IMS Collections
Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková
Vol. 7 (2010) 12–24
© Institute of Mathematical Statistics, 2010
DOI: 10.1214/10-IMSCOLL702

Nonparametric comparison of ROC curves: Testing equivalence*

Jaromír Antoch[1], **Luboš Prchal**[1,2] and **Pascal Sarda**[2]

Charles University of Prague and Université Paul Sabatier Toulouse

Abstract: The problem of testing equivalence of two ROC curves is addressed. A transformation of corresponding ROC curves, which motivates a test statistic based on a distance of two empirical quantile processes, is suggested, its asymptotic distribution found and a simulation scheme proposed that enables us to find critical values.

1. Introduction

Receiver operating characteristic (ROC) curves are a popular and widely used tools that can help to summarize the overall performance of diagnostic methods and/or classifiers assigning individuals $g \in \mathcal{G} = \mathcal{G}_0 \cup \mathcal{G}_1, \mathcal{G}_0 \cap \mathcal{G}_1 = \emptyset$, into one of the groups $\mathcal{G}_0$ or $\mathcal{G}_1$. Typically, the $\mathcal{G}_1$ individuals hold a feature of interest and are referred to as *positives*, while the $\mathcal{G}_0$ individuals are without the feature and are referred to as *negatives*.

Assume that a suitable diagnostic measure Y is available. By convention, the larger values of Y are supposed to be more indicative for an individual to belong to $\mathcal{G}_1$, so that if $Y \geq t$, $t \in R$ is a fixed threshold, then an individual is assigned to $\mathcal{G}_1$. On the contrary, if $Y < t$ then it is assigned to $\mathcal{G}_0$. Let us introduce probabilities $F_0(t) = \mathrm{P}(Y \leq t \,|\, \mathcal{G}_0)$ and $F_1(t) = \mathrm{P}(Y \leq t \,|\, \mathcal{G}_1)$. It is evident that $F_0(t)$ and $F_1(t)$ as functions of t are distribution functions of the diagnostic variable Y for the $\mathcal{G}_0$ and $\mathcal{G}_1$ groups, so that we can denote the corresponding random variables by Y_0 and Y_1. With this notation in mind, one possible way is to define ROC functions as mapping $\varrho(\cdot; F_0, F_1)$, where

$$\begin{aligned} \varrho(\,\cdot\,; F_0, F_1): \ \mathbb{R} &\to \ [0,1] \times [0,1] \\ t &\mapsto \ \big[1 - F_0(t), 1 - F_1(t)\big]. \end{aligned} \tag{1}$$

In other words, it is a curve in a unit square $[0,1] \times [0,1]$ square consisting of $1 - F_1(t)$ on the vertical axis plotted against $1 - F_0(t)$ on the horizontal axis for all $t \in \mathbb{R}$. We refer the readers to the monographs Zhou et al. [32] and Pepe [23] for the properties and applications of ROC curves.

In practice, ROC curves are often used to compare several diagnostic methods (classifiers). It is usually accepted that the method with a corresponding ROC curve

*Work was supported by grant GAČR 201/09/0755 and research project MSM 0021620839.

[1] Charles University in Prague, Department of Probability and Mathematical Statistics, Sokolovská 83, CZ – 186 75 Praha 8, Czech Republic. e-mails: jaromir.antoch@mff.cuni.cz, lubos.prchal@mff.cuni.cz

[2] Université Paul Sabatier, Institut de Mathématiques de Toulouse, UMR 5219, 118 route de Narbonne, F – 310 62 Toulouse cedex, France. e-mail: sarda@cict.fr

AMS 2000 subject classifications: Primary 62G05; secondary 62H30, 62P99.

Keywords and phrases: ROC curves, binary classification, kernel ROC curves.

closest to the point $(0, 1)$ is the best one for the particular problem. However, this oversimplified rule is not easily applicable in practice because ROC curves in many applications are mostly non convex and the effect on the analysis can be non-trivial. Some examples are presented in this paper. Figure 1 displays three plots, each with a pair of ROC curves corresponding to different association measures suitable for the collocation extraction. It illustrates three typical situations that we come across.

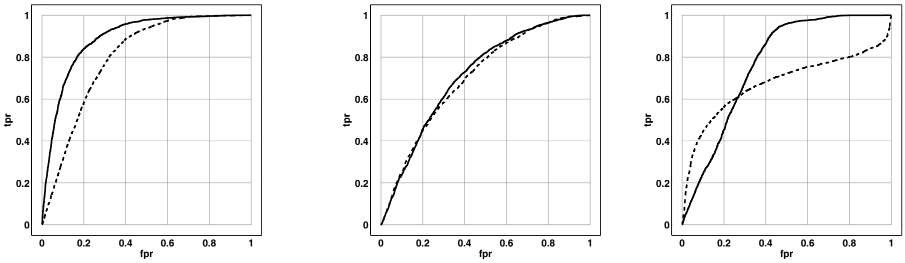

Fig. 1. *Examples of ROC curves for several linguistic measures described in Pecina and Schlesinger [22].*

First, everyone would agree that the solid curve in Figure 1a outperforms the dashed one. Figure 1b seems to be the opposite case, because both association measures provide, at least optically, equivalent ROC curves. Finally, the situation in Figure 1c is not at all clear. On one hand the solid line is much closer to the point $(0, 1)$. On the other hand, the curves are crossing and it is not at all clear which of them we should prefer. In all three cases, nevertheless, a natural question arises: *Are these ROC curves significantly different?*

Several methods exist for testing the equivalence of two ROC curves. The pioneer work, proposed for normally distributed variables, was Greenhouse and Mantel [11], later extended by Weiand et al. [30] and Beam and Wieand [2]. The most widely used current approach is based on the AUC (area under the curve), proposed by Bamber [1] and developed further by, e. g., Hanley and McNeil [13] and Delong et al. [6]. A totally different approach to testing is based on a permutation principle suggested by Venkatraman and Begg [29]. Additional parametric methods, mainly connected to the binormal ROC curves and their transformations, have been also developed. We refer to Zhou et al. [32] for the review of the parametric ROC curve modeling.

In practice it is usual that we do not have any a priori information about the form of the underlying distribution of Y. In such a case a parametric approach is not appropriate. Since we often deal with curves possibly crossing each other as in Figure 1c, the AUC test does not work, since the crossing curves may have the same AUC but represent diagnostic methods with completely different properties. However, in case of large sample sizes, large numbers of considered ROC curves disqualify the use of the permutation principle or other resampling techniques because they are unsupportable from a computational point of view.

All of these considerations motivated us to suggest a new test of equivalence of two ROC curves. The basic idea is to transform the testing problem and consider the methods separately in groups $\mathcal{G}_0$ and $\mathcal{G}_1$ rather than to compare the ROC curves themselves. We believe that this alternative approach covers a large field of ROC settings and might open new perspectives of a ROC curve analysis as a whole. It leads to a test statistic based on the difference between the quantile processes associated with diagnostic variables of each group, and enables us to determine the asymptotic distribution under the null hypothesis of ROC curves equivalence. These points are discussed in Section 2, where a more precise setting for ROC curves and their estimators are presented as well.

Regarding estimation of $F_0(t)$ and $F_1(t)$, we use the empirical cumulative distribution function (CDF). The main competitor of the empirical CDF is the smooth kernel CDF estimator that possesses some theoretical and "visual" advantages for CDF and ROC curve estimation. For details see, e. g., Falk [9] or Zhou et al. [33]. However, in the case of large sample size of data the possible advantage of the kernel ROC curve appears to be completely negligible. On the other hand, estimating ROC curves and testing their equivalence are totally different tasks. In our experience, the kernel estimator does not substantially improve the testing procedure, whereas the empirical CDF estimator is easier to apply. Nevertheless, in other practical situations the kernel approach can be useful, at least as an alternative to the empirical CDF. It is shown that all theoretical results remain true when testing is based on either the empirical or the kernel estimators.

The rest of the paper is organized as follows. Section 2 contains the hypothesis formulation, description of the test procedure, discussion about finding critical values and the use of the kernel estimators instead of the empirical ones. The proofs of the theoretical results formulated in Section 2 are given in Appendix.

2. Test of equivalence of two ROC curves

2.1. *Hypothesis formulation*

Let Y be a diagnostic variable with distribution functions $F_0(t)$ and $F_1(t)$, and let Y_0 and Y_1 denote corresponding random variables as introduced in Section 1 above formula (1). Denote, according to (1), the ROC curve associated to Y by

$$\mathrm{ROC}_Y = \big\{ \boldsymbol{r} \in [0,1]^2 : \exists\, t \in \mathbb{R} \quad \varrho(t; F_0, F_1) = \boldsymbol{r} \big\}. \tag{2}$$

Moreover, assume that:

(C1) Y_0 and Y_1 have continuous distributions with densities $f_0(t)$ and $f_1(t)$ such that $f_0(t) > 0$ and $f_1(t) > 0$ on the same interval $\mathcal{I}_Y \subseteq \mathbb{R}$, and that the densities are equal to zero outside $\mathcal{I}_Y$.

(C2) Y_0 and Y_1 are independent.

Remarks.

(i) Model assumption (C1) on supports assures one-to-one mapping between the thresholds and ROC points in the unit square $[0,1] \times [0,1]$ square. This technically simplifies the notation used later, but it can be relaxed if one properly takes into account the relationship between t and the ROC curve

(ii) Assumption (C2) means that diagnostic variable Y_0 keeps only the information assuring that negatives belong to $\mathcal{G}_0$, while diagnostic variable Y_1 keeps only the information assuring that positives belong to $\mathcal{G}_1$.

Let us introduce another diagnostic variable Z with the distribution functions $G_0(t)$ and $G_1(t)$, denoting the corresponding ROC curve by

$$\mathrm{ROC}_Z = \big\{ \boldsymbol{r} \in [0,1]^2 : \exists\, t \in \mathbb{R} \quad \varrho(t; G_0, G_1) = \boldsymbol{r} \big\}, \tag{3}$$

and assume that Z_0 and Z_1 also satisfy conditions (C1) and (C2) with densities $g_0(t)$ and $g_1(t)$ on some $\mathcal{I}_Z$. Our main goal is to compare these two ROC curves; more precisely, we aim to test equivalence of ROC_Y and ROC_Z.

Taking into account the definition of ROC curves, the equivalence of ROC_Y and ROC_Z means that for any particular point $\boldsymbol{r}_Y \in \mathrm{ROC}_Y$ there exists an "identical" point $\boldsymbol{r}_Z \in \mathrm{ROC}_Z$, i. e. $\boldsymbol{r}_Y = \boldsymbol{r}_Z$. Equivalently, for any threshold $t_Y \in \mathcal{I}_Y$ the equivalence of the curves assures that we can find a threshold $t_Z \in \mathcal{I}_Z$ such that $\varrho(t_Y; F_0, F_1) = \varrho(t_Z; G_0, G_1)$. This allows us to express the ROC equivalence in terms of distribution functions, i. e.

$$\text{(4)} \quad \mathrm{ROC}_Y \equiv \mathrm{ROC}_Z \iff \forall\, t_Y \in \mathcal{I}_Y\ \exists\, t_Z \in \mathcal{I}_Z:\ F_0(t_Y) = G_0(t_Z)\ \&\ F_1(t_Y) = G_1(t_Z).$$

Due to (C1), all considered distribution functions are strictly increasing on $\mathcal{I}_Y$, $\mathcal{I}_Z$ respectively, so that there exist increasing transformation functions $\tau_0, \tau_1 : \mathcal{I}_Y \to \mathcal{I}_Z$ relating separately distribution functions in group $\mathcal{G}_0$ and $\mathcal{G}_1$. Define functions $\tau_0(t)$ and $\tau_1(t)$ such that $F_0(t) = G_0\big(\tau_0(t)\big)$ and $F_1(t) = G_1\big(\tau_1(t)\big)$, i. e.,

$$\text{(5)} \qquad \tau_0(t) = G_0^{-1}\big(F_0(t)\big) \quad \text{and} \quad \tau_1(t) = G_1^{-1}\big(F_1(t)\big) \quad \forall\, t \in \mathcal{I}_Y.$$

ROC curves consist of the values of distribution functions evaluated simultaneously at the same thresholds. Therefore, they are equivalent if and only if the groups $\mathcal{G}_0$ and $\mathcal{G}_1$ are related by the same threshold transformations $\tau_0(t) \equiv \tau_1(t)$. Hence, we may formulate the null hypothesis of the two ROC curves equivalence as

$$\text{(H)} \qquad \tau_0(t) = \tau_1(t) \quad \forall\, t \in \mathcal{I}_Y,$$

which we aim to test against the alternative

$$\text{(A)} \qquad \exists\ \mathcal{J}_Y \subseteq \mathcal{I}_Y, \mathcal{J}_Y \neq \emptyset, \quad \text{such that} \quad \tau_0(t) \neq \tau_1(t) \quad \forall\, t \in \mathcal{J}_Y.$$

Before deriving a test statistic, let us have a look at the transformations used. First, notice that the original problem of comparing two ROC curves is transformed into the problem of comparing behavior of the involved diagnostic methods on $\mathcal{G}_0$ and $\mathcal{G}_1$. Indeed, in order to have identical ROC curves it is not necessary that considered diagnostic methods behave exactly in the same manner, but that their behavior globally agrees both on the "positive" and the "negative" parts of the population. Globally it means that both methods correctly recognize the same proportion of $\mathcal{G}_0$ and $\mathcal{G}_1$ individuals, even though not necessarily the same individuals. Moreover, note that the transformations are not only technical tools but provide an interesting diagnostic approach as well. They have been studied extensively, e. g., by Doksum [7] and Doksum and Sievers [8], who proposed confidence regions and statistical inference about their shape.

To get insight into this concept, the upper row of plots in Figure 2 display empirical estimators

$$\text{(6)} \qquad \widehat{\tau}_0(t) = \widehat{G}_0^{-1}\big(\widehat{F}_0(t)\big) \quad \text{and} \quad \widehat{\tau}_1(t) = \widehat{G}_1^{-1}\big(\widehat{F}_1(t)\big), \qquad \forall\, t \in \mathcal{I}_Y,$$

of the transformation functions used for the three ROC pairs presented in Figure 1. The empirical CDF's $\widehat{F}_k(t)$ and $\widehat{G}_k(t), k = 0, 1$, are based on the samples $Y_{01}, \ldots, Y_{0n_0^Y}, Y_{11}, \ldots, Y_{1n_1^Y}, Z_{01}, \ldots, Z_{0n_0^Z}$, and $Z_{11}, \ldots, Z_{1n_1^Z}$, with a total sample size $n = n_0 + n_1 = n_0^Y + n_0^Z + n_1^Y + n_1^Z$. The quantile functions used are defined as $\widehat{G}_k^{-1}(u) = \inf\ \big\{t : \widehat{G}_k(t) > u\big\},\ k = 0, 1.$

We clearly see almost identical transformations in the central plot as expected in the case of equivalent ROC curves, while $\widehat{\tau}_0(t)$ and $\widehat{\tau}_1(t)$ have rather different forms in the other two cases. Another point of view is presented in lower plots of Figure 2. The transformation functions are plotted one against the other. Under the null hypothesis the obtained cloud of points should lie along the straight line with the unit-slope. In the central plot we see that a majority of points, with respect to supports of transformations, touches the line indicating ROC equivalence, while the points on the other plots are considerably far away from the expected null hypothesis line.

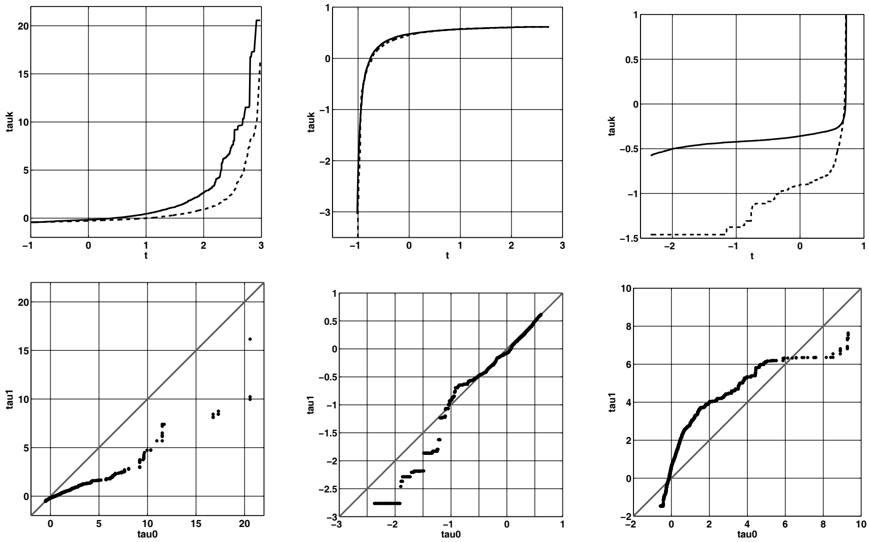

Fig. 2. *Transformation functions corresponding to the ROC curves plotted in Figure 1. The upper plots presents the form of $\widehat{\tau}_0(t)$ (solid lines) and $\widehat{\tau}_1(t)$ (dashed lines) depending on the threshold t, while the lower plots show transformation $\widehat{\tau}_0(t)$ plotted against $\widehat{\tau}_1(t)$.*

2.2. Test statistic

As illustrated by the graphs in Figure 2, transformation functions $\tau_0(t)$ and $\tau_1(t)$ indicate (non)equivalence of two ROC curves. Therefore, we suggest basing a decision on the distance between them. Precisely, we suggest a test statistic of the form

$$T_n = n \int_{\mathcal{I}_Y^*} \big(\widehat{\tau}_0(t) - \widehat{\tau}_1(t)\big)^2 \,\mathrm{d}t, \tag{7}$$

where the integral is on a closed interval $\mathcal{I}_Y^* \subseteq \mathcal{I}_Y$ such that the densities $g_0(s)$ and $g_1(s)$ are positive and finite for all s in the images of $\tau_0(t)$ and $\tau_1(t)$, $t \in \mathcal{I}_Y^*$, i.e.

$$\text{(C3)} \qquad 0 < g_0\big(\tau_0(t)\big) < \infty \quad \text{and} \quad 0 < g_1\big(\tau_1(t)\big) < \infty \quad \forall\, t \in \mathcal{I}_Y^*.$$

There is a lack of symmetry as concerns cdf's $F(x)$ and $G(x)$ in the definition of T_n inherited from the genesis of ROC curves. Our numerical calculations both with real and simulated data show, however, that its influence on the p values is quite negligible, especially when the size of the data is large.

As expected, test statistic T_n should be small under the null hypothesis and increase with growing difference between $\tau_0(t)$ and $\tau_1(t)$ under the alternative. Hence, if an appropriate critical value $c(\alpha)$ is available, the decision rule rejects the null hypothesis whenever $T_n > c(\alpha)$. Theorem 2.1 stated below establishes the asymptotic distribution of T_n under the null hypothesis (H).

Theorem 2.1. *Assume the setting described in Subsection 2.1 and the test statistic T_n defined by (7). Let conditions (C1) – (C3) hold and Y_0, Y_1, Z_0 and Z_1 be mutually independent. Let n_0 and n_1 tend to infinity such that $n_0^Y/n_0 \to \kappa_0$, $n_1^Y/n_1 \to \kappa_1$, $\kappa_0, \kappa_1 \in (0,1)$, and n tends to infinity such that $n/n_0 \to \kappa^0$ and $n/n_1 \to \kappa^1$, where $1/\kappa^0, 1/\kappa^1 \in (0,1)$. Then, under the null hypothesis (H), the test statistic T_n converges for $n \to \infty$ in distribution to the infinite weighted sum of independent χ_1^2 variables $\eta_1^2, \eta_2^2, \ldots$, i. e.*

$$T_n \xrightarrow{D} T^{\boldsymbol{B}} = \sum_{j=1}^{\infty} \lambda_j \eta_j^2, \tag{8}$$

where $\{\lambda_j\}$ represent the eigenvalues of the covariance operator of the zero-mean Gaussian process $B(t)$ with the covariance structure

$$\operatorname{cov}\big(B(s), B(t)\big) = c_0 \frac{F_0(s)\big(1 - F_0(t)\big)}{g_0\big(\tau_0(s)\big) g_0\big(\tau_0(t)\big)} + c_1 \frac{F_1(s)\big(1 - F_1(t)\big)}{g_1\big(\tau_1(s)\big) g_1\big(\tau_1(t)\big)}, \tag{9}$$

$s \leq t \in \mathcal{I}_Y^*$, $c_0 = \kappa^0/\big(\kappa_0(1-\kappa_0)\big), c_1 = \kappa^1/\big(\kappa_1(1-\kappa_1)\big)$.

Proof. Postponed to Appendix A. □

Asymptotic distribution of T_n is stated in Theorem 2.1 for independent realizations of independent diagnostic variables Y and Z. However, this condition is not always realistic in practice.

We think that the above test procedure behaves well for weakly dependent variables. However, when strong dependence is suspected, we suggest to use following two-step approach. The first step consists of determining separately critical values based on the limit processes of $\widehat{F}_k(t)$ and $\widehat{G}_k^{-1}(t)$, $k = 0, 1$ (see appendix A). A critical value for T_n can then be obtained by using a Bonferroni inequality as derived in Horváth et al. [14]. Of course, the accuracy of this procedure, and more generally the problem of dependence between diagnostic variables, should warrant a deep study of its own.

Taking into account the genesis of the test statistics, which is data dependent, its power against any alternative is of natural interest. Thus, the following theorem assures the consistency of the suggested test statistic.

Theorem 2.2. *Assume the setting and assumptions of Theorem 2.1 and the test statistic T_n defined by (7). Then this test is consistent against any alternative for which the conditions of Theorem 2.1 are satisfied.*

Proof. Postponed to Appendix A. □

2.3. Critical values

We have seen that the distribution of the test statistic can be approximated by the distribution of an infinite weighted sum of χ_1^2 variables $T^{\boldsymbol{B}} = \sum_{j=1}^{\infty} \lambda_j \eta_j^2$. As a practical matter, several problems have to be solved. First, we need to estimate unknown eigenvalues $\{\lambda_j\}$. Second, even if the eigenvalues were known, we would need to set an appropriate cut-off point and consider only a finite approximation of (8). Finally, even the finite approximation of $T^{\boldsymbol{B}}$ may still be quite complex and great attention has to be paid to obtain reliable critical values.

We start with estimating the eigenvalues of the covariance operator, say Γ, of the limit process $B(t)$. The covariance operator is a kernel operator whose kernel is formed by the covariance structure (9) of the underlying process, i. e.,

$$\Gamma\xi(t) = \int_{\mathcal{I}_Y^*} \operatorname{cov}\big(B(s), B(t)\big)\xi(s)\,\mathrm{d}s, \qquad \xi \in L^2(\mathcal{I}_Y^*). \tag{10}$$

Therefore, estimators of the eigenvalues of Γ can be based on the estimated $\operatorname{cov}\big(B(s), B(t)\big)$. For that purpose, we suggest using a plug-in estimator

$$\widehat{\operatorname{cov}}\big(B(s), B(t)\big) = c_0 \frac{\widehat{F}_0(s)\big(1-\widehat{F}_0(t)\big)}{\widetilde{g}_0\big(\widehat{\tau}_0(s)\big)\widetilde{g}_0\big(\widehat{\tau}_0(t)\big)} + c_1 \frac{\widehat{F}_1(s)\big(1-\widehat{F}_1(t)\big)}{\widetilde{g}_1\big(\widehat{\tau}_1(s)\big)\widetilde{g}_1\big(\widehat{\tau}_1(t)\big)},$$

where $s, t \in \{t_1, \ldots, t_p\} \subset \mathcal{I}_Y^*$ and $\widehat{F}_k(t),\ k = 0, 1$, are the empirical CDFs, $\widehat{\tau}_k(t)$ are given by (6), and $\widetilde{g}_k(t)$ stands for the kernel estimators of the densities $g_k(t)$. For details see, e. g., Silverman [26]. The covariance operator Γ then can be approximated by its discrete estimated version

$$\widehat{\Gamma}_{n,p} = \Big(\omega_i\, \widehat{\operatorname{cov}}\big(B(t_i), B(t_j)\big)\Big)_{i,j=1}^{p}, \tag{11}$$

where ω_i stands for the weights used for the numerical quadrature replacing theoretical integration in (10) by discrete summation over $\{t_1, \ldots, t_p\}$. Another possibility is to use $\omega_i = t_i - t_{i-1}$. Spectral decomposition of the matrix $\widehat{\Gamma}_{n,p}$ then provides consistent estimators $\big\{\widehat{\lambda}_j\big\}$ of the asymptotic eigenvalues $\big\{\lambda_j\big\}$.

Values of c_i's are in practice established by the data, as seen in Theorem 2.1. The real problem can arise when the proportion of $\mathcal{G}_0$ elements – and therefore also of $\mathcal{G}_1$ elements – is extreme, i. e., very close to zero or n. Regarding the value of p, it follows from our calculations that it is preferable to keep the grid of values t_i as dense as possible, of course, to be able to estimate $\Gamma\xi(t)$. We used $p = 10^3$ for our calculations.

Theorem 2.3. *Assume that kernel density estimators $\widetilde{g}_k(t)$ are based on continuous, bounded, compactly supported kernels and on bandwidths $\{h_k\}$ such that, $h_k \to 0$ and $h_k n_k^Z / \log(n_k^Z) \to \infty$ for $n_k^Z \to \infty, k = 0, 1$. Then, under the conditions of Theorem 2.1, it holds*

$$\Big|\widehat{\lambda}_j - \lambda_j\Big| \xrightarrow{P} 0 \quad \textit{as} \quad n \to \infty, \quad j = 1, 2, \ldots$$

Proof. Postponed to Appendix A. □

Suppose that the described estimation procedure results in J positive eigenvalues that allow approximation of the infinite representation of $T^{\boldsymbol{B}}$ by its first J components, i. e.,

$$T^{\boldsymbol{B}} \approx \sum_{j=1}^{J} \widehat{\lambda}_j \eta_j^2 \equiv S^J, \tag{12}$$

where $\eta_1^2, \ldots, \eta_J^2$ stand for independent χ^2 variables with one degree of freedom. In our calculations we set J in such a way that we have used all eigenvalues larger than 10^{-10}.

As distribution of S^J is not explicitly known, we can perform Monte Carlo simulation to obtain the desired critical value. The simulation scheme is straightforward:

```
1. FOR $k = 1 : K$
2.      Simulate $J$ independent $\chi_1^2$ variables $\eta_1^2, \dots, \eta_J^2$
3.      Calculate the value of $S^J$ and store it to $S_k^J$
4. ENDFOR
```

Once the sample $S_1^J, \dots, S_K^J$ is available, we form standard empirical distribution and quantile functions and use estimated quantiles instead of the unknown exact ones. If extreme quantiles are required, more sophisticated rare event methods based on properly tuned importance sampling or saddle point approximation should be used to obtain reliable results.

Concerning computational costs, performing sufficiently many ($K \approx 10^6$) simulations for $J \approx 1000$ components is feasible on a standard “home” computer in a couple of seconds. We point out that taking squares of standard normal variables is considerably faster, mainly for a large J value, than a direct simulation of χ^2 variables, especially if a matrix language such as Matlab, e. g., is available. Notice that far fewer simulations are required to get critical values for the test statistic (8) at standards α-levels. Typically $K = 10^4$ is enough. However, in our context one needs reasonably exact p-values for small values of p, making it necessary to run a large number of simulations in order to obtain a reliable estimator of the tail of the distribution.

Kac and Siegert [16] have shown that the characteristic function of $T^{\boldsymbol{B}}$ takes the form

$$\psi_{T^{\boldsymbol{B}}}(\varsigma) = \mathrm{E}\,\exp\{i\varsigma T^{\boldsymbol{B}}\} = \prod_{j=1}^{\infty}(1 - i\varsigma\lambda_j)^{-1/2}, \quad \varsigma \in \mathbb{R},$$

so that the inverse formula by Gil-Pelaez [10] provides the distribution function of $T^{\boldsymbol{B}}$, i. e.,

$$\mathsf{P}(T^{\boldsymbol{B}} \le s) = H_{T^{\boldsymbol{B}}}(s) = \frac{1}{2} - \frac{1}{\pi}\int_0^{\infty} \Im\left(\frac{e^{-i\varsigma s}\psi_{T^{\boldsymbol{B}}}(\varsigma)}{\varsigma}\right) \mathrm{d}\varsigma, \quad s \ge 0, \tag{13}$$

where $\Im(z)$ stands for the complex part of a complex number $z \in \mathbb{C}$.

If $T^{\boldsymbol{B}}$ is approximated by S^J, Imhof [15] suggested to represent its distribution function by

$$\mathsf{P}(S^J < s) = \frac{1}{2} - \frac{1}{\pi}\int_0^{\infty} \frac{\sin\theta(s,u)}{u\rho(u)}\,\mathrm{d}u, \tag{14}$$

where $2\theta(s,u) = \sum_{j=1}^J \arctan\left(\widehat{\lambda}_j u\right) - su$, $\rho(u) = \prod_{j=1}^J \left(1 + \widehat{\lambda}^2 u^2\right)^{1/4}$. In practice, the integration in (14) has to be carried over a finite range $0 \le u \le U$. Imhof [15] claims that the truncation error is satisfactorily small and provides its upper bound $\left(JU^J\right)^{-1} \prod_{j=1}^J \widehat{\lambda}_j^{-1/2}$. However, our numerical experiments show that the integration of (14) must be performed extremely carefully with either a very fine step of the order 10^{-6} or rather tricky weighting. We point out that a naive use of numerical quadrature often leads to the values of distribution function greater than one, which is, of course, an unacceptable property. As one does not obtain an adequate precision gain with respect to the computational costs of Imhof’s procedure, simulations turn out to be the most favorable in practice.

2.4. Kernel estimator

The methodology described above is based on the use of the empirical estimators of distribution and quantile functions $F_k(t), G_k^{-1}(p), k = 0, 1$. Evidently, to estimate cdf's $F_k(t), k = 0, 1$, one might use the kernel estimators instead, i. e.,

$$\widetilde{F}_k(t) = \frac{1}{n_k^Y} \sum_{i=1}^{n_k^Y} H\left(\frac{t - Y_{ki}}{h_k}\right), \qquad t \in \mathbb{R},\ k = 0, 1, \tag{15}$$

where $H(\cdot)$ is an appropriate cumulative kernel function and the bandwidth parameters h_0 and h_1 control the smoothness of estimators. Analogously, kernel estimators $\widetilde{G}_k^{-1}(p)$ might be used to estimate quantile functions $G_k^{-1}(p)$, where $\widetilde{G}_k^{-1}(p) = \inf\left\{t : \widetilde{G}_k(t) > p\right\},\ p \in (0, 1),\ k = 0, 1$.

Combining these two kernel estimators and following ideas of Section 2.1 we naturally come to the kernel analogue of the empirical transformation functions (6), i. e., to

$$\widetilde{\tau}_k(t) = \widetilde{G}_k^{-1}\big(\widetilde{F}_k(t)\big), \qquad t \in \mathcal{I}_Y,\ k = 0, 1. \tag{16}$$

Consequently, in the definition (7) one can replace the empirical transformations $\widehat{\tau}_k(t)$ with the kernel ones $\widetilde{\tau}_k(t)$ and obtain the kernel analogue of the test statistic T_n. As one might expect, both Theorem 2.1 and Theorem 2.3 hold for the kernel type test statistic as well (see Appendix A for a formal proof). Hence, in the practice of performing the test procedure, one may follow the same "lines" both for the empirical and the kernel estimators.

It is well known that the kernel estimators offer some advantages compared to their empirical analogues. The most important is probably the fact that kernel smoothing typically brings a better "visual" effect as it provides a continuous curve in the ROC square instead of discrete points of an empirical ROC curve. On the other hand, if smoothing parameters are not properly chosen, the kernel type test statistic may lead to irrelevant and unreliable results.

The kernel CDF estimator has been proposed and studied for the first time by Nadaraya [20]. Concerning the kernel ROC curves, one finds the proposals in, e. g., Zhou et al. [33] or Lloyd [19]. The last paper has been followed-up by an interesting paper of Hall et al. [12]. Later, Prchal [24] suggested an automatic procedure that, by means of data transformation, improves accuracy of kernel ROC curves.

Appendix A: Proofs

Theorem 2.1 is stated for the test statistic T_n defined by (7), which is based on the empirical estimators of the distribution and quantile functions. However, we provide its proof for a more general class of estimators satisfying conditions (P1) and (P2) listed below.

Let $Y_1, \ldots, Y_m$ and $Z_1, \ldots, Z_n$ be i.i.d. samples with respective continuous distribution functions $F(t)$ and $G(t)$ such that the supports of their densities are real intervals $\mathcal{I}_Y$ and $\mathcal{I}_Z$. Let $\mathcal{I}_Y^* \subseteq \mathcal{I}_Y$ be a closed interval such that $0 < g\left(G^{-1}\big(F(t)\big)\right) < \infty,\ \forall\, t \in \mathcal{I}_Y^*$. Let $\widehat{F}(t)$ and $\widehat{G}(t)$ be the estimators of $F(t)$ and $G(t)$, and $\widehat{G}^{-1}(u) = \inf\{t : \widehat{G}(t) > u\}$ be an estimator of the quantile function $G^{-1}(u)$, such that

$$\sup_{t \in \mathcal{I}_Y^*} \left|\widehat{F}(t) - F(t)\right| \xrightarrow{a.s.} 0, \tag{P1}$$

(P2) $\sqrt{m}\big(\widehat{F}(t)-F(t)\big) \xrightarrow{D} W_1\big(F(t)\big)$ & $\sqrt{n}\big(\widehat{G}(t)-G(t)\big) \xrightarrow{D} W_2\big(G(t)\big)$, $\forall\, t \in \mathcal{I}_Y^*$,

where W_1 and W_2 stand for independent Brownian bridges.

The first step of proving Theorem 2.1 concerns a weak convergence result of an estimated quantile process.

Lemma A.1. *Let m and n tend to infinity such that $m/(n+m) \to \kappa \in (0,1)$. Then, under the conditions (P1) and (P2),*

$$\sqrt{m+n}\Big(\widehat{G}^{-1}\big(\widehat{F}(t)\big)-G^{-1}\big(F(t)\big)\Big) \xrightarrow{D} \frac{1}{\sqrt{\kappa(1-\kappa)}}\frac{1}{g\big(G^{-1}\big(F(t)\big)\big)}W\big(F(t)\big), \qquad t \in \mathcal{I}_Y^*, \tag{17}$$

where $\big\{W(s), s \in [0,1]\big\}$ denotes a Brownian bridge defined on $[0,1]$.

Proof. First, notice that $\sqrt{m+n}\Big(\widehat{G}^{-1}\big(\widehat{F}(t)\big) - G^{-1}\big(F(t)\big)\Big)$ can be decomposed as

$$\sqrt{m+n}\Big(\widehat{G}^{-1}\big(\widehat{F}(t)\big) - G^{-1}\big(\widehat{F}(t)\big)\Big) + \frac{G^{-1}\big(\widehat{F}(t)\big) - G^{-1}\big(F(t)\big)}{\widehat{F}(t) - F(t)}\sqrt{m+n}\big(\widehat{F}(t) - F(t)\big), \qquad t \in \mathcal{I}_Y^*. \tag{18}$$

The second term, using (P1), (P2) and the same arguments as in the proof of Theorem 4.1 by Doksum [7], converges in distribution to

$$\frac{1}{\sqrt{\kappa}}\frac{1}{g\big(G^{-1}\big(F(t)\big)\big)}W_1\big(F(t)\big), \quad \forall\, t \in \mathcal{I}_Y^*.$$

Further, from (P2) and (3.4) in Ralescu and Puri [25] we can deduce that

$$\sup_{u=F(t),\ t\in\mathcal{I}_Y^*}\Big|\sqrt{m+n}\Big(\widehat{G}^{-1}(u) - G^{-1}(u)\Big) - U(u)\Big| \xrightarrow{P} 0, \tag{19}$$

where $U(u) \equiv \Big(\sqrt{1-\kappa}g\big(G^{-1}(u)\big)\Big)^{-1}V(u)$ and V stands for a Brownian bridge independent of W_1. Note that when $\widehat{G}(.)$ is the empirical function, (19) can be deduced from results stated by Kiefer (1970, 1972), see Theorems 4.3.2 and 5.2.1 in Csörg¨ and Révész [5]. Together with (P1) and continuity arguments we obtain

$$\begin{aligned}\sup_{t\in\mathcal{I}_Y^*}&\Big|\sqrt{m+n}\Big(\widehat{G}^{-1}\big(\widehat{F}(t)\big) - G^{-1}\big(\widehat{F}(t)\big)\Big) - U\big(\widehat{F}(t)\big) + U\big(\widehat{F}(t)\big) - U\big(F(t)\big)\Big|\\ &\le \sup_{0\le u\le 1}\Big|\sqrt{m+n}\Big(\widehat{G}^{-1}(u) - G^{-1}(u)\Big) - U(u)\Big| + \sup_{t\in\mathcal{I}_Y^*}\Big|U\big(\widehat{F}(t)\big) - U\big(F(t)\big)\Big|,\end{aligned}$$

that converges to 0 in probability. □

Proof of Theorem 2.1

Proof. If $\widehat{F}(t)$ stands for the empirical CDF estimator, the property (P1) is satisfied due to the well-known Glivenko–Cantelli theorem, whereas the proof of (P2) can be found, e. g., in Billingsley [4]. Hence, Lemma A.1 holds for this case and with continuity of L_2 norm with respect to the Skorochod topology it assures

$$T_n \xrightarrow{D} \int_{\mathcal{I}_Y^*} B^2(t)\,\mathrm{d}t, \tag{20}$$

where $\{B(t), t \in \mathcal{I}_Y^*\}$ is a zero-mean Gaussian process with the covariance structure given by (9). As $\mathrm{E}\, B^2(t) < \infty\ \forall t \in \mathcal{I}_Y^*$, $B(t)$ admits the Karhunen-Loève decomposition

$$B(t) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\eta_j v_j(t),$$

where η_j are real random variables following the standard normal distribution and $\{v_j\}$ is the orthonormal system of the eigenfunctions corresponding to the eigenvalues $\{\lambda_j\}$ of the covariance operator Γ of $\{B(t), t \in \mathcal{I}_Y^*\}$. It follows from Kac and Siegert [16] that

$$\int_{\mathcal{I}_Y^*} B^2(t)\,\mathrm{d}t = \int_{\mathcal{I}_Y^*} \left(\sum_{j=1}^{\infty} \sqrt{\lambda_j}\eta_j v_j(t)\right)^2 \mathrm{d}t = \sum_{j=1}^{\infty} \lambda_j \eta_j^2, \tag{21}$$

which assures the statement of Theorem 2.1. □

Proof of Theorem 2.2

Proof. We have shown above that, under the assumptions of Theorem 2.1, Lemma A.1 holds. Thus

$$n \int_{\mathcal{I}_Y^*} \left(\widehat{\tau}_0(t) - \tau_0(t) - \widehat{\tau}_1(t) + \tau_1(t)\right)^2 \mathrm{d}t \xrightarrow{D} \int_{\mathcal{I}_Y^*} B^2(t)\,\mathrm{d}t, \tag{22}$$

where $\{B(t),\ t \in \mathcal{I}_Y^*\}$ is a zero-mean Gaussian process with the covariance structure given by (9). Under an alternative hypothesis and a given critical value t_α, the probability of rejecting the null hypothesis is $P(T_n > t_\alpha)$. Using (22) we have

$$\lim_{n\to\infty} P\big(T_n > t_\alpha\big) \longrightarrow 1,$$

what proves consistency of the test. □

Remark. As pointed out in Section 2.4, Theorem 2.1 remains valid when the kernel CDF estimators are used. Indeed, property (P1) is due to Nadaraya [20], while Nixdorf [21] has shown (P2).

Proof of Theorem 2.3

Proof. According to the Glivenko–Cantelli theorem one has for $k = 0, 1$

$$\begin{aligned} &\sup_{s,t\in\mathcal{I}_Y^*} \left|\widehat{F}_k(s)\big(1-\widehat{F}_k(t)\big) - F_k(s)\big(1-F_k(t)\big)\right| \\ &\quad = \sup_{s,t\in\mathcal{I}_Y^*} \left|\big(1-\widehat{F}_k(t)\big)\big(\widehat{F}_k(s)-F_k(s)\big) + F_k(s)\big(F_k(t)-\widehat{F}_k(t)\big)\right| \xrightarrow{a.s.} 0. \end{aligned} \tag{23}$$

Further, Bertrand-Retali [3] has shown that

$$\sup_{t\in\mathcal{I}_Y^*} \left|\widetilde{g}_k(t) - g_k(t)\right| \xrightarrow{a.s.} 0. \tag{24}$$

For validity of

$$\sup_{\epsilon<u<1-\epsilon} \left|\widehat{G}_k^{-1}(u) - G_k^{-1}(u)\right| \xrightarrow{a.s.} 0 \tag{25}$$

see Van der Vaart and Wellner [28] and the references therein. Combining (23), (24) and (25) leads to

$$\sup_{s,t\in\mathcal{I}_Y^*}\left|\widehat{\operatorname{cov}}\big(B(s),B(t)\big)-\operatorname{cov}\big(B(s),B(t)\big)\right| \xrightarrow{a.s.} 0. \tag{26}$$

The statement of Theorem 2.3 now follows from (26) and result (15) in Yao et al. [31]. □

References

[1] Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psych.* **12** 387 – 415.

[2] Beam, C. A. and Wieand, H. S. (1991). A statistical method for the comparison of a discrete diagnostic test with several continuous diagnostic tests. *Biometrics* **47** 907 – 919.

[3] Bertrand-Retali, M. (1978). Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Rev. Roumaine Math. Pures Appl.* **23** 361 – 385. (In French)

[4] Billingsley, P. (1968). *Convergence of Probability Measures.* J. Wiley, New York.

[5] Csörgő, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics.* Academic Press, New York.

[6] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrika* **44** 837 – 846.

[7] Doksum, K. A. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Statist.* **2** 267 – 277.

[8] Doksum, K. A. and Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika* **63** 421 – 434.

[9] Falk, M. (1983). Relative efficiency and deficiency of kernel type estimator of smooth distribution functions. *Statist. Neerlandica* **37** 73 – 83.

[10] Gil-Pelaez, J. (1951). Note on the inversion theorem. *Biometrika* **38** 481 – 482.

[11] Greenhouse, S. W. and Mantel, N. (1950). The evaluation of diagnostic tests. *Biometrics* **6** 399 – 412.

[12] Hall, P., Hyndman, R. J., and Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curves, *Biometrika* **91** 743 – 750.

[13] Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the area under two ROC curves derived from the same cases. *Radiology* **148** 839 – 843.

[14] Horváth, L., Horváth, Z., and Zhou, W. (2008). Confidence bands for ROC curves. *J. Stat. Plann. Inference* **138 (6)** 1894 – 1904.

[15] Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48** 419 – 426.

[16] Kac, M. and Siegert, J. F. (1947). An explicit representation of a stationary gaussian process. *Ann. Math. Stat.* **18** 438 – 442.

[17] Kiefer, J. (1970). Deviations between the sample quantile process and the sample distribution functions. *In: Nonparametric Techniques in Statistical Inference (M. L. Puri, Ed.) 299 – 319, Cambridge University Press, London.*

[18] Kiefer, J. (1972). Skorohod embedding of multivariate rvs and the sample df. Z. Wahrscheinlichkeitstheorie verw. Gebiete **24** 1 – 35.

[19] Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *J. Amer. Statist. Assoc.* **93** 1356 – 1364.

[20] Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory Probab. Appl.* **15** 497 – 500.

[21] Nixdorf, R. (1985). Central limit theorem in $C[0,1]$ for a class of estimators of a distribution function. *Statist. Neerlandica* **39** 251 – 260.

[22] Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction. *In: Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Poster Sessions, Sydney.*

[23] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, Oxford.

[24] Prchal, L. (2007). Kernel ROC curve estimator for skewed diagnostic variables. *Preprint, UPS Toulouse.*

[25] Ralescu, S. S. and Puri, M. L. (1996). Weak convergence of sequence of first passage processes and applications. *Stochastic Process. Appl.* **62** 327 – 345.

[26] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, New York.

[27] Strawderman, R. L. (2004). Computing tail probabilities by numerical Fourier inversion: The absolutely continuous case, *Statistica Sinica* **14** 175 – 201.

[28] Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics.* Springer, New York.

[29] Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* **83** 835 – 848.

[30] Wieand, S., Gail, M. H., James, B. R., and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired and unpaired data. *Biometrika* **76** 585 – 592.

[31] Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577 – 590.

[32] Zhou, X. H., McClish, D. K., and Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine.* J. Wiley, New York.

[33] Zou, K. H., Hall, W. J., and Shapiro, D. E. (1997). Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Stat. Med.* **16** 2143 – 2156.