-c}^{c} (1 - W(-t))\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt\\
&= h\int_{-1}^{-c} W(t)\left\{\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} - \frac{f\left(g_2^{-1}(-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}\right\}dt\\
&\quad + F\left(g_1^{-1}(2ch)\right) - h\int_{-c}^{c} W(t)\frac{f\left(g_1^{-1}((c+t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c+t)h)\right)}dt\,.
\end{aligned}$$

Using a Taylor expansion of order 2 on the function $F\left(g_1^{-1}(\cdot)\right)$ we have

$$F\left(g_1^{-1}(2ch)\right) = F(0) + f(0)2ch + \left(f^{(1)}(0) - f(0)g_1^{(2)}(0)\right)2c^2h^2 + o(h^2)\,.$$

By the existence and continuity of $F^{(2)}(\cdot)$ near 0, we obtain for $x = ch$

$$\begin{aligned}
F(0) &= F(x) - f(x)ch + \frac{1}{2}f^{(1)}(x)c^2h^2 + o(h^2)\\
f(x) &= f(0) + f^{(1)}(0)ch + o(h)\\
f^{(1)}(x) &= f^{(1)}(0) + o(1)\,.
\end{aligned}$$

Therefore,

$$F\left(g_1^{-1}(2ch)\right) = F(x) + f(0)ch + \left(\frac{3}{2}f^{(1)}(0) - 2f(0)g_1^{(2)}(0)\right)c^2h^2 + o(h^2)\,. \tag{7}$$

Now, (7) and a Taylor expansion of order 1 of the functions

$$\frac{f\left(g_1^{-1}(\cdot)\right)}{g_1^{(1)}\left(g_1^{-1}(\cdot)\right)} \quad \text{and} \quad \frac{f\left(g_2^{-1}(\cdot)\right)}{g_2^{(1)}\left(g_2^{-1}(\cdot)\right)}$$

give

$$
\begin{aligned}
&\mathrm{E}\left(\widetilde{F}_{h,K}(x)\right) - F(x)\\
&= h\int_{-1}^{-c} W(t)\left\{2f^{(1)}(0)ch - f(0)h\left((c-t)g_1^{(2)}(0) + (c+t)g_2^{(2)}(0)\right) + o(h)\right\} dt\\
&\quad + f(0)ch + \left\{\frac{3}{2}f^{(1)}(0) - 2f(0)g_1^{(2)}(0)\right\} c^2h^2 + o(h^2)\\
&\quad - h\int_{-c}^{c} W(t)\left\{f(0) + \left(f^{(1)}(0) - f(0)g_1^{(2)}(0)\right)(c+t)h + o(h)\right\} dt\\
&= h\left\{f(0)c - f(0)\int_{-c}^{c} W(t)dt\right\} + h^2\left\{\frac{3}{2}f^{(1)}(0)c^2 + 2f^{(1)}(0)c\int_{-1}^{-c} W(t)dt\right.\\
&\quad - 2f(0)g_1^{(2)}(0)c^2 - f(0)g_1^{(2)}(0)\int_{-1}^{-c}(c-t)W(t)dt - f(0)g_2^{(2)}(0)\int_{-1}^{-c}(c+t)W(t)dt\\
&\quad \left. - \left(f^{(1)}(0) - f(0)g_1^{(2)}(0)\right)\int_{-c}^{c}(c+t)W(t)dt\right\} + o(h^2)\,.
\end{aligned}
$$

From the symmetry of K and the definition $W(x)$, one can write $W(x) = \frac{1}{2} + b(x)$, where $b(x) = -b(-x)$ for all x such that $|x| \le 1$. Thus $\int_{-c}^{c} W(t)dt = c$ and therefore the coefficient of h is zero. So after some algebra we obtain the bias expression as

$$
\begin{aligned}
\mathrm{E}\left(\widetilde{F}_{h,K}(x)\right) - F(x) = h^2 &\left\{f^{(1)}(0)\left(\frac{c^2}{2} + 2c\int_{-1}^{-c} W(t)dt - \int_{-c}^{c} tW(t)dt\right)\right.\\
&\left. -f(0)g_1^{(2)}(0)\int_{-1}^{c}(c-t)W(t)dt - f(0)g_2^{(2)}(0)\int_{-1}^{-c}(c+t)W(t)dt\right\} + o(h^2)\,.
\end{aligned}
$$

Proof of (5). Observe that for $x = ch$, $0 \le c \le 1$, we have

$$
\begin{aligned}
n\mathrm{var}\left(\widetilde{F}_{h,K}(x)\right) &= \frac{1}{n}\mathrm{var}\left\{\sum_{i=1}^{n}\left[W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right]\right\}\\
&= \mathrm{E}\left\{W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right\}^2\\
&\quad - \left\{\mathrm{E}\left[W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right]\right\}^2\\
&= A_1 - A_2\,,
\end{aligned}
$$

where

$$
\begin{aligned}
A_1 &= \mathrm{E}\left\{W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right\}^2 \\
&= \int_0^\infty \left\{W\left(\frac{x-g_1(y)}{h}\right) - W\left(-\frac{x+g_2(y)}{h}\right)\right\}^2 f(y)dy \\
&= \int_0^\infty \left\{W^2\left(\frac{x-g_1(y)}{h}\right) + W^2\left(-\frac{x+g_2(y)}{h}\right)\right\} f(y)dy \\
&\quad - \int_0^\infty 2W\left(\frac{x-g_1(y)}{h}\right) W\left(-\frac{x+g_2(y)}{h}\right) f(y)dy \\
&= h\int_{-1}^{-c} W^2(t)\left\{\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} + \frac{f\left(g_2^{-1}((-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}\right\}dt \\
&\quad + h\int_{-c}^{c} W^2(t)\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt \\
&\quad - \int_0^\infty 2W\left(\frac{x-g_1(y)}{h}\right) W\left(-\frac{x+g_2(y)}{h}\right) f(y)dy \\
&= A_{1,1} + A_{1,2} - A_{1,3}\,.
\end{aligned}
$$

Using a Taylor expansion as in the last proof, it can be shown that

$$
\begin{aligned}
A_{1,1} &= h\int_{-1}^{-c} W^2(t)\left\{\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} + \frac{f\left(g_2^{-1}((-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}\right\}dt \\
&= h\int_{-1}^{-c} W^2(t)\left(2f(0) + o(1)\right)dt\,.
\end{aligned}
$$

For $A_{1,2}$ we use the identity $W(t) = 1 - W(-t)$ and similarly as in the last proof we get

$$
\begin{aligned}
A_{1,2} &= h\int_{-c}^{c} W^2(t)\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt \\
&= h\int_{-c}^{c} \left(1 - 2W(-t) + W^2(-t)\right)\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt \\
&= h\int_{-c}^{c} \frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt - 2h\int_{-c}^{c} W(t)\frac{f\left(g_1^{-1}((c+t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c+t)h)\right)}dt \\
&\quad + h\int_{-c}^{c} W^2(t)\frac{f\left(g_1^{-1}((c+t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c+t)h)\right)}dt \\
&= F\left(g_1^{-1}(2ch)\right) - 2h\int_{-c}^{c} W(t)\left(f(0) + o(1)\right)dt + h\int_{-c}^{c} W^2(t)\left(f(0) + o(1)\right)dt \\
&= F(x) - f(0)ch + hf(0)\int_{-c}^{c} W^2(t)dt + o(h)\,.
\end{aligned}
$$

Using the continuity of $g_i^{(2)}$, $g_i(0) = 0$, and $g_i^{(1)}(0) = 1$, $i = 1, 2$, and by a Taylor expansion of order 2 on $g_2\left(g_1^{-1}(\cdot)\right)$, we have

$$\begin{aligned}
g_2\left(g_1^{-1}((c-t)h)\right) &= g_2\left(g_1^{-1}(0)\right) + \frac{g_2^{(1)}\left(g_1^{-1}(0)\right)}{g_1^{(1)}\left(g_1^{-1}(0)\right)}(c-t)h + o(h) \\
&= (c-t)h + o(h)\,.
\end{aligned}$$

With the preceding expansion we obtain

$$\begin{aligned}
A_{1,3} &= \int_0^{\infty} 2W\left(\frac{x-g_1(y)}{h}\right) W\left(-\frac{x+g_2(y)}{h}\right) f(y)dy \\
&= 2h\int_{-1}^{c} W(t)W\left(-\frac{x}{h} - \frac{g_2\left(g_1^{-1}((c-t)h)\right)}{h}\right) \frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt \\
&= 2h\int_{-1}^{c} W(t)W\left(\frac{-ch-(c-t)h-o(h)}{h}\right)\left(f(0)+o(1)\right)dt \\
&= 2hf(0)\int_{-1}^{c} W(t)W(t-2c)dt + o(h)\,.
\end{aligned}$$

Now we can express A_1 as

$$\begin{aligned}
A_1 &= A_{1,1} + A_{1,2} - A_{1,3} \\
&= 2hf(0)\int_{-1}^{-c} W^2(t)dt + F(x) - f(0)ch + hf(0)\int_{-c}^{c} W^2(t)dt \\
&\quad -2hf(0)\int_{-1}^{c} W(t)W(t-2c)dt + o(h) \\
&= F(x) + hf(0)\left\{2\int_{-1}^{-c} W^2(t)dt - c + \int_{-c}^{c} W^2(t)dt - 2\int_{-1}^{c} W(t)W(t-2c)dt\right\} \\
&\quad +o(h)\,.
\end{aligned}$$

With the expression obtained for the bias we obtain the expression for A_2 as

$$\begin{aligned}
A_2 &= \left\{\mathrm{E}\left[W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right]\right\}^2 \\
&= \left\{\mathrm{E}\left(\widetilde{F}_{h,K}(x)\right)\right\}^2 \\
&= F^2(x) + o(h)\,.
\end{aligned}$$

Finally, we obtain the variance of the estimator as

$$\begin{aligned}
& n\mathrm{var}\left(\widetilde{F}_{h,K}(x)\right) = A_1 - A_2 \\
&= F(x) + hf(0)\left\{2\int_{-1}^{-c} W^2(t)dt - c + \int_{-c}^{c} W^2(t)dt - 2\int_{-1}^{c} W(t)W(t-2c)dt\right\} \\
&\quad -F^2(x) + o(h) \\
&= F(x)(1-F(x)) \\
&\quad +hf(0)\left\{2\int_{-1}^{-c} W^2(t)dt - c + \int_{-c}^{c} W^2(t)dt - 2\int_{-1}^{c} W(t)W(t-2c)dt\right\} + o(h)\,.
\end{aligned}$$

References

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

Horová, I., Koláček, J., Zelinka, J., and El-Shaarawi, A. H. (2008). Smooth estimates of distribution functions with application in environmental studies. *Advanced topics on mathematical biology and ecology*, 122-127.

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

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

Karunamuni, R. J., and Alberts, T. (2006). A locally adaptive transformation method of boundary correction in kernel density estimation. *Journal of Statistical Planning and Inference*, *136*, 2936-2960.

Karunamuni, R. J., and Zhang, S. (2008). Some improvements on a boundary corrected kernel density estimator. *Statistics & Probability Letters*, *78*, 497-507.

Lejeune, M., and Sarda, P. (1992). Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis*, *14*, 457-471.

Lloyd, C. J. (1998). The use of smoothed ROC curves to summarise and compare diagnostic systems. *Journal of the American Statistical Association*, *93*, 1356-1364.

Lloyd, C. J., and Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical. *Statistics and Probability Letters*, *44*, 221-228.

Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory of Probability and its Application*, *15*, 497-500.

Reiss, R. D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, *8*, 116-119.

Sheather, S. J., and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, *53*, 683-690.

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

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

Zhang, S., and Karunamuni, R. J. (1998). On kernel density estimation near endpoints. *J. Statist. Planning and Inference*, *70*, 301–316.

Zhang, S., and Karunamuni, R. J. (2000). On nonparametric density estimation at the boundary. *Nonparametric Statistics*, *12*, 197–221.

Zhang, S., Karunamuni, R. J., and Jones, M. C. (1999). An improved estimator of the density function at the boundary. *Journal of the American Statistical Association*, *94*, 1231–1241.

Authors' addresses:

Jan Koláček
Department of Mathematics and Statistics
Faculty of Science
Kotlářská 2
611 37 Brno
Czech Republic

E-Mail: kolacek@math.muni.cz

Rohana J. Karunamuni
Department of Mathematical and Statistical Sciences
University of Alberta
T6G 2G1 Edmonton
Canada

E-Mail: R.J.Karunamuni@ualberta.ca