Efficient non-parametric confidence bands for ROC curves

Pablo Martínez-Camblor*

The Dartmouth Institute for Health Policy and Clinical Practice, NH, USA and Universidad Autonoma de Chile, Santiago, Chile

Sonia Pérez-Fernández

Universidad de Oviedo, Asturies, Spain

Norberto Corral

Universidad de Oviedo, Asturies, Spain

Summary

ROC curve is a popular graphical method frequently used in order to study the diagnostic capacity of continuous (bio)markers. In spite of the existence of a huge number of papers devoted to both theoretical and practical aspects of this topic, the construction of confidence bands has had little impact in the specialized literature. As far as the authors know, in the `CRAN` there are only three `R` packages providing ROC curve confidence regions: `plotROC`, `pROC` and `fbroc`. This work tries to fill this gap studying and proposing a new non-parametric method to build confidence bands for both the standard ROC curve and its generalization for non-monotone relationships. The behaviour of the proposed procedure is studied via Monte Carlo simulations and the methodology is applied on two real-world biomedical problems. In addition, an `R` function to compute the proposed and some of the previously existing methodologies is provided as *online supplementary material*.

*Pablo Martínez Camblor, Ph.D., Department of Biomedical Data Science, The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth. Hanover, NH 03755, USA

Email addresses: `Pablo.Martinez.Camblor@Dartmouth.edu` (Pablo Martínez-Camblor), `perezsonia@uniovi.es` (Sonia Pérez-Fernández), `norbert@uniovi.com` (Norberto Corral)

Key words: Confidence bands; ROC curve; Bootstrap method; Sensitivity; Specificity.

1. Introduction

The receiver operating-characteristic (ROC) curve is a popular graphical method frequently used in order to study and compare the diagnostic capacity of continuous (bio)markers. It displays in a plot the false-positive rate, FPR (i.e., the inability of the marker to recognize a normal subject, without the studied characteristic, as normal) against the true-positive rate, TPR (i.e., the ability of the marker to detect the characteristic of interest, in biomedicine, frequently one disease) for all possible thresholds. Conventionally, it is assumed that larger values of the marker indicate larger confidence that a given subject is positive/diseased. Therefore, let χ and ξ be two continuous random variables representing the values of the diagnostic test for negative and positive subjects, respectively. For a fixed point t (FPR), the ROC curve is defined by

$$\begin{aligned}\mathcal{R}(t) =& 1 - F_\xi(F_\chi^{-1}(1-t)) \\ =& \mathcal{P}\{\xi > F_\chi^{-1}(1-t)\} = \mathcal{P}\{1 - F_\chi(\xi) \leq t\} = F_{1-F_\chi(\xi)}(t), \end{aligned} \tag{1}$$

where F_χ and F_ξ denote the cumulative distribution functions (CDF) of variables χ and ξ, respectively. The ROC curve was developed during the World War II in the context of radar signal detection and popularized in the 60s [1]. Since then, it has received great attention in the specialized literature, there is a great amount of references which cover both theoretical and practical aspects of this topic. In a non exhaustive list, we want to highlight the monographs of Pepe [2], Zhou, Obuchowski and McClish [3], Krzanowski and Hand [4] and Broemeling [5, 6]. All of them provide a complete overview of the ROC curves and some related problems from different points of view.

From an inference point of view, the problem of estimating the ROC curve lies on the estimation of the involved CDFs. With this goal, different approaches have been considered (see Gonçalvez, Subtil, Rosário, et al. [7] for a recent

overview of this particular topic). From a non-parametric point of view; Zou, Hall and Shapiro [8], among others, explored the use of kernel density estimators in order to obtain a smooth estimation of the ROC curve. Recently, Cheam and McNicholas [9] proposed a complex algorithm which estimates the unknown CDFs from Gaussian mixture models. In spite of all these procedures, the most used non-parametric estimator is still the direct empirical one, which employs the empirical cumulative distribution function (ECDF) in order to approximate the unknown CDFs. The most common parametric estimator assumes that both positive and negative subjects are normally distributed [10]. Hsieh and Turnbull [11] considered the so-called semiparametric *binormal* model, in which it is assumed that the distributions F_χ and F_ξ are normal after the same unknown monotonic transformation of the measurement scale. The asymptotic properties of the ROC curve estimators [11] can be used in order to obtain pointwise confidence intervals for $\mathcal{R}(\cdot)$ and even for the well-known area under the curve, AUC [12]. On the other hand, for a fixed threshold, the binomial distribution can be employed for deriving confidence intervals for both the sensitivity and the specificity. However, when the focus is the whole ROC curve, one should construct confidence bands, i.e., two random curves $\mathfrak{L}_{\alpha_1}(\omega,\cdot)$ and $\mathfrak{U}_{\alpha_2}(\omega,\cdot)$ (ω denotes the random component) such that, given a fixed confidence level $1-\alpha$ ($\alpha \in [0,1]$):

$$\mathcal{P}\left\{\inf_{t\in[0,1]}\left[\mathcal{R}(t)-\mathfrak{L}_{\alpha_1}(\omega,t)\right]<0\right\}=\alpha_1,$$
$$\mathcal{P}\left\{\sup_{t\in[0,1]}\left[\mathcal{R}(t)-\mathfrak{U}_{\alpha_2}(\omega,t)\right]>0\right\}=\alpha_2,$$

with $\alpha_1+\alpha_2=\alpha$. Note that, with this proviso, the probability that all the points of the ROC curve are within the region between the curves $\mathfrak{L}_{\alpha_1}(\omega,t)$ and $\mathfrak{U}_{\alpha_2}(\omega,t)$ ($t\in[0,1]$) is $1-\alpha$.

This problem has already been considered in the specialized literature. Based on the asymptotic distribution of the ROC curve [11] and using Monte Carlo simulations for approximating the distribution of the limit process, Jensen, Müller and Schäfer [13] developed symmetrical non-parametric confidence bands for

the ROC curve. More recently, Horváth, Horváth and Zhou [14] considered a non-parametric method for confidence bands building based on bootstrapping. This procedure, unlike Jensen, Müller and Schäfer one, does not require estimating density functions. From a parametric point of view, Ma and Hall [15] adapted the Working-Hotelling-type confidence bands [16] used in linear regression for building confidence bands for ROC curves. Demidenko [17] revised this method and proposed the so-called *ellipse-envelope* procedure, which obtains similar but a little bit better coverage percentages than Ma and Hall [15]. From a machine-learning perspective, Macskassy, Provost and Rosset [18] made an empirical revision of different methods for the ROC confidence bands construction and pointed out the difficulty of translating methods for building pointwise confidence intervals, such that those implemented by the three mentioned `R` packages, into methods to obtain confidence bands.

It is interesting to highlight that for a fixed confidence level, $1-\alpha$, given a confidence band, $\{\mathfrak{L}_{\alpha_1}(\omega,\cdot),\mathfrak{U}_{\alpha_2}(\omega,\cdot)\}$, for a fixed specificity, S_P $(=1-p)$, the interval $(\mathfrak{L}_{\alpha_1}(\omega,p),\mathfrak{U}_{\alpha_2}(\omega,p))$ (vertical lines) contains a confidence interval for the sensitivity at level $1-\alpha$. And, for a fixed sensitivity, S_E $(=\mathcal{R}(p))$, the interval (p_1,p_2) satisfying that $\mathfrak{U}_{\alpha_2}(\omega,p_1)=\mathfrak{L}_{\alpha_1}(\omega,p_2)=\mathcal{R}(p)$ (horizontal lines), contains a confidence interval for $1-S_P$ at level $1-\alpha$ (proofs are straightforward). On the other hand, it is easy to check that $(\int\mathfrak{L}_{\alpha_1}(\omega,s)ds,\int\mathfrak{U}_{\alpha_2}(\omega,s)ds)$ contains a confidence interval at level $1-\alpha$ for the true area under the ROC curve. Figure 1 stands for a situation scheme.

In this paper the authors propose to build non-parametric confidence bands for the ROC curve by using the pivotal function $\sqrt{n}\cdot\sigma_n^{-1}(t)\cdot[\hat{\mathcal{R}}(\omega,t)-\mathcal{R}(t)]$ $(t\in[0,1])$, where $\sigma_n^2(t)$ is the variance of $\sqrt{n}\cdot[\hat{\mathcal{R}}(\omega,t)-\mathcal{R}(t)]$ and n the number of included positive subjects, and approximating its distribution by the smoothed bootstrap method. Section 2 provides some technical guidelines for the confidence bands construction for general stochastic curves. This support is then applied to developing confidence bands for both the ROC curve and the ROC curve generalization for non-monotone relationships [19] (onwards gROC). An

`R` function to compute the proposed methodology is provided as *online supplementary material*. This `R` function allows to obtain the most efficient confidence band for the ROC and the gROC curves (in terms of the area between the curves), not only the symmetrical one. The coverage percentages and the efficiency of the proposed confidence bands are studied via Monte Carlo simulations and employed in two real-world datasets.

2. Confidence bands for general curves

Given a target continuous function, $\mathcal{C}(t)$ ($t \in \mathbb{R}$) and a fixed confidence level, $1-\alpha$, we are looking for random curves, $\mathfrak{L}_{\alpha_1}(\omega,t)$ and $\mathfrak{U}_{\alpha_1}(\omega,t)$, satisfying:

$$\mathcal{P}\left\{\sup_{t\in\mathbb{R}}\left[\mathfrak{L}_{\alpha_1}(\omega,t)-\mathcal{C}(t)\right]>0\right\}=\alpha_1, \tag{2}$$

$$\mathcal{P}\left\{\sup_{t\in\mathbb{R}}\left[\mathcal{C}(t)-\mathfrak{U}_{\alpha_2}(\omega,t)\right]>0\right\}=\alpha_2, \tag{3}$$

with $\alpha_1+\alpha_2=\alpha$. Let $\hat{\mathcal{C}}_n(\omega,t)$ be a suitable estimator for $\mathcal{C}(t)$ which, for each $t\in\mathbb{R}$, satisfies the following weak convergence result,

$$n^{\delta}\cdot\{\hat{\mathcal{C}}_n(\omega,t)-\mathcal{C}(t)\}\xrightarrow{\mathcal{L}}_n\ \mathcal{X}(\omega,t), \tag{4}$$

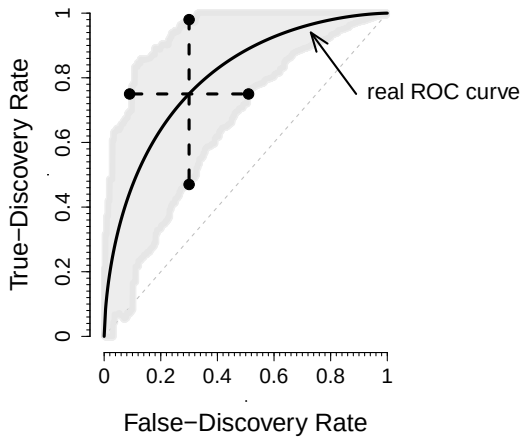

Figure 1: Confidence bands for a particular ROC curve and confidence intervals for a particular sensitivity (vertical lines), 1-specificity (horizontal lines).

where $\delta > 0$ stands for the convergence ratio and $\mathcal{X}$ is a stochastic process satisfying:

(i) $\sup_{t\in\mathbb{R}} |\mathcal{X}(\omega,t)|$ follows some particular probability distribution $\mathcal{D}_{\mathcal{C}}$, and (ii) $\mathbb{E}[\mathcal{X}(\omega,t)] = 0 \quad \forall t \in \mathbb{R}$ ($\mathbb{E}$ denotes the *mean operator*).

Then, for fixed α_1, α_2 ($\alpha_1 + \alpha_2 = \alpha$) and if c_{α_1}, c_{α_2} are such that

$$\mathcal{P}\left\{\sup_{t\in\mathbb{R}} \left\{\sigma(t)^{-1}\cdot\mathcal{X}(\omega,t)\right\} > c_{\alpha_1}\right\} = \alpha_1, \tag{5}$$

$$\mathcal{P}\left\{\inf_{t\in\mathbb{R}} \left\{\sigma(t)^{-1}\cdot\mathcal{X}(\omega,t)\right\} < c_{\alpha_2}\right\} = \alpha_2, \tag{6}$$

where $\mathbb{V}[\mathcal{X}(\omega,t)] = \sigma^2(t)$ ($\mathbb{V}$ denotes the *variance operator*), the region between the random curves $\mathfrak{L}_{\alpha_1}(\omega,t) = \hat{\mathcal{C}}_n(\omega,t) - c_{\alpha_1}\cdot\sigma(t)\cdot n^{-\delta}$ and $\mathfrak{U}_{\alpha_2}(\omega,t) = \hat{\mathcal{C}}_n(\omega,t) - c_{\alpha_2}\cdot\sigma(t)\cdot n^{-\delta}$ is an asymptotic confidence band at level $1-\alpha$ for the target function $\mathcal{C}(\cdot)$. Notice that, for *non-pathological* distributions $\mathcal{D}_C$, from condition (ii), $c_{\alpha_2} < 0$.

3. Confidence bands for the ROC curve

Hsieh and Turnbull [11] proved that, if F_χ and F_ξ have continuous densities, f_χ and f_ξ, respectively, $f_\xi(F_\chi^{-1}(t))/f_\chi(F_\chi^{-1}(t))$ is bounded in any subinterval (a,b) of $(0,1)$ and $n/m \to \lambda$ as $\min\{n,m\} \to \infty$ (n and m stand for the sample size of positive and negative subjects, respectively), then

$$\sqrt{n}\cdot\{\hat{\mathcal{R}}_n(\omega,t) - \mathcal{R}(t)\} \xrightarrow[n]{\mathcal{L}} \mathcal{X}(\omega,t), \tag{7}$$

where $\hat{\mathcal{R}}_n(\omega,\cdot)$ is the empirical ROC curve estimator (ω denotes the random component i.e., the sample) and

$$\mathcal{X}(\omega,t) = \lambda^{1/2}\cdot r(t)\cdot\mathcal{B}_1(1-t) + \mathcal{B}_2(1-\mathcal{R}(t)),$$

where $r(t) = f_\xi(F_\chi^{-1}(1-t))/f_\chi(F_\chi^{-1}(1-t))$ and $\{\mathcal{B}_1(t),\, 0 \le t \le 1\}$ and $\{\mathcal{B}_2(t),\, 0 \le t \le 1\}$ are two independent Brownian bridges. This result guarantees that the region between the random functions $\mathfrak{L}_{\alpha_1}(\omega,t) = \hat{\mathcal{R}}_n(\omega,t) - c_{\alpha_1}\cdot\sigma(t)/\sqrt{n}$ and $\mathfrak{U}_{\alpha_2}(\omega,t) = \hat{\mathcal{R}}_n(\omega,t) - c_{\alpha_2}\cdot\sigma(t)/\sqrt{n}$, with $\sigma^2(t) = \mathbb{V}[\mathcal{X}(\omega,t)]$

and c_{α_1} and c_{α_2} satisfying (5) and (6) above, respectively, is an asymptotic confidence band for $\mathcal{R}(t)$ for any subinterval (a, b) of $(0, 1)$, particularly for $(1/n, 1 - 1/n)$. Then, taken into account that $\mathcal{R}(0) = 0$ and $\mathcal{R}(1) = 1$, the above one is in fact an asymptotic confidence band for $\mathcal{R}(t)$ in $[0, 1]$.

Jensen, Müller and Schäfer [13] approximated the distribution of $\mathcal{X}(\omega, \cdot)$ via Monte Carlo simulations. Unfortunatelly, in order to estimate this distribution, both density functions for positive and negative subjects must also be estimated. These authors proposed to use kernel density estimators with this goal. However, as it is well-known, kernel density estimators are strongly dependent on the selected bandwidth making complex its use in inference (see, for instance, Martínez-Camblor and de Uña-Álvarez [20]). In this context, due to one of the densities appears as denominator, the final obtained estimation could be unstable in those values close to zero [21]. Horváth, Horváth and Zhou [14] based their confidence bands on the pivotal function $\sqrt{n} \cdot \{\hat{\mathcal{R}}(\omega, \cdot) - \mathcal{R}(\cdot)\}$ and approximated its distribution via bootstrapping. In that paper, it is proved the asymptotic convergence of the smoothed bootstrap (SB) approximation. These results can be directly used in order to prove the convergence of the smoothed bootstrap approximation for the pivotal function $\sqrt{n} \cdot \sigma_n^{-1}(\cdot) \cdot \{\hat{\mathcal{R}}(\omega, \cdot) - \mathcal{R}(\cdot)\}$ where $\sigma_n^2(\cdot) = \mathbb{V}[\sqrt{n} \cdot \{\hat{\mathcal{R}}(\omega, \cdot) - \mathcal{R}(\cdot)\}]$. Notice that, when the distribution of the considered estimator directly depends on local properties (the distribution of $\mathcal{X}(\cdot)$ depends on density functions), it is advisable to use smoothed resampling instead of the standard naïve bootstrap one [22]. The sole difference between the SB and the standard naive bootstrap procedures is that, in the first one, the bootstrap samples are run from the smoothed cumulative distribution function estimation, SCFE, instead of from the empirical one. Given a random sample Z_n, the well-known SCFE [23] is defined by $\tilde{F}(Z_n, \cdot) = (1/n) \cdot \sum_{i=1}^{n} K((z_i - \cdot)/h_n)$ where $K(\cdot)$ is a kernel function usual chosen to be a distribution function such that $K(x) = 1 - K(-x)\ \forall x \in \mathbb{R}$, with finite variance and h_n is a real positive number usually called *bandwidth*. Using the standard normal distribution function as kernel, the SB is equivalent to the naive bootstrap and adding white

noise, with variance h_n^2, to the bootstrap samples. Hence, given X_n and Y_m random samples from ξ (positives) and χ (negatives), respectively; we propose to approximate $\sigma_n^2(\cdot)$, c_{α_1} and c_{α_2} using the following algorithm:

A_1 From the observed sample, compute the empirical ROC curve estimation, $\hat{\mathcal{R}}(\omega, \cdot)$ (remember that ω denotes the random component: the sample).

A_2 Compute $\tilde{F}(X_n, \cdot)$ and $\tilde{F}(Y_m, \cdot)$ (SCFE of F_ξ and F_χ, respectively) and for each $b \in \{1, \ldots, B\}$ (B is an arbitrary large number) generate the smoothed random samples $X_n^{*,b}$ and $Y_m^{*,b}$ and their respective empirical ROC curve estimation, $\hat{\mathcal{R}}(\omega^{*,b}, \cdot)$.

A_3 Approximate $\sigma_n^2(\cdot)$ by $\sigma_n^{*,2}(\cdot) = \mathbb{V}[\sqrt{n} \cdot (\hat{\mathcal{R}}(\omega^{*,}, \cdot) - \hat{\mathcal{R}}(\omega, \cdot))]$ and, for $b \in \{1, \ldots, B\}$, compute $U^b = \sup_{t \in [0,1]}\{\sigma_n^{*,-1}(t) \cdot \sqrt{n} \cdot [\hat{\mathcal{R}}(\omega^{*,b}, t) - \hat{\mathcal{R}}(\omega, t)]\}$ and $L^b = \inf_{t \in [0,1]}\{\sigma_n^{*,-1}(t) \cdot \sqrt{n} \cdot [\hat{\mathcal{R}}(\omega^{*,b}, t) - \hat{\mathcal{R}}(\omega, t)]\}$.

A_4 Approximate c_{α_1} by the $(1-\alpha_1)$-percentile of $\{U^1, \ldots, U^B\}$ and c_{α_2} by the α_2-percentile of $\{L^1, \ldots, L^B\}$.

A_5 Finally, compute:

$$\begin{aligned} \mathfrak{L}_{\alpha_1}(\omega, t) =& \hat{\mathcal{R}}(\omega, t) - c_{\alpha_1} \cdot \sigma_n^*(t)/\sqrt{n}, \\ \mathfrak{U}_{\alpha_2}(\omega, t) =& \hat{\mathcal{R}}(\omega, t) - c_{\alpha_2} \cdot \sigma_n^*(t)/\sqrt{n}. \end{aligned}$$

The area between the curves ($a = \int_0^1 [\mathfrak{U}_{\alpha_2}(\omega, t) - \mathfrak{L}_{\alpha_1}(\omega, t)]dt$) could be used to measure the precision of the estimation. From the above algorithm, we obtain that this precision is $a = (c_{\alpha_1} - c_{\alpha_2}) \cdot n^{-1/2} \cdot \int \sigma_n^*(t)dt$. Hence, the most precise estimation is the one which minimizes the quantity $(c_{\alpha_1} - c_{\alpha_2})$, considering that $\alpha_1 + \alpha_2 = \alpha$.

Remark. We realize that, in step A_2, the computation of the SCFEs, $\tilde{F}(X_n, \cdot)$ and $\tilde{F}(Y_m, \cdot)$, requires the selection of adequate bandwidths. In kernel density estimation, the selection of an optimal bandwidth was a really hot topic in the 80s (see, for instance, Wand and Jones [24], and references therein). However, in the SB methodology, bandwidth selection usually has minor impact on the

observed results [25]. In this paper, we consider the bandwidth $h = s \cdot N^{-1/5} \cdot \hat{\sigma}$, where N is the minimum between the positive and the negative sample sizes ($N = \min\{m, n\}$), $\hat{\sigma}$ the sample standard deviation and s is an arbitrary scale parameter. In the following simulations $s = 1$. Observed results for different values of s (1/2, 3/2 and 2 were considered) are provided as online supplementary material. Results suggest that, for s values around 1, the method is stable in most of the considered models. However, expected coverage percentage could be far from the expected one for extreme values of s. For instance, for the considered scenarios (fully explained in the next subsection), the worst observed results was in the Table S2 (online supplementary files), for the model $m_{A,2}$, with a coverage percentage of 70.8% ($n = 50$, $m = 50$).

3.1. Monte Carlo simulation study

The behaviour of the proposed method is studied by Monte Carlo simulations. Three different scenarios have been considered. In the first one (Scenario I), the negative subjects were drawn from a standard normal distribution, $\mathcal{N}_{0,1}$, while four different distributions were considered for the positives: normal distribution with mean 0.95 and standard deviation 1 ($m_{N,1}$), normal distribution with mean 2.13 and standard deviation 3 ($m_{N,2}$), log-normal distribution with both parameters equal to 1/2 ($m_{N,3}$) and the mixture of normal distributions $0.15 \cdot \mathcal{N}_{0,1} + 0.85 \cdot \mathcal{N}_{3,0.75}$ ($m_{N,4}$). In the second scenario (Scenario II), negative subjects were drawn from a centered log-normal distribution with parameters 1/10 and 1 (it was centered in order to be a zero mean distribution) and four different distributions were considered for the positives: normal distribution with mean 2 and standard deviation 1 ($m_{A,1}$), normal distribution with mean 3.75 and standard deviation 3 ($m_{A,2}$), log-normal distribution with both parameters equal to 1/2 ($m_{A,3}$) and the mixture of normal distributions $0.15 \cdot \mathcal{N}_{0,1} + 0.85 \cdot \mathcal{N}_{3,0.75}$ ($m_{A,4}$). Finally, in the third scenario (Scenario III) gamma distributions have been studied. Negative subjects were always drawn from a $\Gamma_{2,3/4}$ ($\Gamma_{a,b}$ stands for a gamma distribution with shape a and scale $1/b$)

while the positives were drawn from the distributions $\Gamma_{5,1}$ ($m_{G,1}$), $\Gamma_{10,2}$ ($m_{G,2}$), $\Gamma_{5/2,1/2}$ ($m_{G,3}$) and $\Gamma_{5/2,1/3}$ ($m_{G,4}$).

Figure 2 depicts the different shapes of both densities and ROC curves considered in the Scenario I. Table 1 shows the observed coverage percentages (% C) and the confidence band areas (mean±standard deviation) for the proposed method (PSN), the one proposed by Jensen, Müller and Schäfer [13] (JMS) and the Demidenko's one [17] (DEK), in 2,000 Monte Carlo simulations for the models described above. All computations were performed using the `R` function `ROCbands` provided as online supplementary material. As it was described above, distribution of PSN pivotal function was approximated by the smoothed bootstrap method with $B = 500$ and $s = 1$. In the models $m_{N,1}$ and $m_{N,3}$ and $m_{N,4}$, the areas between the bands obtained by the proposed method are often larger than those obtained by the other two considered methods, especially compared to DEK. However, the PSN procedure is the only which achieved the expected confidence level in most of the considered situations, even it shows itself mostly conservative. Results obtained by JMS and DEK suggest that, in these considered models and for the whole curve, they do not work adequately. In the model $m_{N,2}$, JMS obtained good results although, in this case, the averages of the area inside the confidence bands were slightly larger than PSN ones. In this model, DEK obtained low but competitive coverage percentages and, as usual, with small areas between the confidence bands.

Table 2 and Figure 3 are similar to Table 1 and Figure 2 but considering the models in Scenario II. Due to negative subjects in the models considered in this scenario are never normal distributed, DEK obtained really poor results. In fact, the provided confidence regions never contained the whole real ROC curve. Notice that, in this context, even the obtained ROC curve estimation is usually far from its target. Although with better results than DEK, JMS also achieved really poor coverage percentages for the whole curve; it is worth remarking that in the original paper, JMS has just proved in subintervals. The proposed method obtained in general good results; although it had some problems in models $m_{A,2}$

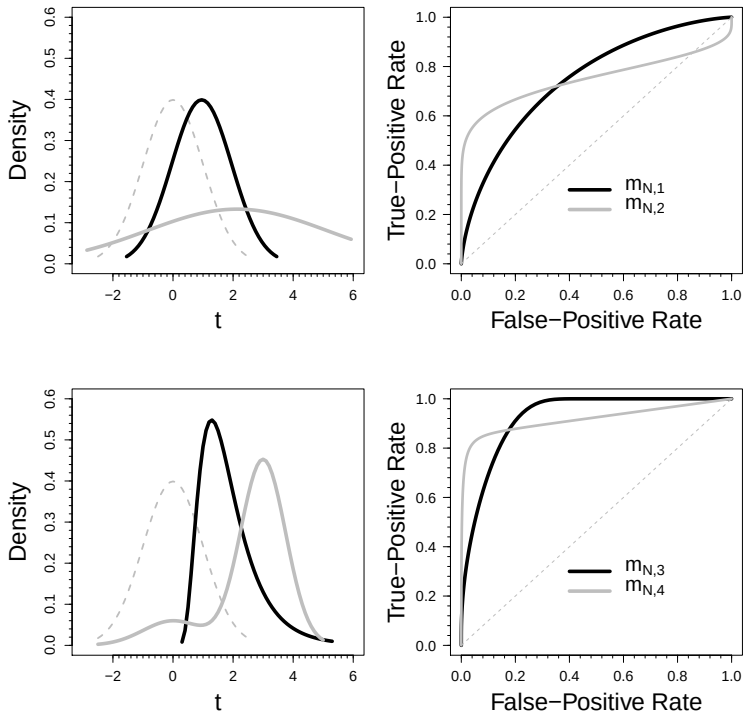

Figure 2: At left, densities for the models considered in the Scenario I: dashed lines denotes the distribution for negatives, in black, models $m_{N,1}$ and $m_{N,3}$, in gray, models $m_{N,2}$ and $m_{N,4}$.

Table 1: Coverage percentages and mean±standard deviation for the area between the 95% confidence bands for the proposed method, PSN; Jensen, Müller and Schäfer, JMS; and Demidenko procedure, DEK, for the Scenario I models.

			PSN		**JMS**		**DEK**	
Model	n	m	**% C**	**Area**	**% C**	**Area**	**% C**	**Area**
$m_{N,1}$	50	50	96.7	0.42 ± 0.04	61.2	0.37 ± 0.05	56.6	0.16 ± 0.01
		100	97.3	0.39 ± 0.04	52.7	0.33 ± 0.04	63.2	0.15 ± 0.01
	100	100	96.8	0.34 ± 0.03	28.9	0.29 ± 0.03	55.0	0.11 ± 0.01
		200	99.6	0.32 ± 0.02	46.7	0.26 ± 0.02	59.8	0.10 ± 0.01
$m_{N,2}$	50	50	97.4	0.36 ± 0.04	95.5	0.47 ± 0.07	89.0	0.21 ± 0.02
		100	97.3	0.36 ± 0.03	93.4	0.43 ± 0.06	94.1	0.22 ± 0.02
	100	100	97.4	0.27 ± 0.02	95.6	0.38 ± 0.05	88.2	0.15 ± 0.01
		200	98.4	0.27 ± 0.02	96.1	0.33 ± 0.03	93.9	0.15 ± 0.01
$m_{N,3}$	50	50	94.1	0.22 ± 0.05	46.1	0.11 ± 0.03	1.1	0.11 ± 0.02
		100	98.5	0.21 ± 0.04	43.9	0.08 ± 0.02	0.9	0.09 ± 0.02
	100	100	98.2	0.18 ± 0.03	36.6	0.08 ± 0.02	1.2	0.08 ± 0.01
		200	99.4	0.17 ± 0.02	35.7	0.08 ± 0.01	0.8	0.06 ± 0.01
$m_{N,4}$	50	50	88.4	0.18 ± 0.05	52.1	0.17 ± 0.05	3.5	0.11 ± 0.02
		100	88.6	0.18 ± 0.05	41.8	0.17 ± 0.05	5.6	0.09 ± 0.02
	100	100	93.8	0.16 ± 0.03	55.4	0.15 ± 0.03	0.2	0.07 ± 0.02
		200	94.4	0.16 ± 0.03	43.4	0.14 ± 0.03	0.5	0.06 ± 0.01

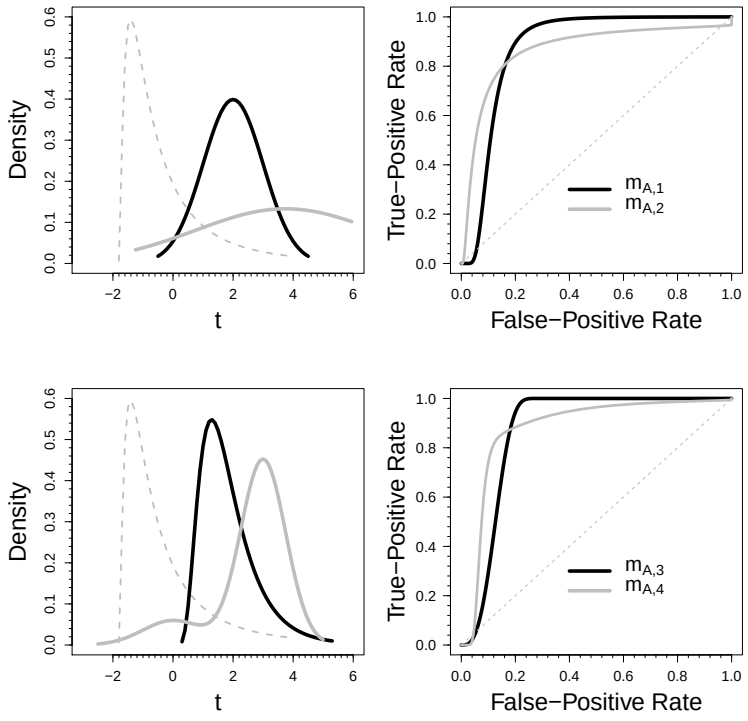

Figure 3: At left, densities for the models considered in the Scenario II: dashed lines denotes the distribution for negatives, in black, models $m_{A,1}$ and $m_{A,3}$, in gray, models $m_{A,2}$ and $m_{A,4}$.

and $m_{A,4}$, the observed results are far from the reference procedures.

Finally, Table 3 and Figure 4 are similar to the previous ones for the models considered in the Scenario III. Observed results support previous conclusions. PSN produced the widest confidence bands and, although a little bit conservative, it was the only method which reached the fixed confidence level in all the considered situations. JMS algorithm often had problems in the extremes of the curves; in models $m_{G,1}$ and $m_{G,2}$, it provided a little bit tighter confidence bands which frequently obtained less confidence than it was expected. In models $m_{G,3}$ and $m_{G,4}$ both JMS and PSN performed similarly. As in the Scenario II models, DEK suffers the lack of normality: this procedure did not work in this

Table 2: Coverage percentages and mean±standard deviation for the area between the 95% confidence bands for the proposed method, PSN; Jensen, Müller and Schäfer, JMS; and Demidenko procedure, DEK, for the Scenario II models.

			PSN		**JMS**		**DEK**	
Model	n	m	**% C**	**Area**	**% C**	**Area**	**% C**	**Area**
$m_{A,1}$	50	50	93.5	0.30 ± 0.06	7.2	0.16 ± 0.04	0.0	0.20 ± 0.04
		100	97.1	0.28 ± 0.04	4.7	0.15 ± 0.03	0.0	0.13 ± 0.02
	100	100	97.9	0.26 ± 0.03	1.6	0.14 ± 0.03	0.0	0.13 ± 0.02
		200	98.7	0.24 ± 0.03	0.7	0.12 ± 0.02	0.0	0.09 ± 0.01
$m_{A,2}$	50	50	80.5	0.30 ± 0.07	53.3	0.23 ± 0.06	0.0	0.16 ± 0.02
		100	91.0	0.30 ± 0.06	58.5	0.22 ± 0.05	0.0	0.14 ± 0.02
	100	100	89.1	0.26 ± 0.04	49.9	0.20 ± 0.04	0.0	0.10 ± 0.01
		200	94.1	0.26 ± 0.03	60.5	0.19 ± 0.03	0.0	0.10 ± 0.01
$m_{A,3}$	50	50	95.5	0.30 ± 0.05	32.9	0.15 ± 0.04	0.0	0.16 ± 0.02
		100	97.8	0.28 ± 0.04	34.9	0.13 ± 0.03	0.0	0.14 ± 0.02
	100	100	99.4	0.24 ± 0.03	24.1	0.13 ± 0.02	0.0	0.10 ± 0.01
		200	94.1	0.25 ± 0.03	17.8	0.11 ± 0.01	0.0	0.10 ± 0.01
$m_{A,4}$	50	50	86.4	0.29 ± 0.06	9.8	0.18 ± 0.05	0.0	0.18 ± 0.04
		100	94.0	0.28 ± 0.05	6.0	0.17 ± 0.04	0.0	0.12 ± 0.02
	100	100	93.9	0.24 ± 0.04	4.5	0.16 ± 0.03	0.0	0.12 ± 0.03
		200	97.1	0.23 ± 0.03	1.8	0.15 ± 0.03	0.0	0.08 ± 0.01

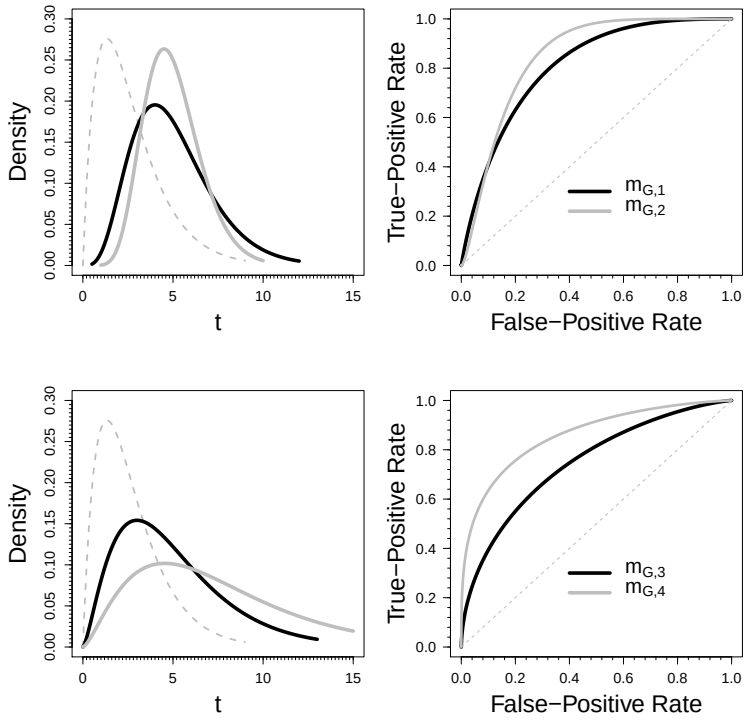

Figure 4: At left, densities for the models considered in the Scenario II: dashed lines denotes the distribution for negatives, in black, models $m_{G,1}$ and $m_{G,3}$, in gray, models $m_{G,2}$ and $m_{G,4}$.

scenario and obtained coverage percentage of zero in all the studied cases.

4. Confidence bands for the gROC curve

Martínez-Camblor, Corral N, Rey C, et al. [19] proposed a ROC curve generalization, $\mathcal{R}_g(\cdot)$, for situations in which both lower and larger values of the marker are associated with higher probability of being positive (called here gROC). The main properties of $\mathcal{R}_g$, defined by

$$\mathcal{R}_g(t) = \sup_{\gamma\in(0,1)} \{1 - \mathcal{R}(1-\gamma\cdot t) + \mathcal{R}([1-\gamma]\cdot t)\} \tag{8}$$

Table 3: Coverage percentages and mean±standard deviation for the area between the 95% confidence bands for the proposed method, PSN; Jensen, Müller and Schäfer, JMS; and Demidenko procedure, DEK, for the Scenario III models.

			PSN		**JMS**		**DEK**	
Model	n	m	**% C**	**Area**	**% C**	**Area**	**% C**	**Area**
$m_{G,1}$	50	50	95.3	0.38 ± 0.05	84.3	0.34 ± 0.04	0.0	0.15 ± 0.01
		100	98.4	0.36 ± 0.04	90.0	0.30 ± 0.03	0.0	0.15 ± 0.01
	100	100	97.8	0.31 ± 0.03	82.1	0.28 ± 0.02	0.0	0.11 ± 0.01
		200	99.5	0.28 ± 0.02	88.1	0.23 ± 0.02	0.0	0.11 ± 0.01
$m_{G,2}$	50	50	96.2	0.35 ± 0.06	58.8	0.25 ± 0.04	0.0	0.16 ± 0.02
		100	98.6	0.32 ± 0.04	70.8	0.21 ± 0.03	0.0	0.11 ± 0.01
	100	100	99.2	0.29 ± 0.03	60.2	0.20 ± 0.02	0.0	0.10 ± 0.01
		200	99.9	0.27 ± 0.02	67.5	0.17 ± 0.02	0.0	0.08 ± 0.01
$m_{G,3}$	50	50	96.0	0.40 ± 0.04	92.8	0.45 ± 0.05	5.1	0.18 ± 0.02
		100	97.7	0.38 ± 0.03	95.5	0.39 ± 0.04	4.4	0.18 ± 0.01
	100	100	96.1	0.32 ± 0.02	95.0	0.36 ± 0.03	0.1	0.13 ± 0.01
		200	99.6	0.29 ± 0.02	95.0	0.30 ± 0.02	0.1	0.12 ± 0.01
$m_{G,4}$	50	50	92.4	0.31 ± 0.05	94.9	0.37 ± 0.06	0.3	0.17 ± 0.06
		100	96.4	0.30 ± 0.04	97.1	0.32 ± 0.04	0.0	0.17 ± 0.01
	100	100	92.7	0.25 ± 0.03	95.9	0.30 ± 0.03	0.0	0.12 ± 0.01
		200	97.4	0.24 ± 0.02	99.0	0.29 ± 0.02	0.0	0.12 ± 0.01

and of its direct empirical estimator, $\hat{\mathcal{R}}_g$, were investigated in that paper. In particular, from the Theorem 2, under the same assumptions that in the standard ROC curve context and with the same notation, for any subinterval $(a, b) \subset (0, 1)$, can be derived the weak convergence that follows:

$$\sqrt{n} \cdot \{\hat{\mathcal{R}}_g(\omega, t) - \mathcal{R}_g(t)\} \xrightarrow{\mathcal{L}}_n \mathcal{X}(\omega, t), \tag{9}$$

with

$$\begin{aligned}\mathcal{X}(\omega, t) =& \lambda^{1/2} \cdot [1 - \gamma_t - \gamma_t' \cdot t] \cdot r([1 - \gamma_t] \cdot t) \cdot \mathcal{B}_1(1 - [1 - \gamma_t] \cdot t) \\ &- \lambda^{1/2} \cdot [-\gamma_t - \gamma_t' \cdot t] \cdot r(1 - \gamma_t \cdot t) \cdot \mathcal{B}_1(\gamma_t \cdot t) \\ &+ \mathcal{B}_2(1 - \mathcal{R}([1 - \gamma_t] \cdot t)) - \mathcal{B}_2(1 - \mathcal{R}(1 - \gamma_t \cdot t)),\end{aligned}$$

where $\gamma_t = \arg\sup_{0 \leq \gamma \leq 1}\{1 - \mathcal{R}(1 - \gamma \cdot t) + \mathcal{R}([1 - \gamma] \cdot t)\}$, γ_t' its derivative and $\{\mathcal{B}_1(t),\, 0 \leq t \leq 1\}$ and $\{\mathcal{B}_2(t),\, 0 \leq t \leq 1\}$ are two independent Brownian bridges.

This result guarantees the good asymptotic behavior of the above algorithm, $A_1 - A_5$, for building confidence bands at level $1 - \alpha$ for the gROC curve. In the next subsection the algorithm performance in the context of finite samples is studied via Monte Carlo simulations.

4.1. Monte Carlo simulation study

The studied situations are similar to the Scenario I and II previously considered in the standard ROC case. In the first scenario (Scenario I), negative subjects were drawn from a standard normal distribution while the positives were generated from: normal distribution with mean 0.95 and standard deviation 1 ($m_{N,1}$), normal distribution with mean zero and standard deviation 2.38 ($m_{N,2}$), log-normal distribution with both parameters equal to 1/2 ($m_{N,3}$) and the mixture of normal distributions $0.4 \cdot \mathcal{N}_{-2,1} + 0.6 \cdot \mathcal{N}_{2,0.75}$ ($m_{N,4}$). In the second considered situation (Scenario II), the negative subjects were drawn from a centered log-normal distribution with parameters 1/10 and 1 and, as in the standard ROC case, four different distributions were considered for the

positives: normal distribution with mean 2 and standard deviation 1 ($m_{A,1}$), normal distribution with mean 0 and standard deviation 3 ($m_{A,2}$), log-normal distribution with both parameters equal to 1/2 ($m_{A,3}$) and the mixture of normal distributions $0.4 \cdot \mathcal{N}_{-4,1} + 0.6 \cdot \mathcal{N}_{0,0.75}$ ($m_{A,4}$). Figure 5 depicts the different shapes of both proposed densities and their respective ROC curves.

Table 4 shows the observed coverage percentage ($\%\, C$) and the confidence band area (mean±standard deviation) for the proposed methodology computed on 1,000 Monte Carlo iterations. Distributions were approximated from 200 smoothed bootstrap replications using $s = 1$. The proposed methodology obtains good, although quite conservative, results for most of the considered models. It should be mentioned that both the lower and upper bound curves were adequately truncated (see the `R` function `ROCbands` enclosed as online supplementary material) for the largest and the smallest values, respectively. It is worth noticing that, when the ROC curve is the most adequate, that is, in the models $m_{N,1}$, $m_{N,3}$, $m_{A,1}$ and $m_{A,3}$, as it is desirable, the observed results for both the gROC and the ROC curves were similar.

5. Real-world practical applications

5.1. Treatment outcome prediction in patients with chronic HCV infection

Vidal-Castiñeira, López-Vázquez, Alonso-Arias et al. [26] studied the effect of several single nucleotide polymorphisms of PD-1 gene and several previously associated factors such as IL28B and KIR receptors on the treatment of chronic hepatitis C virus (HCV) responses. With this goal, the 407 patients finally collected were classified as sustained virological response (SVR, positives) or with non sustained virological response (NR, negatives). More information about the 210 SVR and the 197 NR patients can be found in [26]. In that paper, one of the findings was a predictive model based on a logistic regression including the variables: IL28B rs12979860, PD-1.3, HCV genotype 1-4, KIR-HLA genotype and viral load before treatment. Standard ROC curve was used in order to study the predictive capacity of this model.

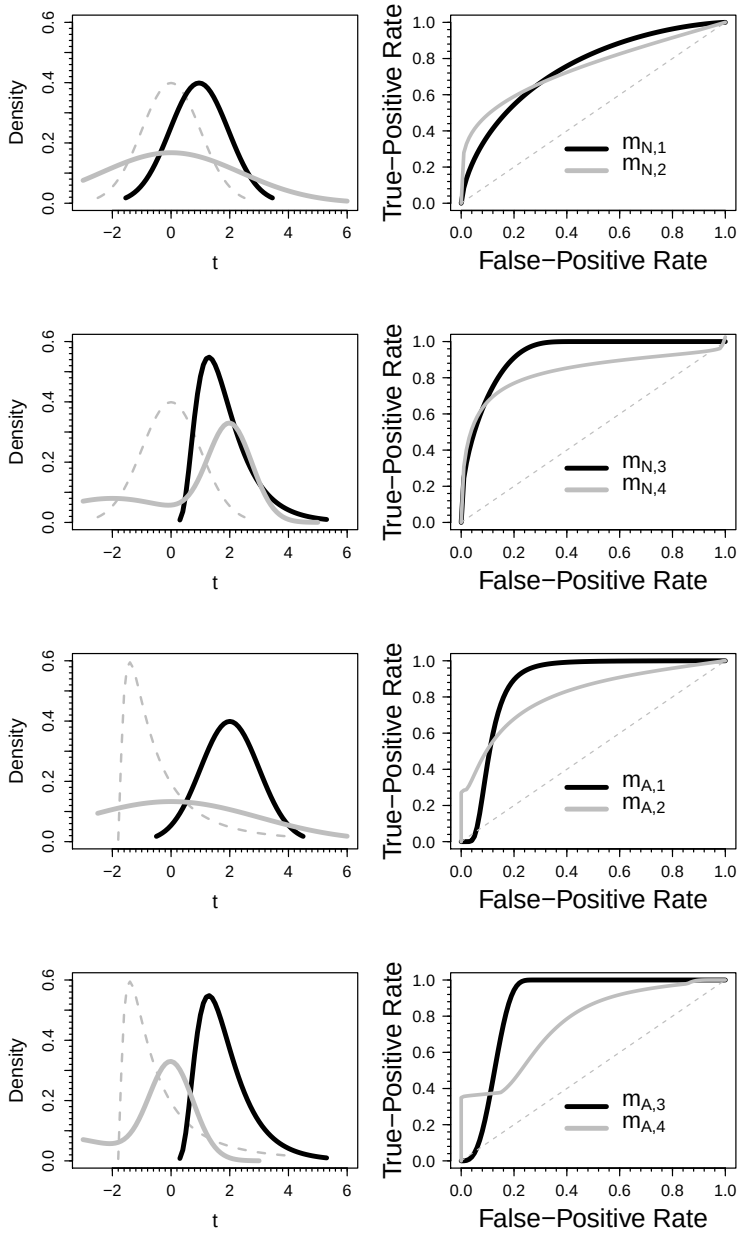

Figure 5: At left, densities for the considered models: dashed lines denotes the distribution for negatives; in black, models $m_{N,1}$, $m_{N,3}$, $m_{A,1}$ and $m_{A,3}$; in gray, models $m_{N,2}$, $m_{N,4}$, $m_{A,2}$ and $m_{A,4}$.

Table 4: Coverage percentages and mean±standard deviation for the area between 95% confidence bands for the gROC curve.

		$m = n$		$m = 2 \cdot n$	
Model	n	**% C**	**Area**	**% C**	**Area**
Scenario I					
$m_{N,1}$	50	96.4	0.41 ± 0.04	98.4	0.38 ± 0.03
	100	96.8	0.34 ± 0.03	93.5	0.31 ± 0.02
$m_{N,2}$	50	92.1	0.35 ± 0.05	94.2	0.34 ± 0.04
	100	92.2	0.29 ± 0.04	92.7	0.27 ± 0.04
$m_{N,3}$	50	99.1	0.25 ± 0.05	99.6	0.22 ± 0.03
	100	99.7	0.20 ± 0.03	99.7	0.18 ± 0.02
$m_{N,4}$	50	99.1	0.28 ± 0.03	98.9	0.28 ± 0.03
	100	99.4	0.25 ± 0.02	99.8	0.24 ± 0.02
Scenario II					
$m_{A,1}$	50	93.9	0.29 ± 0.06	96.1	0.26 ± 0.04
	100	98.6	0.24 ± 0.04	98.9	0.20 ± 0.04
$m_{A,2}$	50	99.4	0.37 ± 0.06	99.6	0.37 ± 0.06
	100	99.8	0.34 ± 0.05	98.3	0.33 ± 0.05
$m_{A,3}$	50	96.6	0.30 ± 0.05	98.2	0.25 ± 0.03
	100	99.2	0.22 ± 0.03	98.4	0.21 ± 0.02
$m_{A,4}$	50	99.5	0.41 ± 0.06	99.9	0.41 ± 0.05
	100	99.1	0.36 ± 0.05	99.9	0.36 ± 0.04

Figure 6-A shows kernel density estimations for the model punctuations in both positive and negative patients. In spite of the fact that, apparently, the density shapes are non-normal, the standard ROC curve seems to be the most appropriate one. Both ROC and gROC curves have the same shape although a small difference between them is observed for false-positive rates around 0.6 (Figure 6-B). The variability approximation, $\sigma^*(t)$, was also similar for both the ROC (Figure 6-C) and the gROC (Figure 6-E) curves and both were robust respect to the s-value. While, at 95% level, the optimal confidence band for the ROC curve is reached for $\alpha_1 = 0.015$ ($\alpha_2 = 0.035$) and the area between the upper and lower band is 0.282, for the gROC curve, the symmetrical confidence band is the optimal one ($\alpha_1 = \alpha_2 = 0.025$) with an area of 0.267. It is worth mentioning that area between the confidence bands considering different α_1 and α_2 (always $\alpha_1 + \alpha_2 = 0.05$) shifted less in the ROC (with a maximum area of 0.305) than in the gROC curve (with a maximum of 0.454).

Figure 7 shows the resulting confidence bands for the JMS, DEK and square pointwise algorithms. The last method is the one provided for the three `R` packages which currently perform ROC curve confidence bands (`pROC`, `fbroc` and `plotROC`). All of them compute pointwise square confidence bands based on bootstrapping. The three methods drew tighter confidence bands than PSN. Areas between the lower and the upper lines were 0.183, 0.068 and 0.143, for JMS, DEK and square pointwise, respectively. It should be noted that the difference between the JMS and the PSN methods which, in spite of being based on the same pivotal function, used different approximations to its distribution. Main difference between these procedures is located at left, for the smallest FPRs.

5.2. Postoperative infection dataset

The objective of this study was to evaluate whether the glucose levels are useful in order to predict the appearance of infection in the inmediate post-operative period. With this goal, López-Ratón [27] considered a subgroup of

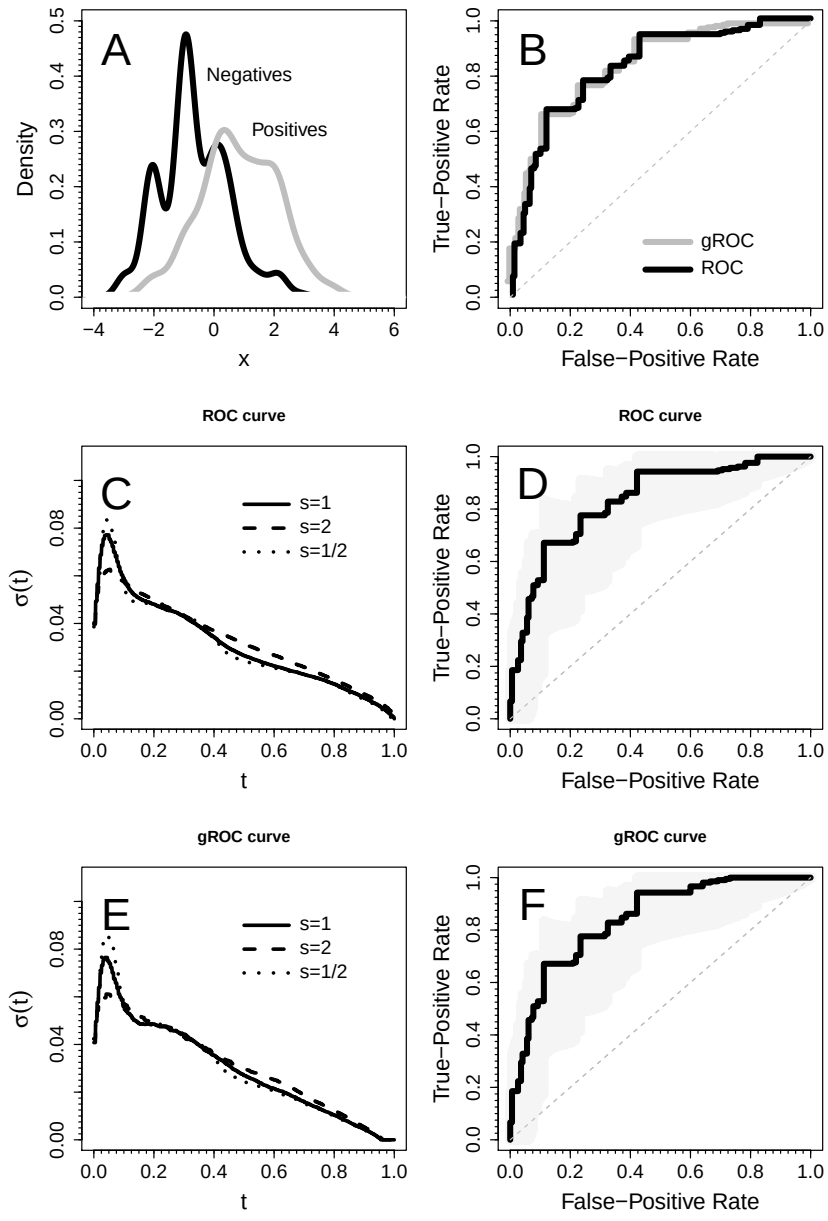

Figure 6: ROC curve analysis for the HCV infection data: density estimations for the model punctuation in positive and negative subjects, ROC and gROC curves, estimated variability for different bandwidths, and 95% confidence bands.

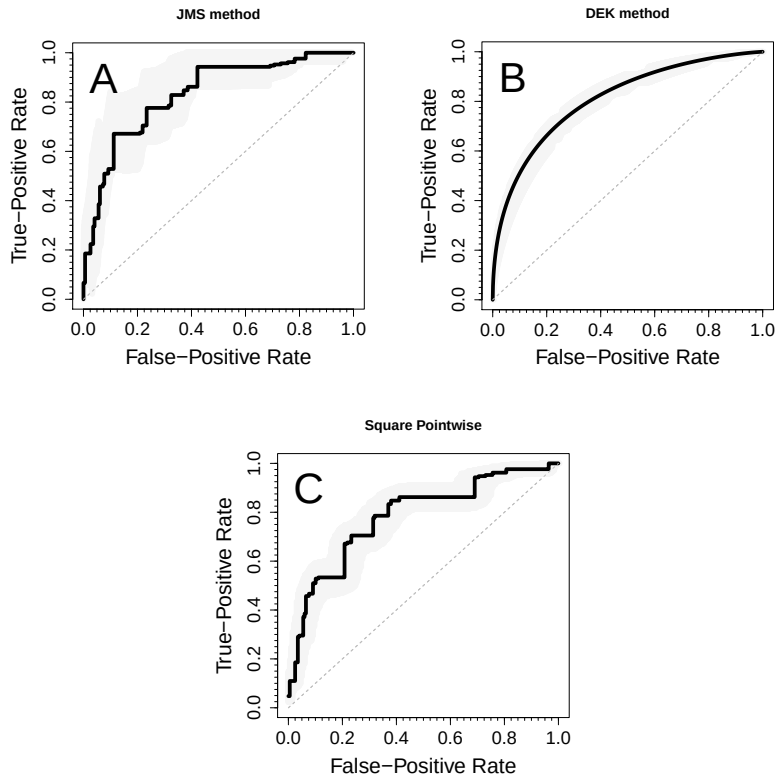

Figure 7: 95% Confidence bands for the JMS and DEK methods computed with the R function rocBands provided as online supplementary material and square pointwise confidence bands computed with the R package pROC for the HCV infection dataset.

non-diabetic individuals who underwent clean surgical interventions at the Hospital Clínico Universitario de Santiago (Santiago de Compostela, north of Spain) from January 1996 to March 1997. A total of 836 patients were finally included, 45 (5%) of them suffered a postoperative infection (POI). Figure 8-A depicts the kernel density estimations for both positive (POI) and negative (Non POI) patients.

Due to the main difference between the POI and Non POI distributions is the variability, in this case the gROC curve seems to be the most appropriate. Figure 8-B shows both ROC and gROC curves. For the smallest specificities the gROC curve yields larger sensitivities than the right-side ROC curve. Variability functions are again robust respect to s-values. However, in this case, the observed approximations for the variability, $\sigma^*(t)$, for the ROC and the gROC curves are quite different each other and also really different from those ones observed in the previous considered example (see Figures 8-C and 8-E). The small number of positive subjects and the large variability of the glucose levels provoke large confidence bands: the optimal one for the right-side ROC curve, achieved for $\alpha_1 = 0.04$ ($\alpha_2 = 0.01$), has an area between the curves of 0.484 (the largest one had an area of 0.751) and the diagonal line is contained within this region (Figure 8-D). The gROC curve does not work quite different: the area of the optimal band, reached for $\alpha_1 = 0.015$ ($\alpha_2 = 0.035$), is 0.380 (maximum of 0.446) and, in spite of the fact that the gROC curve is better than the standard ROC, the diagonal line is mostly still within the confidence region (Figure 8-F).

Figure 9 shows the resulting confidence bands for the JMS, DEK and square pointwise methods. In this case, DEK and square pointwise procedures also led on to tighter confidence bands than the PSN one. Their areas between the lower and the upper lines were 0.251 and 0.255, respectively. JMS and PSN produced similar results with areas of 0.425 and 0.484, respectively. Again, main difference between both confidence bands is located at the left side of the curve. It is worth mentioning that the four confidence bands included, total or partially, the main diagonal within the confidence region.

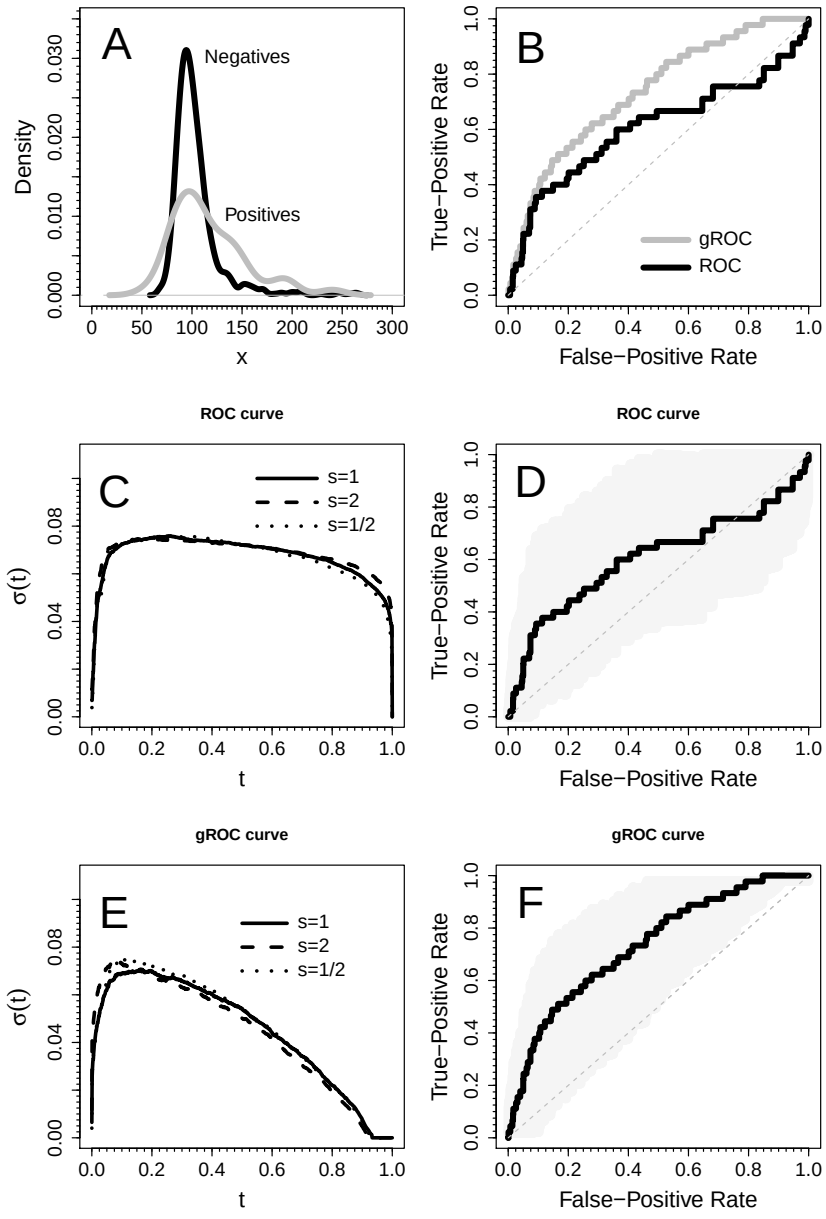

Figure 8: ROC curve analysis for the Postoperative infection data: density estimations for the glucose levels in positive and negative subjects, ROC and gROC curves, estimated variability for different bandwidths, and 95% confidence bands.

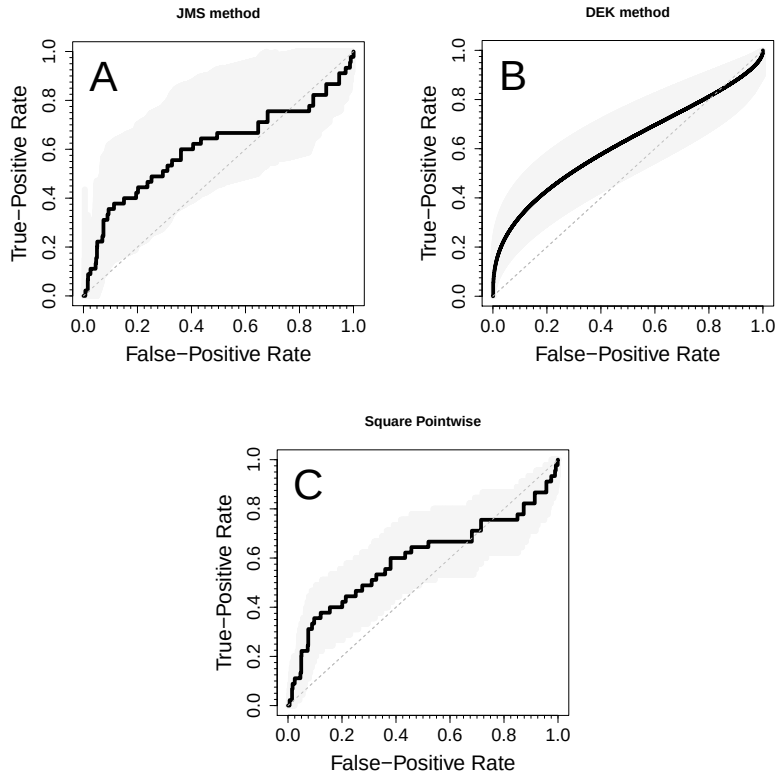

Figure 9: 95% Confidence bands for the JMS and DEK methods computed with the R function rocBands provided as online supplementary material and square pointwise confidence bands computed with the R package pROC for the Postoperative infection dataset.

6. Discussion

The ROC curve is a popular graphical tool often used in order to study the diagnostic capacity of continuous (bio)markers. Both its theoretical and practical aspects have been deeply studied in the specialized literature. In spite of the huge amount of published works, confidence bands construction has deserved less attention. Particularly, the authors have selected four main papers: two of them (Jensen, Müller and Schäfer [13] and Horváth, Horváth, Zhou [14]) dealing with the problmen from the non-parametric point of view, and two more (Ma and Hall [15] and Demidenko [17]) from a parametric approach. Regarding to the non-parametric methods: in the approach proposed by Jensen, Müller and Schäfer [13] the estimation of a quotient dependent on a smoothed parameter is required in order to approximate the pivotal function distribution, and this directly brings some well-known associated problems with it [21]. The Horváth, Horváth and Zhou [14] proposal generates parallel confidence bands which are not usually the most efficient ones, specially in the extremes of the curve, where, usually, the variability is smaller. The parametric approach has the usual limitation of fixed distribution models assumption.

The confidence bands are a natural generalization to the confidence interval concept for the *curves context*. The idea is to build a region in which we can likely draw the real curve, that is, all points of the real curve are contained within this region. Due to the ROC curve provides a graphical overview of the diagnostic capacity of a marker, an appropriate confidence band lets us know the accuracy of this overview, allowing to see the weaknesses and strengths parts (those ones wider and narrower, respectively) of the provided analysis.

In this paper, the authors use the pivotal function proposed in [13] but they take advantage of some theoretical results developed by Horváth, Horváth and Zhou [14]. The result is an efficient non-parametric procedure for building asymptotic confidence bands for both ROC and gROC curves. Note that similar procedures to the proposed one can be used in order to develop confidence bands for more general curves based on estimators satisfying the convergence given

in (4). In particular, for other ROC curve generalizations such as the time-dependent case (see Martínez-Camblor, Bayón and Pérez-Fernández [28] for a recent output of this topic).

While the asymptotic behaviour of the proposed method is directly derived from [14], Monte Carlo simulations have been used to check its finite sample performance. Observed results suggest that the proposed methodology provides conservative confidence bands for both the ROC and the gROC curves in most studied models. Even though the area between the bands was usually larger than the one observed in the reference methods [17] [13], we must highlight that it was the only one which achieved the fixed coverage percentage in most of models considered. In order to get appropriate coverages, the extremes of the curves are the main problem, since for those values with high/low sensitivity/specificity the computed confidence regions work worse. This was the main problem in the JMS procedure, so that one must be careful making inferences at these points, especially, when the slope of the curve is high. The procedures used as references (Demidenko [17] and Jensen, Müller, Schäfer [13]) obtained really poor results in most of the considered models. The observed results suggest that parametric method [17] is seriously affected by the lack of normality.

Although the used pivotal function does not include any smoothed parameter, due to the fact that its function distribution directly depends on local properties, it is advised to use smoothed bootstrap in order to approximate it (see Hall, DiCiccio, Romano [22]). Hence, a bandwidth parameter must be selected with this goal. In this paper, we chose bandwidth in the way $h = s \cdot \min\{n, m\}^{-1/5} \cdot \hat{\sigma}$. Even though the method obtained similar and good results for s-values around 1, it could not work for too large or too small s-values. While larger s-values are always associated with larger confidence regions, the relationship is not always the same for the coverage percentages. In the Scenario I, where the negative subjects were symmetrically distributed, the observed coverage percentages were larger, and similar, for $s = 3/2$ and $s = 2$ than for $s = 1/2$ (Table S1 in the online supplementary material). In

the Scenario II, with the negative subjects asymmetrically distributed, the observed impact of the parameter s is greater; the coverage percentages decreased with bandwidth. Here, we observed the worst result: a coverage percentage of 70.8% in the model $m_{A,2}$ ($n = 50, m = 50$) and $s = 2$ (Table S2 in the online supplementary material). Finally, in the Scenario III (Table S3), where both positives and negatives follow gamma distributions, s-value has a minor impact on the observed coverage percentages: all considered s-values performed similarly althought the biggest, $s = 2$, showed itself a little more conservative than the others. It is worth to remark that, although the obtained confidence bands perform well for the main body of the curve (central part), the procedure requires some calibration for the largest and smallest ROC curve values (right-upper and left-lower extremes: highest and lowest sensitivities) in order to get suitable coverage percentages: lower and upper bounds must be truncated.

Both Monte Carlo simulations and the studied real-world datasets results suggest the well performace of the gROC curve. It does not overestimate diagnostic capacity when the standard ROC curve is the proper one. Moreover the inferences based on this curve seem to have the expected behaviour. Note that in the second real-word example, conclusions do not differ quite much respect to those ones derived from the ROC curve: sample size of positives does not allow to make any definitive decision, which seems to be the most expected conclusion.

Finally, it is clear that the popularization of statistical methodologies strongly depends on the existence of friendly and available computational tools. Among more than twenty `R` packages in the `CRAN` including some procedures related to ROC curve, just three of them provide confidence regions: the `fbroc` package developed by Erik Peter and published on 12/10/2015 contains a function which computes confidence regions for the right-side ROC curve but no information about the method used to build these regions is provided. The package `plotROC` developed by Michael C. Sachs and published on 05/02/2016 also contains a function which, literally *displays rectangular confidence regions for the ROC*

curve. And finally perhaps the most known and commonly used in practice, the `pROC` package: developed by Robin, Turk, Haimard et al. [29] and published on 05/05/2015. This package, labelled as confidence intervals for *shape*, computes square pointwise confidence bands. In fact, these three package produce the same square pointwise confidence bands. As we mentioned in the introduction of this paper, an empirical revision of these methods performance has already been carried out by Macskassy, Provost and Rosset [18] pointed out the difficulty of translating methods for building pointwise confidence intervals into methods to obtain confidence bands. We provided, as online supplementary material, the results obtained by this method in our three scenarios. Obtained results (Table S5) support the conclusions presented by Macskassy, Provost and Rosset [18]. Produced confidence bands were too tight. The maximum observed coverage percentage was 78.9% in the model $m_{N,2}$ ($N = 100$, $M = 100$). Coverage percentages were close to zero for models $m_{N,3}$, $m_{A,1}$, $m_{G,1}$ and $m_{G,2}$. In this work, also as online supplementary material, it is included the `R` function `ROCbands` which allows computing and plotting the proposed confidence bands. This function also computes the confidence bands proposed by Demidenko [17] and by Jensen, Müller and Schäfer [13]. However, we realize that these procedures were computed based on our reading of those papers. We have tried to contact the authors in order to ensure the accurateness of their procedure implementations but, unfortunately, we have not got it.

Acknowledgements

The authors are grateful to José Ramón Vidal Castañeira and to Mónica López-Ratón for their permission to use the HCV infection and the Postoperative infection data, respectively. To Xabier Robin for his clarifications about the `pROC` package and to the anonymous reviewers for their comments and suggestions. This paper was financial supported by the Grants MTM2014-55966-P and MTM2015-63971-P from the Spanish Ministerio of Economía y Competitividad and by FC-15-GRUPIN14-101 from the Principado de Asturias.

References

[1] Green DM, Swets JA, *Signal detection theory and psychophysics*, New York, USA, Wiley, 1966.

[2] Pepe MS, it *The statistical evaluation of medical tests for classification and prediction*, Oxford, Oxford University Press, 2003.

[3] Zhou XH, Obuchowski NA, McClish DK, *Statistical methods in diagnostic medicine*, New York, USA, Wiley, 2002.

[4] Krzanowski WJ, Hand DJ, *ROC curves for continuous data*, Boca Raton, FL, Chapman & Hall/CRC Press, 2009.

[5] Broemeling LD, *Bayesian biostatistics and diagnostic medicine*, Boca Raton, FL, Chapman & Hall/CRC Press, 2007.

[6] Broemeling LD, *Advanced bayesian methods for medical test accuracy*, Boca Raton, FL, Chapman & Hall/CRC Press, 2011.

[7] Gonçalvez L, Subtil A, Rosário Oliveira M, et al., ROC curve estimation: an overview, *REVSTAT-Statistical Journal*, 2014, **12**(1), 1-20.

[8] Zou KH, Hall WJ, Shapiro DE, Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests, *Statistics in Medicine*, 1997, **16**(19), 2143-2156.

[9] Cheam ASM, McNicholas PD, Modelling receiver operating characteristic curves using Gaussian mixtures, *Computational Statistics & Data Analysis*, 2016, **93**, 192-208.

[10] Hanley JA, The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests, *Statistics in Medicine*, 1996, **15**, 1575-1585.

[11] Hsieh F, Turnbull BW, Nonparametric and semiparametric estimation of the receiver operating characteristic curve, *Annals of Statistics*, 1996, **24**, 25-40.

[12] Hanley JA, NcNeil BJ, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, 1982, **148**, 29-36.

[13] Jensen K, Müller H-H, Schäfer H, Regional confidence bands for ROC curves, *Statistics in Medicine*, 2000, **19**, 493-509.

[14] Horváth L, Horváth Z, Zhou W, Confidence bands for ROC curves, *Journal of Statistical Planning and Inference*, 2008, **138**, 1894-1904.

[15] Ma G, Hall WJ, Confidence bands for receiver operating characteristic curves, *Medical Decision Making*, 1993, **13**, 191-198.

[16] Working H, Hotelling H, Application of the theory of error to the interpretation of trends, *Journal of the American Statistical Society*, 1929, **24**, 73-85.

[17] Demidenko E, Confidence intervals and bands for the binormal ROC curve revisited, *Journal of Applied Statistics*, 2012, **39**(1), 67-79.

[18] Macskassy S, Provost F, Rosset S, ROC confidence bands: an empirical evaluation, *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[19] Martínez-Camblor P, Corral N, Rey C, et al., ROC curve generalization for non-monotone relationships, *Statistical Methods in Medical Research*, 2014, `doi:10.1177/0962280214541095`.

[20] Martínez-Camblor P, de Uña-Álvarez J, Studying the bandwidth in k-sample smooth tests, *Computational Statistics*, 2013, **28**, 875-892.

[21] Martínez-Camblor P, Nonparametric cutoff point estimation for diagnostic decisions with weighted errors, *Revista Colombiana de Estadística*, 2011, **34**(1), 133-146.

[22] Hall P, DiCiccio JT, Romano JP, On smoothing and the bootstrap, *Annals of Statistics*, 1989, **17**, 692-704.

[23] Nadaraya EA, (1964) Some new estimates for distribution functions, *Theory of Probability and its Applications*, **9**, 497-500.

[24] Wand MP, Jones MC, *Kernel Smoothing*, New York, USA. Chapman & Hall, New York: 1996.

[25] Martínez-Camblor P, Carleos C, Corral N, General nonparametric ROC curves comparison, *Journal of the Korean Statistical Society*, 2013, **42**, 71-81.

[26] Vidal Castañeira JR, López-Vázquez A, Alonso-Arias R, et al., A predictive model of treatment outcome in patients with chronic HCV infection using IL28B and PD-1 genotyping, *Journal of Hepatology*, 2012, **56**(6), 1230-1238.

[27] López-Ratón M, *Optimal cutoff points for classification in diagnostic studies: new contributions and software developments*, PhD dissertation, Universidade de Santiago de Compostela, Spain: 2015.

[28] Martínez-Camblor P, F. Bayón G, Pérez-Fernández S, Cumulative/dynamic ROC curve estimation, *Journal of Statistical Computation and Simulation*, 2016, `doi: 10.1080/00949655.2016.1175442`.

[29] Robin X, Turck N, Hainard A, et al., pROC: an open source package for R and S+ to analyze and compare ROC curves, *BMC Bioinformatics*, **12**(77), 1-8.