Journal of Statistical Modelling: Theory and Applications
Vol. 5, No. 1, 2024, pp. 203-215
Yazd University Press 2024

JSMTA

Research Paper

Boundary modified kernel estimator for the ROC curve

Habiballah Mombeni, Behzad Mansouri* and Mohammad Reza Akhoond
Department of Statistics, Faculty of Mathematics and Computer Sciences,
Shahid Chamran University of Ahvaz, Ahvaz, Iran

Received: September 10, 2024/ Revised: December 19, 2024/ Accepted: January 04, 2025

Abstract: The receiver operating characteristic curve is a simple graphical tool used to assess the accuracy of diagnostics tests. Pulit (2016) proposed an innovative approach for estimating the receiver operating characteristic curves based on kernel smoothing. Although his proposed estimator is highly appealing in several aspects, it suffers from the well-known boundary bias effect. In this paper, we highlight this drawback and propose a new modified estimator that uses an appropriate boundary kernel. The asymptotic convergence of the proposed estimator at boundary points is demonstrated. Using both simulated and real data sets, we illustrate the performance of the proposed estimator. The results show that the proposed estimator outperforms not only the Pulit's estimator but also other commonly used estimators.

Keywords: Boundary effects, Distribution function, Kernel-type estimator, Receiver operating characteristic curve.
Mathematics Subject Classification (2010): 62G30, 62G05.

1 Introduction

A receiver operating characteristic (ROC) curve is a simple graphical tool employed for evaluating and comparing the accuracy of discrimination rules. Today, many scholars have shown the effectiveness of the ROC curves in a variety of scientific fields such as psychology, medicine, and machine learning, to name a few (Fawcett, 2006).

Let us consider a discrimination task to assign an individual into two separate groups based on a continuous measured score. For the sake of simplifying, assume a medical situation in which two groups are the healthy and the diseased groups.

*Corresponding author:j.estabraqi@yahoo.com

In addition, suppose that X and Y as the scores in the healthy and the diseased groups are random variables with absolutely continuous unknown cumulative distribution functions of $F(x)$ and $G(y)$, respectively. Based on the discrimination rule, an individual is assigned into the diseased groups if his or her score is greater than a cut-off point c, $-\infty \leq c \leq \infty$ and into the healthy group, otherwise. This corresponds to a hypothesis testing in which the null hypothesis (H_0) is "the individual is healthy" and the alternative hypothesis (H_1) is "the individual is diseased". The probability of type I error is $\alpha = P(\text{reject } H_0 | \text{ the individual is diseased}) = P_{H_0}(X > c) = 1 - F(c)$ and the probability of type II error is $\beta = P(\text{accept } H_0 | \text{ the individual is diseased}) = P_{H_1}(Y < c) = G(c)$. The sensitivity of the test is defined as $SE(c) = 1 - \beta = 1 - G(c)$ and the specificity of the test is defined by $SP(c) = 1 - \alpha = F(c)$. The ROC graph is a two-dimensional graph in which $SE(c)$ is plotted on the vertical axis and $\alpha = 1 - SP(c)$ is plotted on the horizontal axis. Note that for all possible cut-off points c, $0 \leq SE(c) \leq 1$ and $0 \leq SP(c) \leq 1$, the ROC curve depicts $SE(c)$ versus $1 - SP(c)$ (see Fawcett, 2006, for more details and some useful discussions). Let $t = 1 - SP(c)$ or $c = F^{-1}(1 - t)$ then we have

$$R(t) = SE(c) = SE(F^{-1}(1-t)) = 1 - G(F^{-1}(1-t)), \qquad 0 \leq t \leq 1.$$

A highly important issue would be the estimation of the ROC curve based on random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from the two populations, i.e. the healthy group and the diseased, respectively. The empirical ROC curve is a natural choice; however, it would not be smooth. Another method for estimating the ROC curve is to use kernel-type estimators. Zou et al. (1997); Lloyd (1998) and Lloyd and Yong (1999) are among pioneers in this field. However, kernel-type estimators have their own restrictions and weaknesses one of which is that they are not invariant under monotone data transformations. In order to remedy this drawback, Pulit (2016) proposed an innovative kernel-type estimator for the ROC curve which is not only invariant under non-decreasing data transformations, but it also has a single smoothing parameter. Although Pulit's estimator has its own advantages (see Pulit (2016) for more details), it suffers from a boundary effect near the boundary points, which is due to using a symmetric kernel in estimating the distribution function $G(\cdot)$ (see Section 2 of this paper). In kernel estimation, boundary effects are quite known and several approaches have so far been proposed to deal with them in regression and density estimation tasks (Chen, 1999, 2000; Gasser and Müller, 1979; Gasser et al., 1985; Hirukawa and Sakudo, 2014, 2015; John, 1984; Müller, 1991; Zhang et al., 1999). Koláček and Karunamuni (2009) have considered the boundary effect in the ROC curve estimation and proposed a boundary-corrected estimator for the ROC curve. Their estimator is a combination of the reflection method and the transformation method, an approach which was originally introduced by Zhang et al. (1999) in boundary kernel density estimation. Although many researchers have considered the boundary effect in kernel density estimation, not many have shown interest in the analogy issue in kernel distribution estimation task. Tenreiro (2013) and Tenreiro (2018) proposed boundary kernels for estimating a cumulative distribution function with bounded support. The advantage of the approach suggested by Tenreiro is that it estimates cumulative distribution function directly while Koláček and Karunamuni (2009) estimate density and sum it up to provide an estimation for the cumulative distribution function.

In this paper, we have tried to exploit the advantages of both Pulit's method in ROC curve estimation and Tenreiro's approach in boundary correction. More precisely, one of the boundary kernels proposed by Tenreiro (2013) has been used to modify Pulit's estimator. The proposed estimator has no boundary problem when the advantages of Pulit's estimator are added.

The rest of the paper is organized as follows. In Section 2, we demonstrate the Pulit's estimator boundary problem, that the convergence rate of its bias in the boundary points is slower than the interior points. Section 3 introduces the proposed estimator. In this section, we will show the asymptotic superiority of the proposed estimator. In Section 4, we conduct a numerical study in order to illustrate the performance of our proposed estimator and compare it with some other frequently-used estimators. In Section 5, we applied our estimator to a set of real data which come from a clinical study. Finally, some conclusions and discussions are given in Section 6.

2 Pulit's estimator boundary bias

In this section, we briefly describe Pulit's estimator and argue that it suffers from a boundary bias problem. For two independent samples $\boldsymbol{X} = (X_1, \ldots, X_m)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ from two unknown distribution functions F and G with the same supports, i.e., $[0, \infty)$, Pulit (2016) has proposed estimating $R(t)$ based on the vector $\hat{\boldsymbol{Z}} = (1 - F_m(Y_1), \ldots, 1 - F_m(Y_n))$, where F_m denotes the empirical distribution function of the sample X_m. The Pulit's estimator is given by

$$\hat{R}_P(t) = n^{-1} \sum_{i=1}^{n} K\left(\frac{t - 1 + F_m(Y_i)}{h}\right),$$

where t is the design point, $K(u)$ equals $\int_{-1}^{u} k(s)ds$ for $u \in R$, $k(\cdot)$ is a bounded and symmetric probability density function with support $[-1, 1]$, and h is the smoothing parameter. Pulit (2016) used the Epanechnikov kernel $k(s) = \frac{3}{4}(1 - s^2)$, $-1 \leq s \leq 1$, which is the best choice among symmetric kernels (Silverman, 2018). In this paper, we denote a design point t by the 'boundary point' if $t = ch$ for some $c \in (0, 1)$ and by the 'interior point', otherwise. Pulit (2016), under some mild assumptions, shows that $\hat{R}_P$ is a consistent estimator. He has proved that $\hat{R}_P$ is asymptotically unbiased and, for the interior points, its bias is of order $o(h^2)$. However, in what follows we show that this is not the case for the boundary points. In fact, $\hat{R}_P$ suffers from a boundary problem in that for the boundary points, the bias of $\hat{R}_P$ is of order $O(h)$ rather than $o(h^2)$. Suppose that R_P is absolutely continuous and has two continuous and bounded derivatives with the smoothing parameter h which satisfies $h = h_n \longrightarrow 0$ as $n \longrightarrow \infty$. Then Pulit (2016) has shown that

$$\begin{aligned} E\left(\hat{R}_P(t)\right) &\approx E_{F,G}(K(T_1)) + \sum_{j=1}^{3} E_{F,G}(K^{(j)}(T_1)(T_{1,m} - T_1)^j) \\ &\approx I_0 + I_1 + I_2 + I_3, \end{aligned} \tag{1}$$

where $K^{(j)}(\cdot)$ is the j'th derivative of $K(\cdot)$ for $j = 1, 2, 3$ and $T_1 = \frac{t-1+F(Y_1)}{h_n}$, $T_{1,m} = \frac{t-1+F_m(Y_1)}{h_n}$ and

$$\begin{aligned}
I_0 &= E_{F,G}(K(T_1)) = K(\frac{t-1}{h}) + R(t)\int_{\frac{t-1}{h}}^{\frac{t}{h}} k(x)dx \\
&\quad -hR^{(1)}(t)\int_{\frac{t-1}{h}}^{\frac{t}{h}} xk(x)dx + \frac{h^2}{2}R^{(2)}(t)\int_{\frac{t-1}{h}}^{\frac{t}{h}} x^2k(x)dx + o(h^2). \\
I_1 &= E_{F,G}(K(T_1)(T_{1,m} - T_1)) = 0, I_2 = E_{F,G}(K^{(2)}(T_1)(T_{1,m} - T_1)^2) = O\left(\frac{1}{m}\right), \\
I_3 &= E_{F,G}(K^{(3)}(T_1)(T_{1,m} - T_1)^3) = O\left(\frac{1}{m^2h^2}\right).
\end{aligned}$$

See Pulit (2016) proof of Theorem 1. Now we consider the left boundary point $t = ch$ for some $c \in (0, 1)$. Since the support of the Epanechnikov kernel is the compact interval $[-1, 1]$, we can conclude

$$\frac{t-1}{h} < x < \frac{t}{h} \Rightarrow c - \frac{1}{h} < x < c.$$

So, $-1 < x < c$ as $h \to 0$. Now we have

$$\begin{aligned}
I_0 &= R(t)\int_{-1}^{c} k(x)dx - hR^{(1)}(t)\int_{-1}^{c} xk(x)dx + \frac{h^2}{2}R^{(2)}(t)\int_{-1}^{c} x^2k(x)dx + o(h^2) \\
&= R(t) + \left(\int_{-1}^{c} k(x)dx - 1\right)R(t) - hR^{(1)}(t)\int_{-1}^{c} xk(x)dx \\
&\quad + \frac{h^2}{2}R^{(2)}(t)\int_{-1}^{c} x^2k(x)dx + o(h^2) \\
&= R(t) + (\mu_{0,c}(k) - 1)R(t) - hR^{(1)}(t)\mu_{1,c}(k) + \frac{h^2}{2}R^{(2)}(t)\mu_{2,c}(k) + o(h^2), \qquad (2)
\end{aligned}$$

and $\mu_{I,c}(k) = \int_{-1}^{c} t^l k(t)dt$, for $l = 0, 1, 2$. Using Taylor expansion, we have

$$\begin{aligned}
R(t) \approx R(0) + tR_+^{(1)}(0) + \frac{t^2}{2}R_+^{(2)}(0) &\implies R(t) \approx chR_+^{(1)}(0) + \frac{(ch)^2}{2}R_+^{(2)}(0), \\
R^{(1)}(t) \approx R_+^{(1)}(0) + tR_+^{(2)}(0) &\implies R^{(1)}(t) \approx R_+^{(1)}(0) + chR_+^{(2)}(0), \\
R^{(2)}(t) \approx R_+^{(2)}(0) + tR_+^{(3)}(0) &\implies R^{(2)}(t) \approx R_+^{(2)}(0) + chR_+^{(3)}(0),
\end{aligned}$$

where $R_+^{(i)}(0)$ for $i = 1, 2$ is the i'th right-derivative of $R(t)$. By substituting these expressions in (2), we can conclude

$$\begin{aligned}
I_0 &= R(t) + (\mu_{0,c}(k) - 1)\left(chR_+^{(1)}(0) + \frac{(ch)^2}{2}R_+^{(2)}(0)\right) - h(R_+^{(1)}(0) \\
&\quad + chR_+^{(2)}(0))\mu_{1,c}(k) + \frac{h^2}{2}(R_+^{(2)}(0) + chR_+^{(3)}(0))\mu_{2,c}(k) + o(h^2)
\end{aligned}$$

$$\begin{aligned} &= R(t) + h(c(\mu_{0,c}(k)-1) - \mu_{1,c}(k))R_+^{(1)}(0) + \frac{h^2}{2}(c^2(\mu_{0,c}(k)-1) \\ &\quad -2c\mu_{1,c}(k) + \mu_{2,c}(k))R_+^{(2)}(0) + o(h^2) \\ &= R(t) + h\delta_1(c)R_+^{(1)}(0) + \frac{h^2}{2}\delta_2(c)R_+^{(2)}(0) + o(h^2). \end{aligned}$$

so at the boundary points we have

$$Bias(\hat{R}_P(t)) = hR_+^{(1)}(0)\delta_1(c) + \frac{h^2}{2}R_+^{(2)}(0)\delta_2(c) + O\left(\frac{1}{m} + \frac{1}{m^2h^2}\right) + o(h^2),$$

where

$$\delta_1(c) = c(\mu_{0,c}(k)-1) - \mu_{1,c}(k), \quad \delta_2(c) = c^2(\mu_{0,c}(k)-1) - 2c\mu_{1,c}(k) + \mu_{2,c}(k).$$

We can see that the order of convergence of $\hat{R}_P$ at the boundary points is different from that of the interior points. A similar result emerges for the right boundary points, i.e. for $t = 1 - ch$, $c \in (0,1)$. In this case, the limits of integrations are $-c$ to 1. In the case where $R_+^{(1)} = 0$, the order of bias at boundary points agrees with the classical ones. In order to remedy the drawback of Pulit's estimator and provide a boundary corrected estimator for the ROC curve, we have proposed a new estimator in the next section.

3 Boundary corrected estimator

In the previous section, we showed that the Pulit's estimator $\hat{R}_P$ has a boundary bias. This drawback is due to the fact that $\hat{R}_P$ uses a non-appropriate kernel function which assigns non-zero weights out of the support of $Z = 1 - F(Y)$ which is $[0,1]$. Tenreiro (2013) introduced some modified kernels for solving the boundary problem in the kernel-estimation of the cumulative distribution function. For the interior points, Tenreiro's estimator is just the ordinary one. However, for the left and right region points, Tenreiro proposed using special well-adjusted left and right boundary kernels (See Tenreiro (2013) for more details and Tenreiro (2018) for the extension of the approach). Our idea is to combine the two approaches introduced by Pulit (2016) and Tenreiro (2018) to achieve a boundary corrected kernel-type estimator for the ROC curve. Our proposed estimator enjoys the advantages of both approaches in that for the interior points, it agrees with the Pulit' estimator and, as to the boundary regions, it is corrected for bias. Our proposed estimator is

$$\tilde{R}(t) = \begin{cases} 0, & t \le 0, \\ n^{-1}\sum_{i=1}^n \tilde{K}_{c,h}(t - 1 + F_m(Y_i)), & 0 < t < 1, \\ 1, & t \ge 1, \end{cases}$$

where $0 < h < \frac{1}{2}$, and

$$c = \begin{cases} \dfrac{t}{h}, & 0 < t < h, \\ 1, & h \le t \le 1-h, \\ \dfrac{1-t}{h}, & h < t < 1, \end{cases}$$

Also $\tilde{K}_{c,h}(U) = \int_{-\infty}^{u} \tilde{k}_{c,h}(v)dv$, $\tilde{k}_{c,h}(v) = k(v/ch)/ch$, and h is the smoothing parameter and $k(\cdot)$ is a bounded and symmetric kernel with support $[-1, 1]$ such that it satisfies in the following conditions:

$$\int \tilde{k}_c(v)dv = 1, \qquad \int v\tilde{k}_c(v)dv = 0, \qquad \int v^2\tilde{k}_c(v)dv \neq 0,$$

where $\tilde{k}_c(v) = k(v/c)/c$ and $\tilde{k}_{c,h}(v) = \tilde{k}_c(v/h)/h$. Then, our estimator is:

$$\tilde{R}(t) = \begin{cases} 0, & t \leq 0, \\ n^{-1}\sum_{i=1}^{n} K\left(\dfrac{t-1+F_m(Y_i)}{t}\right), & 0 < t < h, \\ \hat{R}_P(t), & h \leq t \leq 1-h \\ n^{-1}\sum_{i=1}^{n} K\left(\dfrac{t-1+F_m(Y_i)}{1-t}\right), & h < t < 1, \\ 1, & t \geq 1. \end{cases}$$

Consider the left boundary point $t = ch$, $c \in (0, 1)$. If we use the Epanechnikov kernel with support $[-1, 1]$, then the integral limits $[\frac{t-1}{h}, \frac{t}{h}]$ are converted to $[-1, c] = [-1, -c] \cup [-c, c]$ as $h \longrightarrow 0$. Now we get

$$\begin{aligned} &\int_{-1}^{-c} \tilde{k}_c(v)dv = 0, \qquad \int_{-c}^{c} \tilde{k}_c(v)dv = 1, \qquad \int_{-1}^{-c} v\tilde{k}_c(v)dvdv = 0, \\ &\int_{-c}^{c} v\tilde{k}_c(v)dv = 0, \qquad \int_{-1}^{-c} v^2\tilde{k}_c(v)dv = 0, \qquad \int_{-c}^{c} v^2\tilde{k}_c(v)dv \neq 0. \end{aligned} \tag{3}$$

Suppose that $R^{(1)}(t)$ and $R^{(2)}(t)$ exists and continuous for $t \in (0, 1)$ and $h \to 0$ as $n \to \infty$. Then by substituting these key equations in Pulit's results, it is easy to see that in the boundary points, the bias of $\tilde{R}(\cdot)$ is of order $O(h^2)$. For example, consider I_0 in (2)

$$\begin{aligned} I_0 &= R(t)\int_{-1}^{c} \tilde{k}_c(x)dx - hR^{(1)}(t)\int_{-1}^{c} x\tilde{k}_c(x)dx + \frac{h^2}{2}R^{(2)}(t)\int_{-1}^{c} x^2\tilde{k}_c(x)dx + o(h^2) \\ &= R(t)\left\{\int_{-1}^{-c} \tilde{k}_c(x)dx + \int_{-c}^{c} \tilde{k}_c(x)dx\right\} - hR^{(1)}(t)\left\{\int_{-1}^{-c} x\tilde{k}_c(x)dx + \int_{-c}^{c} x\tilde{k}_c(x)dx\right\} \\ &\quad + \frac{h^2}{2}R^{(2)}(t)\left\{\int_{-1}^{-c} x^2\tilde{k}_c(x)dx + \int_{-c}^{-c} x^2\tilde{k}_c(x)dx\right\} + o(h^2) \\ &= R(t) + \frac{h^2}{2}R^{(2)}(t)\int_{-c}^{c} x^2\tilde{k}_c(x)dx + o(h^2). \end{aligned}$$

By investigating Pulit's proofs, it is easy to check that I_1, I_2 and I_2 in (1), remain unchanged. The variance of $\tilde{R}(\cdot)$ is slightly different from the variance of $\hat{R}_P(\cdot)$. While $\tilde{R}(\cdot)$ is designed such that the calculations in the boundary and the interior points are identical, this is not the case for $\hat{R}_P(\cdot)$. More precisely, the variance of $\hat{R}_P(\cdot)$ which is provided by Pulit (2016) (see Pulit (2016), Equation 7) is valid only in the interior points. However, due to Equations in (3), the variance of $\tilde{R}(\cdot)$ in both the boundary

and the interior points agree and are identical with the variance of $\hat{R}_P(\cdot)$ in the interior points. Since the calculations are very cumbersome and do exist in Pulit, we only show briefly how the variance of $\hat{R}_P(\cdot)$ is different in the boundary points. Pulit[1] showed that $Var\left(\hat{R}_P(t)\right) := \frac{1}{n}J_1 + \frac{n-1}{n}J_2$ where $J_1 = J_{1,0} + \Sigma_{k=1}^{6} J_{1,k} - R^2(t) + O(h^2 + \frac{1}{m} + \frac{1}{m^2h^2} + \frac{1}{m^4h^4})$ and $\Sigma_{k=1}^{6} J_{1,k} = O(\frac{1}{mh} + \frac{1}{m^2h^3} + \frac{1}{m^3h^5})$ and J_2 is provided by Pulit (2016), Equation 27. For our purpose, we concentrate on $J_{1,0}$ which is

$$J_{1,0} = K^2\left(\frac{t-1}{h}\right) + 2R(t)\int_{\frac{t-1}{h}}^{\frac{t}{h}} k(x)K(x)dx - 2hR^{(1)}\int_{\frac{t-1}{h}}^{\frac{t}{h}} xk(x)K(x)dx + o(h),$$

Pulit (2016) in Equation 25 showed that $J_{1,0} = R(t) - \frac{9}{35}R^{(1)}(t)h + o(h)$. Consider the left boundary point $t = ch$ for some $c \in (0,1)$ and let $\gamma_{l,c}(k) = 2\int_{-1}^{c} t^l k(t)K(t)dt$, for $l = 0, 1$ and $\rho_1(c) = (c(\gamma_{0,c}(k) - 1) - \gamma_{1,c}(k))$ Now we have

$$\begin{aligned} J_{1,0} &= 2R(t)\int_{-1}^{c} k(x)K(x)dx - 2hR^{(1)}(t)\int_{-1}^{c} xk(x)K(x)dx + o(h) \\ &= R(t) + (\gamma_{0,c}(k) - 1)\,R(t) - hR^{(1)}(t)\gamma_{1,c}(k) + o(h) \\ &= R(t) + h\left(c(\gamma_{0,c}(k) - 1) - \gamma_{1,c}(k)\right)R_+^{(1)}(0) + o(h) \\ &= R(t) + \rho_1(c)R_+^{(1)}(0)h + o(h). \end{aligned}$$

For the Epanechnikov kernel, $\rho_1(c)$ is

$$\rho_1(c) = \frac{1}{112}c^7 - \frac{3}{40}c^5 - \frac{1}{16}c^4 + \frac{3}{16}c^3 + \frac{3}{8}c^2 - \frac{3}{4}c + \frac{33}{560}.$$

Note that $\rho_1(c)$ is a decreasing function of $c \in (0,1)$ and its smallest value is $\frac{-9}{35}$ for $c = 1$ which is the case for the interior points. Although $J_{0,1}$ is slightly different in the boundary points, it has the same order of h in both the boundary and the interior regions. An analogical investigation could be run for $J_{1,k}$, $k = 1, \ldots, 6$ and J_2.

To sum-up this short discussion, unlike $\hat{R}_P(\cdot)$, the boundary-modified kernel estimator $\tilde{R}(\cdot)$ is designed such that it has the same rate of convergence and the same variance in both the boundary and the interior regions.

4 Numerical study

In this section, we illustrate the performance of our proposed estimator through a simulation study. We compared our proposed estimator with those of Lloyd (1998), Pulit (2016) and K-K (Koláček and Karunamuni, 2009). We used Epanechnikov kernel in all the mentioned estimators. In the proposed estimator, we chose the smoothing parameter using the Beta-reference method proposed by Tenreiro (2013). To choose the smoothing parameter in both Lloyd estimator and Pulit estimator, we used the method proposed by Altman and Leger (1995). Finally, for K-K estimator, we chose the smoothing parameter by the method proposed by Horová et al. (2008). The simulations and plots in this paper were carried out using MATLAB software.

We considered five different cases for combining distributions (F, G) including: a: (Beta(1,1), Beta(3,1)), b: (Gamma(1,2), Gamma(3,2)), c: (Normal(0,1), Normal(1,1)),

Table 1: The mean and standard deviation of the ISE (100 repetitions) in estimating ROC curve via four methods and four different sample sizes (see the text for details).

$(\times 10^{-3})$	Method	cases	case 1	case 2	case 3	case 4	case 5
$(50, 50)$	LIoyd	Mean	5.5108	4.7870	5.0554	6.9369	7.0047
		Std.	4.5171	3.6722	4.4036	4.1141	6.6565
	K-K	Mean	5.5101	4.7124	5.0609	6.7639	7.0338
		Std.	4.5171	3.6280	4.4083	4.0806	6.7018
	Pulit	Mean	5.9723	4.9410	5.3311	4.9179	6.2849
		Std.	5.3245	6.3664	4.2310	4.3733	6.0307
	Proposed	Mean	4.8426	3.8930	4.4208	3.9554	5.8752
		Std.	4.4474	3.8473	3.7942	3.8693	6.2042
$(100, 100)$	Lloyd	Mean	2.6137	2.9466	2.1251	4.6909	4.6658
		Std.	2.0991	2.0730	2.1525	2.6196	4.3771
	K-K	Mean	2.6136	2.9283	2.1247	4.5719	4.6705
		Std.	2.0991	2.0629	2.1537	2.6352	4.7439
	Pulit	Mean	2.6310	2.6596	2.4381	2.9556	3.9979
		Std.	2.2475	2.8549	2.0611	2.5191	4.2617
	Proposed	Mean	2.2939	2.0219	1.9331	2.3280	3.6485
		Std.	2.0989	2.0097	1.8359	2.1759	4.0418
$(200, 200)$	Lloyd	Mean	1.7585	1.4818	1.5014	2.3388	1.8644
		Std.	1.6334	0.9492	1.4481	1.0143	2.0595
	K-K	Mean	1.7585	1.4760	1.5011	2.2862	1.8649
		Std.	1.6334	0.9453	1.4481	1.0068	2.0600
	Pulit	Mean	1.6098	1.1876	1.7231	1.2749	1.5433
		Std.	1.6740	0.8173	1.4743	1.0478	1.8794
	Proposed	Mean	1.4029	0.8133	1.3337	0.9655	1.4297
		Std.	1.6224	0.7191	1.2798	0.9746	1.7763

d: (lognormal(0,1), lognormal(1,1)) and e: (lognormal(1,1), gama(3,2)). In each case, we generated 200 samples of three different sample sizes $(n, m) = \{(50, 50), (100, 100), (200, 200)\}$. In all cases, we used the maximum likelihood method to estimate the unknown parameters. To evaluate and compare the performance of the estimators, we considered the integrated squared error $ISE_i = \int_0^\infty (\hat{R}_i(t) - R(t))^2 dt$ as the error metric, where $\hat{R}_i(x)$, $i = 1, 2, 3, 4$ stands for the ROC curve estimated by the proposed estimator, Lloyd estimator, Pulit estimator, and K-K estimator, respectively. In our setting, we approximate the integral as follow $ISE_i = \frac{1}{100} \sum_{j=1}^{100} (\hat{R}_i(t_j) - R(t_j))^2$.

Table 1 shows the mean and standard deviation $(\times 10^{-3})$ of the ISE in 100 repetitions for different combinations of distributions of X and Y. The simulation results show that, in almost all cases, based on the mean of ISE, our proposed estimator outperforms the other three estimators. Figure 4-5 display 30 estimates of the ROC curve (dotted curves) along with the true ROC curve (bold curve) for the four different cases ($m = 100$ and $n = 100$) via four methods. In all figures, the boundary problem of Pulit estimator is obvious. On the other hand, both Lloyd estimator and K-K estimator suffer from under-estimation. This drawback is clear especially, in Cases 2, 3 and 4. Lloyd (1998) has accepted this drawback and has confirmed that when $R(t)$ is convex, his approach under-estimates the ROC curves. In general, the performance of the proposed estimator in this study is satisfactory.

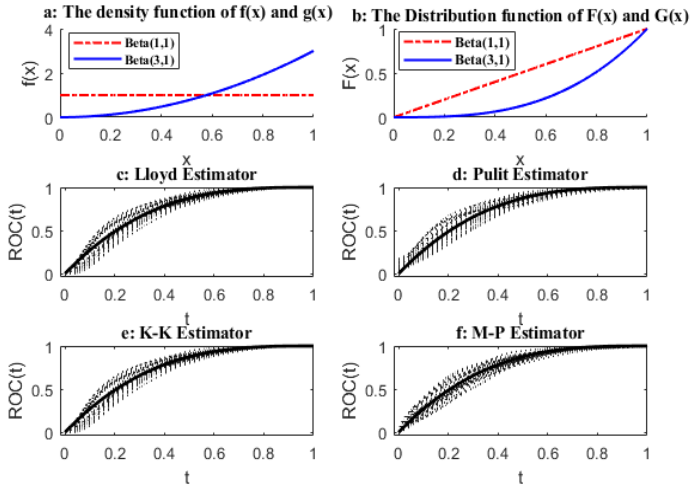

Figure 1: Display 30 estimates of the ROC curve (dotted curves) along with the true ROC curve (bold curve) where F and G are Beta (1,1) and Beta (3,1), respectively ($m = 100$ and $n = 100$) via four methods.

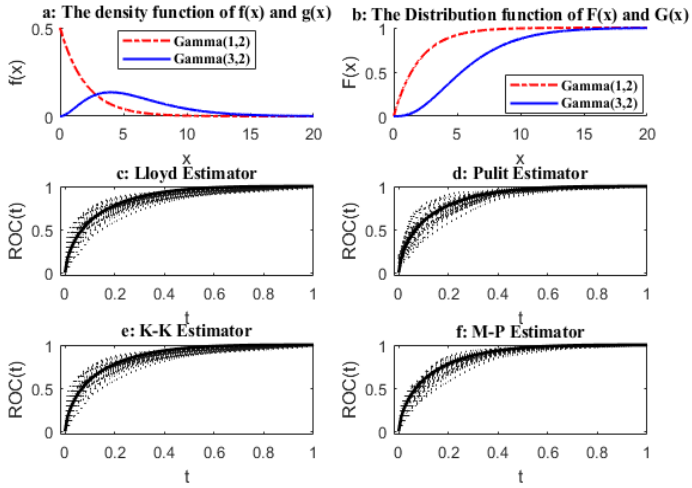

Figure 2: Display 30 estimates of the ROC curve (dotted curves) along with the true ROC curve (bold curve) where F and G are Gamma (1,2) and Gamma (3,2), respectively ($m = 100$ and $n = 100$) via four methods.

5 Real data analysis

To illustrate the performance of our estimator, we applied it to a set of real data which come from a clinical study. The data set results from research done by Turck et al. (2010) to identify the prognostic factors, to predict patient outcomes in patients with an aneurysmal subarachnoid hemorrhage. The data set involves clinical scores

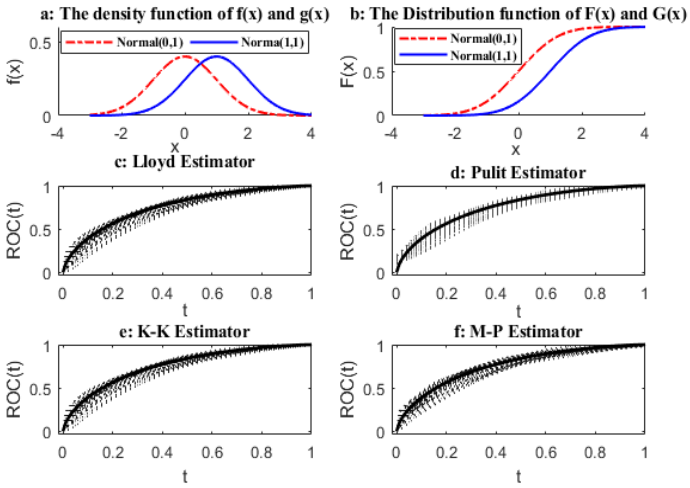

Figure 3: Display 30 estimates of the ROC curve (dotted curves) along with the true ROC curve (bold curve) where F and G are Normal (0,1) and Normal (1,1), respectively ($m = 100$ and $n = 100$) via four methods.

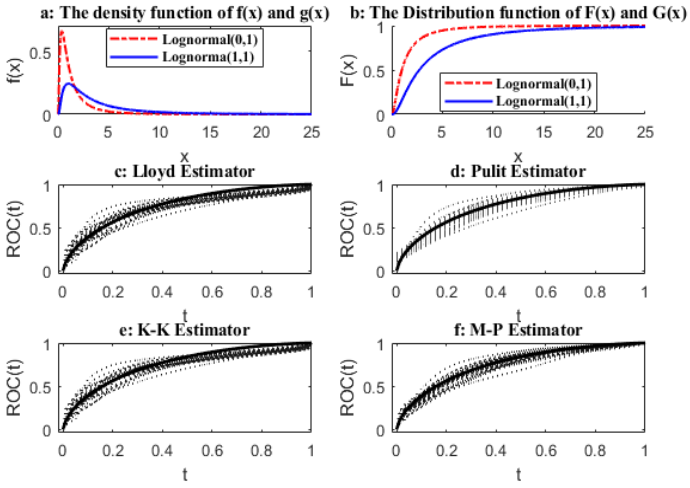

Figure 4: Display 30 estimates of the ROC curve (dotted curves) along with the true ROC curve (bold curve) where F and G are Lognormal $(0, 1)$ and Lognormal $(1, 1)$, respectively ($m = 100$ and $n = 100$) via four methods.

together with several predictive factors (brain injury-related biomarkers): H-FABP, NDKA, UFD1 and S100 β. We take the data from the R package pROC (Robin et al., 2011) which summarizes this data set as "aSAH" for 113 patients. The ROC curve estimators for S100β as the predictive factor is plotted in Figure 6. The estimators are our proposed estimator, the Lloyd estimator, Pulit estimator, K-K estimator and the

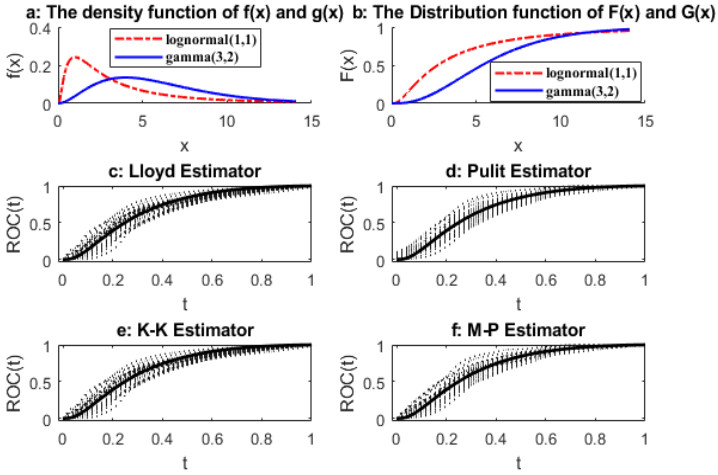

Figure 5: Display 30 estimates of the ROC curve (dotted curves) along with the true ROC curve (bold curve) where F and G are Lognormal $(1, 1)$ and Gamma $(3, 2)$, respectively ($m = 100$ and $n = 100$) via four methods.

empirical estimator. While the empirical ROC curve provides a discontinuous estimate, Pulit estimator suffers from boundary bias. In addition, the under-estimation in both Lloyd and K-K estimators, especially for large t (close to 1), is obvious. It can be seen that, our proposed estimator seems to fit better than the other estimators. Using the proposed method, the area under the curve is 0.72, which indicates that although S100 β is not an excellent predictor in this case but it has the potential to distinguish the outcomes between patients with an aneurysmal subarachnoid hemorrhage.

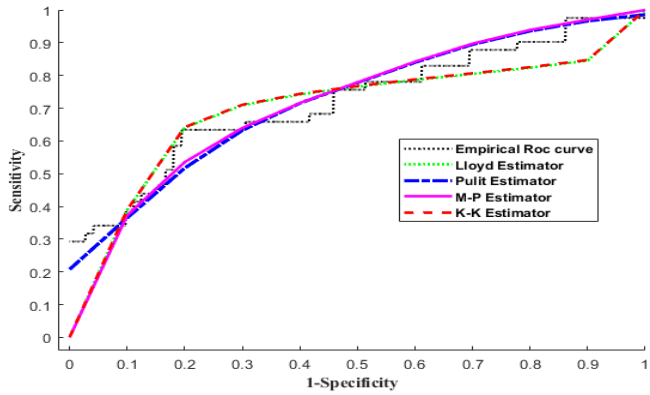

Figure 6: The fitted empirical ROC curve and the estimators of ROC curve for S100β.

6 Conclusion and discussion

The ROC curve is a popular device to assess discriminant rules. The effectiveness of the ROC curve is highly dependent on unbiased distribution estimation. Many researchers employ kernel-based estimators for density and distribution estimation where the smoothing parameter has an essential impact. Pulit (2016) proposed a revolutionary approach for the ROC curve estimation since his estimator is just in need of one smoothing parameter selection. However, his estimator suffers from boundary problem. We proposed an approach to remedy the Pulit estimator. In contrast to other estimators, in our proposed estimator, we estimate the distribution directly and we use a special kernel proposed by Tenreiro (2013) which is designed to be consistent against the boundary problem. Simulation study on different sample sizes and different distributions indicates good performance of our estimator. Also, using a medical dataset, we demonstrated the proposed method's capability to estimate the ROC curve in practice, free from boundary bias.

References

Altman, N. and Leger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, **46**(2):195–214.

Chen, S.X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, **31**(2):131–145.

Chen, S.X. (2000). Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, **52**:471–480.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8): 861-874.

Gasser, T. and Müller, H.G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation: Proceedings of a Workshop held in Heidelberg*, pp. 23–68, Springer Berlin Heidelberg.

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

Hirukawa, M. and Sakudo, M. (2014). Nonnegative bias reduction methods for density estimation using asymmetric kernels. *Computational Statistics & Data Analysis*, **75**:112–123.

Hirukawa, M. and Sakudo, M. (2015). Family of the generalised gamma kernels: A generator of asymmetric kernels for nonnegative data. *Journal of Nonparametric Statistics*, **27**(1):41–63.

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*, **1**:122–127.

John, R. (1984). Boundary modification for kernel regression. *Communications in Statistics-Theory and Methods*, **13**(7):893–900.

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

Lloyd, C.J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, **93**(444):1356–1364.

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

Müller, H.G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika*, **78**(3):521–530.

Pulit, M. (2016). A new method of kernel-smoothing estimation of the ROC curve. *Metrika*, **79**(5):603–634.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C. and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**:1–8.

Silverman, B.W. (2018). *Density Estimation for Statistics and Data Analysis*. New York: Routledge.

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

Tenreiro, C. (2018). A new class of boundary kernels for distribution function estimation. *Communications in Statistics-Theory and Methods*, **47**(21):5319–5332.

Turck, N., Vutskits, L., Sanchez-Pena, P., Robin, X., Hainard, A., Gex-Fabry, M., Fouda, C., Bassem, H., Mueller, M. and Lisacek, F. (2010). A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Medicine*, **36**:107–115.

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

Zou, K.H., Hall, W. and Shapiro, D.E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, **16**(19):2143–2156.