



Visualizing the decision rules behind the ROC curves: understanding the classification process

Sonia Pérez-Fernández¹ · Pablo Martínez-Cambor² · Peter Filzmoser³ · Norberto Corral¹

Received: 30 September 2019 / Accepted: 14 October 2020 / Published online: 13 November 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

The receiver operating characteristic (ROC) curve is a graphical method commonly used to study the capacity of continuous variables (markers) to properly classify subjects into one of two groups. The decision made is ultimately endorsed by a classification subset on the space where the marker is defined. In this paper, we study graphical representations and propose visual forms to reflect those classification rules giving rise to the construction of the ROC curve. On the one hand, we use static pictures for displaying the classification regions for univariate markers, which are specially convenient when there is not a monotone relationship between the marker and the likelihood of belonging to one group. In those cases, there are two options to improve the classification accuracy: to allow for more flexibility in the classification rules (for example considering two cutoff points instead of one) or to transform the marker by using a function whose resulting ROC curve is optimal. On the other hand, we propose to build videos for visualizing the collection of subsets when several markers are considered simultaneously. A compilation of techniques for finding a rule that maximizes the area under the ROC curve is included, with a focus on linear combinations. We present a tool for the R software which generates those graphics, and we apply it to one real dataset. The R code is provided as Supplementary Material.

Keywords Area under the curve · Classification regions · Graphical animations · Multivariate marker · Receiver operating characteristic curve

The authors gratefully acknowledge support by the Grants MTM2015-63971-P from the Spanish Ministerio of Economía y Competitividad and by FC-15-GRUPIN14-101 and Severo Ochoa Grant BP16118 from the Principado de Asturias and Grant from Campus of International Excellence of University of Oviedo (the last two ones for Pérez-Fernández).

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10182-020-00385-2>) contains supplementary material, which is available to authorized users.

✉ Sonia Pérez-Fernández
perezsonia@uniovi.es

Extended author information available on the last page of the article

1 Introduction

As a supervised learning technique, classification is a statistical method whose final objective is to build a grouping rule based on one or various markers collected in a training dataset where the response variable is also known. With that rule, classifications of new subjects can be done on the basis of their marker values (Nielsen et al. 2009). Going into good classifications is important in many fields such as medical diagnosis, machine learning, data mining or business intelligence.

This paper covers binary classifications. The goal is to discriminate between *positive* and *negative* populations based on a continuous variable making the error probabilities as small as possible. There are two potential errors: classifying a positive subject as negative (*false negative*) and classifying a negative subject as positive (*false positive*). The trade-off between those error probabilities is reflected in the trace of the *Receiver Operating Characteristic (ROC) curve* (McClish and Powell 1989).

The decision rules which determine the classification of the subjects depend on the allocation of their marker values inside—classed as positive—or outside—classed as negative—a region called *classification subset* (Martínez-Cambor and Pardo-Fernández 2019). The ROC curve displays the probability of correctly classifying positive subjects (*sensitivity* or *true-positive rate*) versus the probability of incorrectly classifying negative subjects (*1-specificity* or *false-positive rate*) for every classification subset of the form (c, ∞) with $c \in \mathbb{R}$. For each *threshold* or *cut-off point* c of the marker, the sensitivity is $Se(c) = \mathcal{P}\{\xi > c\}$ and the specificity is $Sp(c) = \mathcal{P}\{\chi \leq c\}$, where χ and ξ denote the variables modeling the marker values in the negative and positive population, respectively. The substitution $t = 1 - Sp(c)$ yields the definition of the *standard ROC curve* as a function of t . For each $t \in [0, 1]$, there is a subset, $s_t \in \mathcal{I}_r(t)$, where

$$\mathcal{I}_r(t) = \{s = (c, \infty) \subset \mathbb{R} \text{ such that } \mathcal{P}\{\chi \in s\} \leq t\}, \tag{1}$$

reporting the maximum sensitivity, i.e., $s_t = \arg \sup_{s \in \mathcal{I}_r(t)} \mathcal{P}\{\xi \in s\}$. The supremum of the sensitivity reported by s_t defines the value of the ROC curve in t , $\mathcal{R}(t)$. $\mathcal{I}_r(t)$ is referred to as a particular family of eligible classification subsets and its form comes from assuming that higher values of the marker are associated with a higher likelihood of being a positive subject.

This procedure can be generalized and mathematically formulated as follows:

$$\begin{aligned} [0, 1] &\longrightarrow \mathcal{B}_X \longrightarrow [0, 1] \\ t &\rightsquigarrow \mathcal{I}(t) \rightsquigarrow \mathcal{R}(t) := \sup_{s \in \mathcal{I}(t)} \mathcal{P}\{\xi \in s\} \end{aligned} \tag{2}$$

where \mathcal{B}_X denotes the Borel subsets of the space where the marker X is defined, t denotes the false-positive rate, $\mathcal{I}(t)$ is a particular family of eligible classification subsets and $\mathcal{R}(t)$ is the value of the ROC curve in t . By definition, the regions s in a *family of eligible classification subsets* $\mathcal{I}(t)$ should have a fixed shape and fulfill $\mathcal{P}\{\chi \in s\} \leq t$, i.e., their specificity is, at least, $1 - t$.

The *Area Under the Curve* (AUC) is a measure of the global classification accuracy of a marker (Bamber 1975). It takes values in the interval $[0.5, 1]$, and the closer to 1 the better the accuracy. The AUC is a standard for measuring the global precision to the extent that some authors such as Kauppi (2016) and Chen et al. (2016) seek to find the transformation of the marker which maximizes it. In such sense, McIntosh and Pepe (2002) proved that the transformation which maximizes the sensitivity for any fixed specificity t is the likelihood ratio. From here on, we will refer to the ROC curve resulting from the latter transformation as the *efficient ROC curve*, eROC curve, following the notation used by Kauppi (2016). It should be noted that transforming the marker has an effect on the classification subsets. They are of the form $s_t^* = (c_t, \infty)$ over the transformed marker, but the corresponding regions s_t over the original marker may be difficult to interpret and even useless for the practitioner.

With the goal of achieving a better classification accuracy, an approach different from finding a transformation can be considered: allowing the classification subsets to be more flexible. As an example, Martínez-Cambor et al. (2017) proposed the *general ROC curve*, gROC curve, considering subsets of the form $(-\infty, x^L] \cup (x^U, \infty)$, with $x^L \leq x^U \in \mathbb{R}$, to accommodate those situations where not only higher values of the marker are associated with a higher likelihood of being a positive but both extremes, lower and higher, are. It should be highlighted that in the standard ROC curve, for any $t \in [0, 1]$, the classification region is univocally defined by $s_t = (F_\chi^{-1}(1 - t), \infty)$. However, in the gROC curve, there is not a direct and explicit expression of the classification rule s_t giving rise to the corresponding point $(t, \mathcal{R}(t))$ of the curve because there are infinite subsets of the form $s = (-\infty, x^L] \cup (x^U, \infty)$ with specificity $1 - t$ (i.e., $\mathcal{P}\{\chi \in s\} = t$). Hence, the corresponding s_t involves the search for the subset fulfilling the previous conditions which reports the maximum sensitivity, i.e., $s_t = \arg \sup_{s \in \mathcal{J}_g(t)} \mathcal{P}\{\xi \in s\}$ with $\mathcal{J}_g(t) = \{s = (-\infty, x^L] \cup (x^U, \infty) \subset \mathbb{R} \text{ such that } \mathcal{P}\{\chi \in s\} \leq t\}$. Therefore, in the construction of the gROC curve, illustrating the underlying decision rules is essential not to lose the sight of the problem under study. The *area under the gROC curve*, gAUC, is also a global index of accuracy of the marker for the family of classification rules of the form $(-\infty, x^L] \cup (x^U, \infty)$.

Due to the relevance of visualizing the decision rules behind each point of the ROC curve, in this paper we study and propose procedures for including them in the graphical representation of the outcomes of classification analysis. With the aim of displaying the classification accuracy of a marker, it is common to build the ROC curve, but not to see the underlying rules, suggesting that the classification problem is blurred. A key strength of the proposed graphical solution is the visualization of the curve, $(t, \mathcal{R}(t))$, and the underlying classification subsets, s_t , on a single figure. With this goal, we have developed a tool compressed as an R package called `movieROC`.

In practice, it is common to collect several markers for improving the performance of a binary classification based on them, giving rise to a new problem: defining the ROC curve of a *multivariate marker* to evaluate its classification accuracy. To cope with it, what is mostly done in the literature is to find a transformation from the original

space \mathbb{R}^p ($p \geq 2$) to \mathbb{R} such that the final marker reports a large AUC. In this multivariate scenario, we study how to monitor the change of those regions, which appear to be omitted in the transformation process, generating the ROC curve. Since it is no longer feasible to display the regions s_t for every $t \in [0, 1]$ in one simple 2D graphic, we propose to use dynamic graphical representations.

In order to introduce the background of the graphical procedures studied and the different functionalities of the new R package (provided as Supplementary Material), this article is organized as follows. In Sect. 2, we provide some remarks about the gROC and eROC curves and propose a computational algorithm to ensure a rational assumption over the decision rules. In Sect. 3, we review some techniques to deal with the ROC curves for multidimensional markers and we propose two ways of visualizing different classification regions. In Sect. 4, we apply our new tool to a real dataset, displaying the classification subsets and the ROC curve construction resulting from the methods addressed previously. Finally, Sect. 5 provides some concluding remarks and feedback.

2 ROC curve for univariate markers

The *standard ROC curve*, generally defined in (2) considering $\mathcal{I}_r(t)$ in (1) as the family of eligible classification subsets, has been deeply studied and popularized since it was first developed for radar signal detection (Green and Swets 1966). Several theoretical and practical aspects have been addressed regarding its standard definition; the interested reader is referred to Pepe (2003). Some techniques for non-standard analyses related to the ROC curve have been implemented in the nsROC package (Pérez-Fernández et al. 2018).

With the goal of improving the classification accuracy of a marker when the differences between the two populations are not directly in location but also in dispersion, specially when the relationship between the marker and the likelihood of being a positive subject is not monotone but U-shape, the family of eligible classification subsets is extended to

$$\mathcal{I}_g(t) = \{s = (-\infty, x^L] \cup (x^U, \infty) \subset \mathbb{R} \text{ such that } \mathcal{P}\{\chi \in s\} \leq t\}, \quad (3)$$

resulting in the so-called *general ROC curve* (gROC curve). The gROC curve has been clearly motivated in the existing literature (Martínez-Cambor et al. 2017; Floege et al. 2011; Gardner et al. 2016).

Nevertheless, the vast majority of literature does not address the underlying regions, but tackles the construction of the ROC curve directly using the standard definition. When the relationship between the marker and the likelihood of being a positive subject is not monotone, they find a functional transformation of the marker whose result maximizes the standard AUC. If we consider a continuous transformation h , the family of eligible classification subsets is

$$\mathcal{A}^{h(X)}(t) = \{s^* = (c, \infty) \subset \mathcal{B}(h) \text{ such that } \mathcal{P}\{h(\chi) \in s^*\} \leq t\} \quad (4)$$

where $\mathcal{R}(h)$ denotes the codomain of the function $h(\cdot)$, leading to the ROC curve

$$\mathcal{R}_h(t) = \sup_{s^* \in \mathcal{H}^{(X)}(t)} \mathcal{P}\{h(\xi) \in s^*\}. \tag{5}$$

When the function $h(\cdot)$ considered is the likelihood ratio, such ROC curve is called the *efficient ROC curve* (eROC curve). In practice, the density functions of the marker in both populations are unknown and thus the problem arises in estimating the referred function $h(\cdot)$ or any monotone transformation of it.

If $h(\cdot)$ is a monotone increasing transformation, the classification subsets for the initial marker are of the form $s_t = (h(c_t), \infty)$ where $c_t \in \mathbb{R}$ is such that $\mathcal{P}\{\chi > c_t\} = t$ and the ROC curve is the standard one. However, if $h(\cdot)$ is not monotone, the subsets for the original marker, s_t , may not be only one interval. For example, if $h(\cdot)$ is a polynomial function of degree n , those may be the union of even $\lfloor n/2 + 1 \rfloor$ intervals, where $\lfloor \cdot \rfloor$ denotes the floor function. Figure 1 reflects the classification subsets from a cubic splines (with seven knots) transformation.

In Martínez-Cambolor et al. (2019), an analysis of a particularly intricate real-world dataset is provided, illustrating the differences between the two approaches: the eROC and the gROC curves. The authors claimed that improving the AUC by means of different functional transformations can result in an overfitting or classification regions with no practical interpretability, so it may not be worthwhile to find the eROC curve with the highest area. Besides, they proved an important result linking gROC and eROC curves: *If the variable in both negative and positive population is normally distributed, the gROC curve is based on the optimal classification rules.* That is, in a binormal scenario, $\mathcal{R}_h(t) = \mathcal{R}_g(t) = \sup_{s \in \mathcal{I}_g(t)} \mathcal{P}\{\xi \in s\}$.

2.1 The efficient ROC curve: the eROC curve

Mathematically, the procedure of defining the ROC curve coming from a transformation $h(\cdot)$ of the marker can be formally formulated as follows:

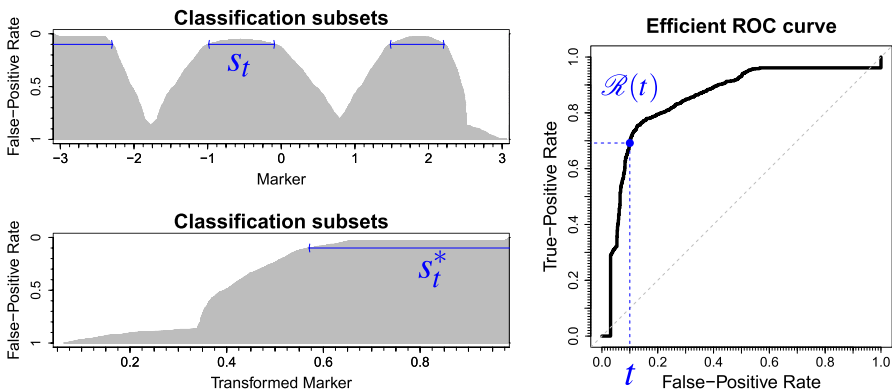


Fig. 1 Efficient ROC curve. Left, the classification subsets (gray) for every false-positive rate for the transformed (bottom) and the original marker (top), s_t^* and s_t , respectively; right, the eROC curve

$$\begin{aligned}
 [0, 1] &\longrightarrow \mathcal{B}_{h(X)} \longrightarrow \mathcal{B}_X \longrightarrow [0, 1] \\
 t &\rightsquigarrow \mathcal{I}^{h(X)}(t) \rightsquigarrow \mathcal{I}^X(t) \rightsquigarrow \mathcal{R}(t) := \sup_{s \in \mathcal{I}^X(t)} \mathcal{P}\{\xi \in s\} = \sup_{s^* \in \mathcal{I}^{h(X)}(t)} \mathcal{P}\{h(\xi) \in s^*\}
 \end{aligned}$$

where \mathcal{B}_X and $\mathcal{B}_{h(X)}$ denote the Borel subsets of the space where the original and the transformed marker, respectively, are defined; t denotes the false-positive rate; $\mathcal{I}^{h(X)}(t)$ is the family of eligible classification subsets for the transformed marker defined in (4); $\mathcal{I}^X(t)$ is the family of regions s for the original marker resulting from applying the inverse function $h^{-1}(\cdot)$ over the subsets $s^* \in \mathcal{I}^{h(X)}(t)$; and $\mathcal{R}(t)$ is the value of the ROC curve in t .

McIntosh and Pepe (2002) proved that the optimal transformation, in terms of achieving the highest sensitivity for any fixed specificity and consequently reporting the optimal AUC, is the likelihood ratio $h(\cdot) = f_\xi(\cdot)/f_\chi(\cdot)$, where f_ξ and f_χ are the density functions of ξ and χ , respectively. As a monotone increasing function of the likelihood ratio, the risk score, $P(D = 1 \mid \cdot)$ where $D = 1$ indicates the real condition of being positive, induces also an optimal transformation. It turns out that binary regression methods can be used to approximate the eROC curve, since f_ξ and f_χ are generally unknown. Several authors have exploited this result: Kauppi (2016) proposed procedures to find the optimal transformation under semiparametric and nonparametric conditions, Chen et al. (2016) presented a semiparametric model by directly modeling the likelihood ratio as an unspecified function of a transformation of the marker, and López-Ratón (2015) made use of logistic GAM regression model to estimate the risk score function, among others.

In practice, the shape of the classification subsets depends on the transformation considered in the regression model. The function `hROC(X, D, formula)` in the package `movieROC` allows any formula to fit a logistic regression model, considering a cubic polynomial fitting by default (`formula = "D ~ pol(X, 3)"`).

2.2 A generalization of the ROC curve: the gROC curve

With a focus on the classification process, a different perspective is proposed to improve the accuracy of the marker, also making use of the distribution of the marker in positive and negative groups, but maintaining control of the classification regions. The idea is to use more flexible decision criteria, defining a family of eligible classification subsets suitable for each particular scenario.

In the scenario where both higher and lower values of the marker are associated with a higher probability of being a positive, the logical family of eligible classification subsets is $\mathcal{I}_g(\cdot)$ [defined in (3)]. For each false-positive rate $t \in [0, 1]$, the classification regions $s_t = (-\infty, x_t^L] \cup (x_t^U, \infty)$ with $x_t^L \leq x_t^U$ reporting the maximum sensitivity fulfill $F_\chi(x_t^U) - F_\chi(x_t^L) = 1 - t$. Therefore, we can write $x_t^L = F_\chi^{-1}(\gamma \cdot t)$ with $\gamma \in [0, 1]$ and consequently $x_t^U = F_\chi^{-1}(1 - (1 - \gamma) \cdot t)$, i.e., γ represents the portion of false-positive rate t which is kept in the subinterval $(-\infty, x_t^L]$ and consequently $1 - \gamma$ represents the portion of false-positive rate t in the subinterval (x_t^U, ∞) . With this notation, we can reformulate the gROC curve definition as

$$\mathcal{R}_g(t) = F_\xi(F_\chi^{-1}(\gamma_t \cdot t)) + 1 - F_\xi(F_\chi^{-1}(1 - (1 - \gamma_t) \cdot t)) \quad t \in [0, 1] \quad (6)$$

where $\gamma_t := \arg \sup_{\gamma \in [0,1]} \{F_\xi(F_\chi^{-1}(\gamma \cdot t)) + 1 - F_\xi(F_\chi^{-1}(1 - (1 - \gamma) \cdot t))\}$.

It should be noted that

- if $\gamma_t = 0 \forall t \in [0, 1]$ (that is, false-positive rate t is totally kept in the subinterval (x^U, ∞)), the gROC curve equals the standard ROC curve, $\mathcal{R}_g(t) = \mathcal{R}_r(t)$;
- if $\gamma_t = 1 \forall t \in [0, 1]$ (that is, false-positive rate t is totally kept in the subinterval $(-\infty, x^L]$), the gROC curve equals the *left-sided ROC curve*, $\mathcal{R}_g(t) = \mathcal{R}_l(t)$, defined as $\mathcal{R}_l(t) := \sup_{s \in \mathcal{I}_l(t)} \mathcal{P}\{\xi \in s\}$ with

$$\mathcal{I}_l(t) = \{s = (-\infty, c] \subset \mathbb{R} \text{ such that } \mathcal{P}\{\chi \in s\} \leq t\}.$$

Related to this, Martínez-Cambolor and Pardo-Fernández (2019) claimed that “the gROC curve is close to the ROC curve (...) when the difference in means dominates over the difference in variances.”

Furthermore, the consideration of the supremum in the definition of the gROC curve is crucial because, in contrast to the standard ROC curve where there is only one subset of the form $s = (c, \infty)$ reporting a particular specificity $1 - t \in [0, 1]$, for the general ROC curve there exist infinite subsets of the form $s = (-\infty, x^L] \cup (x^U, \infty)$ reporting a particular specificity $1 - t \in [0, 1]$. Therefore, among them, the one reporting the maximum sensitivity, i.e., $s_t = \arg \sup_{s \in \mathcal{I}_g(t)} \mathcal{P}\{\xi \in s\}$, is chosen for defining the value of the gROC curve in t , $\mathcal{R}_g(t)$.

Figure 2 displays the construction of the gROC curve for a scenario where considering subsets of the form $(-\infty, x^L] \cup (x^U, \infty)$ is advisable. In this particular scenario, the classification regions are self-contained over the increase in false-positive rate, that is, they fulfill the condition (C):

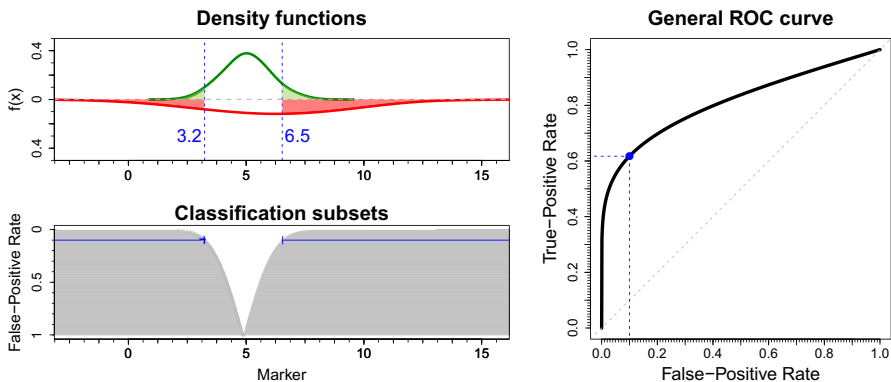


Fig. 2 General ROC curve. Left, density functions of χ (green) and ξ (red) which come from normal distributions $\mathcal{N}(5, 1)$ and $\mathcal{N}(6, 3)$, respectively (top), and the classification subsets (gray) for every false-positive rate t (bottom); right, the gROC curve. Video: <https://doi.org/10.6084/m9.figshare.9891977>

- (C) let $t_1, t_2 \in [0, 1]$ with $t_1 \leq t_2$ and let $s_{t_1} = (-\infty, x_{t_1}^L] \cup (x_{t_1}^U, \infty)$ and $s_{t_2} = (-\infty, x_{t_2}^L] \cup (x_{t_2}^U, \infty)$ be the corresponding eligible subsets such that $\mathcal{R}(t_1) = \mathcal{P}\{\xi \in s_{t_1}\}$ and $\mathcal{R}(t_2) = \mathcal{P}\{\xi \in s_{t_2}\}$, then $s_{t_1} \subseteq s_{t_2}$.

This condition (C) implies that if a subject is classed as a positive for a particular false-positive rate t_1 , it will also be classed as a positive for every $t \geq t_1$.

Imposing it, Martínez-Cambor and Pardo-Fernández (2019) gave the following probabilistic interpretation of the gAUC: *The area under the gROC curve is the probability of selecting randomly and independently two subjects, one positive and one negative, for which there exists a classification subset of the form $(-\infty, x^L] \cup (x^U, \infty)$ which correctly classifies both subjects.*

However, when the condition is not directly fulfilled, imposing it may suppose a loss in the gAUC. In practice, on the basis of a particular dataset, the process of searching the regions which define the gROC curve under the restriction (C) becomes a relevant computational task to address. Finding the optimal solution involves the computation of all possible ways resulting in self-contained subsets built by pairs of values of the marker. However, the computational times are unfeasible and thus a lighter search algorithm to find solutions close to the optimal one is needed. In the next subsection, an algorithm to find those classification regions getting a proficient solution is proposed and a simulation study to examine the impact of imposing the restriction (C) is carried out. The algorithm was also included in the movieROC package, and it makes use of the also implemented method to estimate the gROC curve without restrictions (using the function `gROC(X, D, side = "both")`).

2.2.1 Proposed method to estimate the gROC curve under restriction (C)

The steps of the proposed algorithm are the following:

1. Estimate the gROC curve without restrictions, $\widehat{\mathcal{R}}_g(t)$.
2. Select a false-positive rate (FPR), $t_i \in [0, 1]$, to start from. The point $(t_i, \widehat{\mathcal{R}}_g(t_i))$ will also be a point in the estimated gROC curve under restriction (C), $(t_i, \widehat{\mathcal{R}}_g^{C,t_i}(t_i))$.
3. Take the classification subset reported by $\widehat{\mathcal{R}}_g(t_i)$, $u_{t_i} = (-\infty, x_{t_i}^L] \cup (x_{t_i}^U, \infty)$.
4. Departing from u_{t_i} , the sequence $\{u_{t_{i-1}}, u_{t_{i-2}}, \dots, u_0\}$, where $t_{i-1} = t_i - 1/m$ with m the total number of negative subjects, is built iteratively considering the restriction (C), which comes out into restriction (C.A):

$$0 \leq \gamma_{t_{j-1}} \leq \min \left\{ \gamma_{t_j} \frac{t_j}{t_{j-1}}, 1 - (1 - \gamma_{t_j}) \frac{t_j}{t_{j-1}} \right\} \text{ for every } t_j < t_i.$$

Then, the algorithm used to find the γ_{t_i} reporting the optimal sensitivity for t_i computed in the estimation of $\widehat{\mathcal{R}}_g(t)$ is used over the domain imposed by (C.A). The resulting sequence of regions contained in u_{t_i} will end up in u_0 , a subset $(-\infty, x_0^L] \cup (x_0^U, \infty)$ without any positive subject inside.

5. Departing from u_{t_i} , the sequence $\{u_{t_{i+1}}, u_{t_{i+2}}, \dots, u_1\}$, where $t_{i+1} = t_i + 1/m$, is built iteratively considering the restriction (C) , which comes out into the restriction $(C.B)$:

$$\max \left\{ 1 - (1 - \gamma_{t_j}) \frac{t_j}{t_{j+1}}, \gamma_{t_j} \frac{t_j}{t_{j+1}} \right\} \leq \gamma_{t_{j+1}} \leq 1 \text{ for every } t_j > t_i.$$

Then, the algorithm used to find the γ_{t_i} which reports the optimal sensitivity for t_i computed in the estimation of $\hat{\mathcal{R}}_g(t_i)$ is used over the domain imposed by $(C.B)$. This restriction results in a sequence of classification regions containing u_{t_i} , ending in u_1 , with all the positive subjects inside.

In Step 2, the implemented gROC function gives the option of running the algorithm starting from every FPR $\in [0, 1]$ and keeping that one reporting the maximum empirical AUC, that is *the optimal estimated gROC curve with restriction (C)* (`gROC(X, D, side = "both", restric = TRUE, t0max = TRUE)`). This option is time-consuming since it implies $m + 1$ computations of the algorithm. But if `t0max = FALSE` (default), the initial FPR (`t0/(m + 1)`) can be selected by the user (`gROC(X, D, side = "both", restric = TRUE, t0)`), being the point leading to Youden index (defined as the maximum sum of sensitivity and specificity, Fluss et al. 2005), the option by default if `t0` is not specified.

In order to explore the influence of this restriction and the impact of the initial point in Step 2 on the resulting gROC curves, a simulation study has been carried out (See Appendix for more details). Generally, the restriction (C) does not seem to have a big impact on the final gAUC except for some pathological scenarios. The suggested procedure is: first, to estimate the gAUC without restrictions; second, to estimate the area under the curve with restriction (C) departing from default initial point (Youden index) in Step 2, $\hat{\mathcal{R}}_{g,C,1}$; finally, if the difference between those estimates is high, compute the optimal estimated gROC curve with restriction (C) , $\hat{\mathcal{R}}_g$; otherwise, keep the estimation $\hat{\mathcal{R}}_{g,C,1}$.

3 ROC curve for multivariate markers

When more than one marker is available, combining them may incur a substantial improvement in the classification accuracy. The main difference with the representation for univariate markers (Sect. 2) is that, in the multivariate case, it is not possible to keep all the classification regions (for all specificities) in one simple 2D plot. When the dimension is two, $p = 2$, we propose to display the change of those regions s_t along t by means of dynamic graphical representations in order to add a new time dimension which keeps the change in the false-positive rate, t , from 0 to 1 (Fig. 3). When the dimension is higher, $p > 2$, the solution is not direct since it is no longer feasible to display the subsets on the original space using a 2D graphic. Two solutions are proposed: to project the regions on the plane constituted by a) two variables selected by the user among the p available, or b) the first two components resulting from a dimensionality reduction technique. This section is divided in two:

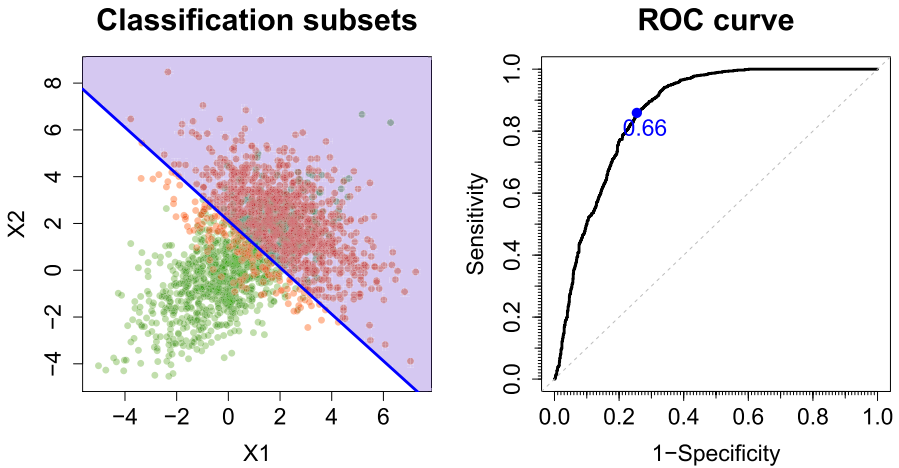


Fig. 3 Bivariate ROC curve. Left, the classification subset (blue) for a particular false-positive rate t ; right, the ROC curve resulting from varying the intercept of the linear classifier function. Data have been simulated from bivariate normal distributions in positive (red) and negative (green) populations. Video: <https://doi.org/10.6084/m9.figshare.9891989>

the first part devoted to bivariate markers and the second one to markers in higher dimension.

The standard ROC curve is only defined for univariate markers; thus, for studying the classification accuracy of a multivariate marker, the available literature suggests reducing it to a univariate one by a transformation, $h : \mathbb{R}^p \rightarrow \mathbb{R}$. As it was in the definition of the efficient ROC curve, the choice of $h(\cdot)$ is not direct and it may incur an overfitting, besides the resulting classification subsets for the original marker may be useless. In this process of reduction to dimension one, the classification problem may often be blurred and, in most references, not even one single region s_t is visualized. We study the graphical representations of the outcomes of that classification process and include them in the presented tool.

The simple linear combination of the different components is frequently considered as $h(\cdot)$, i.e., $h(X) = \mathcal{L}_\beta(X) = \sum_{i=1}^p \beta_i X_i$ where $X = (X_1, \dots, X_p)^\top$ denotes the multivariate marker. These transformations are easy to interpret in terms of classification subsets: For each fixed specificity, there is a hyperplane in the original space separating the two classes (Fig. 3).

Once the type of functional transformation $h(\cdot) = h_\beta(\cdot)$ is chosen, the AUC of the transformed marker $h(X)$ is generally the objective function to maximize for getting the parameters β of the combination which reports the “best classification ability”. The meaning of the latter concept depends on that objective function; for example, if it is the standard AUC, the “best resulting marker” is the one reporting the maximum probability $\mathcal{P}\{h(\xi) > h(\chi)\}$. It should be noted that, optimizing the AUC, the whole ROC curve is relevant and not only one particular specificity, highlighting the importance of visualizing all the classification subsets which generate the ROC curve. There are some references considering different objective functions related to ROC curves: McIntosh and Pepe (2002) aimed to estimate the linear combination

which maximizes the partial AUC (the area under the curve over an interval of specificities); Yin and Tian (2014) and Xu et al. (2015) proposed linear and nonlinear transformations of the marker, respectively, to optimize the Youden index; while Meisner et al. (2017) deals with the search of the linear combination which maximizes the sensitivity for every fixed specificity (which is denoted subsequently by *linear combinations with dynamic parameters*).

3.1 Bivariate markers

Most references cover linear combinations of the original markers, $\mathcal{L}_\beta(X) = \beta_1 X_1 + \beta_2 X_2$, whose resulting ROC curve optimizes an objective function (AUC, partial AUC, Youden index, etc.). Once the parameter β is estimated, classification subsets of the form (c, ∞) over $h(X) = \mathcal{L}_\beta(X)$ are considered. The regions in the original space (two-dimensional) are limited by a line with the same slope, $-\beta_1/\beta_2$, but changing intercept c_i/β_2 depending on t . Then, the lines separating the two classes are parallel over the change in t (Fig. 3). These combinations will be called *linear combinations with fixed parameters*.

The ROC curve coming from the linear combination $\mathcal{L}_\beta(X) = \beta_1 X_1 + \beta_2 X_2$ with fixed parameter β can be defined as (2) with the following family of eligible classification subsets:

$$\mathcal{I}_{\mathcal{L}_\beta}(t) = \{s \subset \mathbb{R}^2 \text{ such that } \beta^\top y \leq c \ \forall y \in s \text{ and } \mathcal{P}\{\chi \in s\} \leq t\}. \tag{7}$$

In this paper, we also cover a different perspective: when the objective function is the true-positive rate for each false-positive rate separately. The goal is to find, for each $t \in [0, 1]$, the combination $\mathcal{L}_{\beta(t)}(X) = \beta_1(t)X_1 + \beta_2(t)X_2$ and classification subset $s_t = (c_t, \infty)$ such that $\mathcal{P}\{\mathcal{L}_{\beta(t)}(\chi) \in s_t\} \leq t$ reporting the highest sensitivity $\mathcal{P}\{\mathcal{L}_{\beta(t)}(\xi) \in s_t\}$. In this case, the parameter $\beta(t)$ depends on t ; hence, the parallel property of the borders of the regions over t may not be fulfilled since the slope is $-\beta_1(t)/\beta_2(t)$ (video linked in Sect. 3.1.2). These combinations will be called *linear combinations with dynamic parameters*.

The ROC curve resulting from the transformation $\mathcal{L}_{\beta(t)}(X) = \beta_1(t)X_1 + \beta_2(t)X_2$ with parameter $\beta(t)$ can be defined as (2) with the family of eligible classification subsets:

$$\mathcal{I}_{\mathcal{L}_{\beta(t)}}(t) = \{s \subset \mathbb{R}^2 \text{ such that } \beta(t)^\top y \leq c \ \forall y \in s \text{ and } \mathcal{P}\{\chi \in s\} \leq t\}. \tag{8}$$

3.1.1 Linear combinations with fixed parameters

Recently, Kang et al. (2016) carried out a complete review of algorithms which seek linear combinations maximizing the AUC. They provided the R code used to estimate the “best parameters” using the different methods revised. We have used it in order to find those parameters β_1 and β_2 and include them as an input parameter in

the function `biROC(X, D, method = "fixedLinear", coefLinear = c(β_1, β_2))` of our R package.

Particularly, the reviewed techniques are the following: Su and Liu (1993) found the linear combination which maximizes the AUC under conditionally multivariate normality; Pepe and Thompson (2000) developed a procedure for bivariate markers where a grid is used to numerically find $\hat{\beta}$ and proposed a distribution-free estimator using the Mann–Whitney U statistics to the AUC estimation (Hanley and McNeil 1982).

In addition to this, logistic regression can be used to find the parameter β maximizing the risk score function fitting the following model:

$$\text{logit}(\mathcal{P}\{D = 1|X = (X_1, X_2)\}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \quad (9)$$

Pepe et al. (2006) proved that, if the model (9) holds, the ROC curve of the linear combination resulting from this method dominates any other ROC curve coming from any other linear or nonlinear transformation of the marker.

In Figure S2, the classification subsets s_t for t reporting the Youden index and the corresponding ROC curves for the three procedures above over two simulated examples are displayed. It can be seen that even if a perfect linear separation is possible (top), since it is far from being multivariate normal, the parametric approach (Su and Liu 1993) results in an AUC of 0.894 instead of 1, the optimal achievable. In this linearly separable case, the outputs from the regression model and Pepe and Thompson (2000) method coincide. However, if a perfect linear separation between groups is not possible, the results may differ, being the estimated AUC from the second method always greater. An example of this scenario is shown in Figure S2 (bottom). The AUC resulting from fitting a logistic linear regression model is 0.716, with an improvement of 9.4% using the Pepe and Thompson (2000) approach. It should be noted the difference in the classification regions resulting from each method.

As a final remark, there is another procedure gathered in Kang et al. (2016): Liu et al. (2011) considered the min–max combination maximizing the Mann–Whitney U estimator of the AUC, i.e., $\mathcal{L}_\alpha(X) := X_{\max} + \alpha X_{\min}$ where $X_{\max} := \max_{k=1,2} X_k$ and $X_{\min} := \min_{k=1,2} X_k$. This is based on the fact that, for every threshold, X_{\max} is the univariate marker with the largest sensitivity and the smallest specificity, while X_{\min} has the smallest sensitivity and the largest specificity. It is a nonparametric approach and thus more robust than Su and Liu (1993) method, but every marker involved is required to be either expressed in the same units or standardized to be unit-less. Besides the regions induced by $\mathcal{L}_\alpha(X) := X_{\max} + \alpha X_{\min}$ are determined by two lines: symmetric with respect to $X_1 - X_2 = 0$ (Figure S3 in Supplementary Material).

3.1.2 Linear combinations with dynamic parameters

There also exist some procedures seeking linear combinations of the markers which may not vary in a parallel way along the construction of the ROC curve since the parameter β depends on t . The two following methods aim to maximize the sensitivity for a particular specificity, that is, for a particular $t \in (0, 1)$, the

objective is to find the coefficient $\beta(t)$ and threshold $c(t)$ for the linear combination $\mathcal{L}_{\beta(t)}(X) = \beta_1(t)X_1 + \beta_2(t)X_2$ such that

$$\begin{aligned}
 (\beta(t), c(t)) &\in \arg \max_{(\beta, c) \in \Omega_t} \mathcal{P}\{\beta_1 \xi_1 + \beta_2 \xi_2 > c\} \quad \text{where} \\
 \Omega_t &:= \left\{ (\beta, c) \in \mathbb{R}^3 \text{ such that } \|\beta\| = 1 \text{ and } \mathcal{P}\{\beta_1 \chi_1 + \beta_2 \chi_2 > c\} \leq t \right\}.
 \end{aligned}
 \tag{10}$$

- Meisner et al. (2017) suggested to estimate the probabilities involved in (10) by means of the empirical distribution functions of $\mathcal{L}_{\beta(t)}(X)$ in both groups for different possible $\beta(t)$ over a grid. But, as the authors highlighted, this estimator involves the indicator function, which is non-smooth with respect to the parameters $(\beta(t), c(t))$, so they proposed to use a smooth approximation of it. Good theoretical asymptotic properties of this estimator under certain weak conditions were derived, and the algorithm was implemented in their R package called `maxTPR`. That implementation is internally used in our package by the function `biROC(X, D, method="dynamicMeisner," alpha, approxh, multiplier)`, where the last input parameters are those for their function `maxTPR` regarding convergence and smoothness of the algorithm.
- We propose to use a grid search method similar to the one proposed in Pepe and Thompson (2000) to estimate $\beta(t)$ and the empirical distribution functions to estimate the probabilities in (10), despite their lack of smoothness. The steps of the proposed algorithm are detailed below:

1. For a fixed false-positive rate, t , consider K equally spaced values $\alpha_l \in [0, 1]$ and calculate $\gamma_l = 1/\alpha_l$. For each element of the grid $G = \{g_1, \dots, g_{4K}\} := \{(1, \alpha_1), \dots, (1, \alpha_K), (1, \gamma_1), \dots, (1, \gamma_K), (-1, \alpha_1), \dots, (-1, \alpha_K), (-1, \gamma_1), \dots, (-1, \gamma_K)\}$, calculate the minimum c_k such that

$$\widehat{\text{FPR}}(g_k, c_k) := \frac{1}{m} \sum_{j=1}^m I(g_{1,k} X_{1,j}^{D=0} + g_{2,k} X_{2,j}^{D=0} > c_k) \leq t$$

where m is the negative sample size and $X_{k,j}^{D=0}$ with $k = 1, 2$ and $j = 1, \dots, m$ is the k th component of the j th negative subject.

2. For each $k \in \{1, \dots, 4K\}$ and c_k in Step 1, calculate the true-positive rate

$$\widehat{\text{TPR}}(g_k, c_k) := \frac{1}{n} \sum_{i=1}^n I(g_{1,k} X_{1,i}^{D=1} + g_{2,k} X_{2,i}^{D=1} > c_k)$$

where n is the positive sample size and $X_{k,i}^{D=1}$ with $k = 1, 2$ and $i = 1, \dots, n$ is the k th component of the i th positive subject.

3. Search the $k \in \{1, \dots, 4K\}$ which reports the maximum $\widehat{\text{TPR}}(g_k, c_k)$. Hence, the estimated parameters are

$$(\hat{\beta}(t), \hat{c}(t)) = (\hat{g}_k, \hat{c}_k) = \arg \max_{k \in \{1, \dots, 4K\}} \widehat{\text{TPR}}(g_k, c_k). \quad (11)$$

The implementation of this algorithm was included in the `movieROC` package by the function `biROC(X, D, method = "dynamicEmpirical", K)`, where `K` is the grid parameter K in Step 1 (201 by default).

In order to estimate the whole ROC curve, for each t on a grid of the unit interval, the parameters $(\beta(t), c(t))$ are estimated using one of the two algorithms above. By definition, the last algorithm results in a monotone ROC curve empirical estimate, but the opposite may happen using the estimates from Meisner et al. (2017) approach.

The construction of the ROC curves from these two procedures for a simulated contaminated bivariate normal scenario is shown in <https://doi.org/10.6084/m9.figshare.9892061>. When the parameters are dynamic, it is specially interesting to visualize how the regions change with the specificity.

3.1.3 Optimal combinations $\mathcal{F}_{\beta(X)}$

It should be noted that the linear combination may not be the optimal transformation in terms of achieving the highest sensitivity uniformly at any specificity. In fact, that is the likelihood ratio, $f_{\xi}(X)/f_{\chi}(X)$, or any monotone transformation, as it was pointed out in Sect. 2.1. In case of a conditionally bivariate normal marker, it is

$$\mathcal{F}_{\beta(X)} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2$$

for some β_1, \dots, β_5 . Under equal covariance matrices, $\beta_3 = \beta_4 = \beta_5 = 0$.

Furthermore, if the model

$$\text{logit}(\mathcal{P}\{D = 1|X = (X_1, X_2)\}) = g(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

holds for some monotone increasing function $g(\cdot)$, then the optimal combination of the marker is a linear combination. But even in this case, if $g(\cdot)$ is unknown, the logistic regression model in (9) may not lead to the referred optimal parameters.

However, the conditionally distribution of the marker is often unknown in practice; hence, finding a proper family of combinations (linear, quadratic, etc.) becomes a difficult task to address. With the purpose of covering different families of combinations and consequently several shapes of classification subsets, the option `biROC(X, D, method = "lrm", formula, stepModel)` is included in the presented package. It computes a logistic regression model considering the combination suggested by `formula`, which is a quadratic function of X by default (`formula = "D ~ poly(X.1, 2) + poly(X.2, 2) + I(X.1 * X.2)"` where $X.i$ is the i th component of the marker). A model based on the Akaike information criterion (AIC) by means of a stepwise algorithm is selected by default; otherwise, the crude model is taken (`stepModel = FALSE`).

Figure S4 (Supplementary Material) collects the classification subsets obtained from the function `biROC` for different choices of families of combinations. Particularly `formula = "D ~ X.1 + X.2"` (linear) and `formula` by default (quadratic), reporting AUCs of 0.705 and 0.854, respectively.

3.2 Multivariate markers

To cope with multivariate markers, the techniques included in the previous subsection to handle bivariate markers are extended to higher dimension. While the approaches proposed by Su and Liu (1993), Liu et al. (2011) and logistic regression model can be directly accommodated to $p > 2$, the extension of the Pepe and Thompson (2000) algorithm, despite being direct, is computationally demanding. To deal with it, they proposed the following stepwise algorithm:

1. Find two markers, X_1^* and X_2^* , whose optimal linear combination, $S^{(1)}(\beta_1, \beta_2) = \beta_1 X_1^* + \beta_2 X_2^*$, reports the maximum AUC among all pairs of markers.
2. Among the rest of variables, find the marker X_3^* whose optimal linear combination with $S^{(1)}(\beta)$ reports the maximum AUC.
3. Repeat Step 2 until all the markers are included, i.e., $S^{(p-1)}(\beta) = \beta_1 X_1^* + \dots + \beta_p X_p^*$.

However, this algorithm implies the computation of Pepe and Thompson (2000) method $1/2 \cdot \left[p(p-1) + \sum_{k=2}^{p-1} (p-k)(p-k-1) \right]$ times, which becomes very large when the dimension increases (129 computations for $p = 10$). Two different solutions for $p > 2$ which are much less computationally intensive ($p-1$ computations of the method are needed) were proposed by Kang et al. (2016) (*step-down* and *step-up* techniques) and Yan et al. (2015).

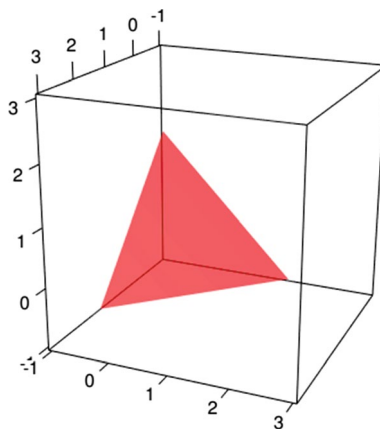
The `multiROC` function in our package allows the user to input the estimated fixed parameters of a linear combination for visualizing the results. Particularly, `multiROC(X, D, method = "fixedLinear", coefLinear = c(\beta_1, \dots, \beta_p))`.

Regarding linear combinations with dynamic parameters, Meisner et al. (2017) technique was designed for multivariate markers. However, as the authors mention, the risk of overfitting is expected to grow as the dimension increases. This method is included in `movieROC` making use of `maxTPR` package: `multiROC(X, D, method = "dynamicMeisner", alpha, approxh, multiplier)`.

Nevertheless, if there does not exist a monotone function $g(\cdot)$ such that $f_{\xi}(X)/f_{\chi}(X) = g(\mathcal{L}_{\beta}(X))$, the optimal combination of the markers is not linear. For covering nonlinear combinations, binary logistic regression procedures have been included in our package: `multiROC(X, D, method = "lrn", formula, stepModel)`.

When $p > 2$, it is not possible to display the regions on the original space, so two options are implemented in the `movieROC` package:

- (a) To show the regions on the plane of the two principal components resulting from a principal component analysis (PCA), using the function `plot.buildROC(obj, display.method = "PCA")` where `obj` is the output of the `multiROC` function;
- (b) To choose two markers among the p involved and project the classification subsets on that plane, by means of the function `plot.buildROC(obj, display.method = "OV", displayOV = c(.,.))` where the vector `displayOV` indicates the two univariate markers to be considered.



An important remark should be mentioned at this point: The colored areas indicating the classification regions below or above the corresponding border (Fig. 3) are not used any more since the meaning is lost with the projection. A clear example is the figure at right, where the red plane defines the border of the classification region in a linearly separable 3D scenario. The solution proposed is to project the border and indicate the classification rule in the original space highlighting the points which are classed as positive for every particular specificity (video in Fig. 8).

4 Real-world data example

The real-world dataset we have taken into consideration to illustrate the use of the proposed tool and also the differences among the approaches discussed in this manuscript is the *Banknote Authentication dataset*, which can be found in the UCI Machine Learning Repository. It contains four features of the wavelet transformed images of a total of 1372 pictures taken from two types of banknotes: genuine and

forged. The goal is to study the performance of those features to detect which ones are counterfeit (44.46% of the sample). The markers are: *variance*, *skewness* and *kurtosis* of wavelet transformed image and *entropy* of image.

First, the performance of *kurtosis* as a single marker is measured. The standard AUC is 0.536 (Fig. 4), suggesting that *kurtosis* is not a good marker to discriminate between genuine and forged banknotes considering the usual classification subsets (c, ∞) .

Considering the transformation $h(\cdot) = \hat{f}_\xi(\cdot)/\hat{f}_\chi(\cdot)$ where $\hat{f}_\xi(\cdot)$ and $\hat{f}_\chi(\cdot)$ are the kernel density estimates for kurtosis of wavelet transformed image in forged and genuine banknotes, respectively, with smoothing bandwidth 1.5 and Gaussian kernel, the resulting area under the curve (eAUC) increases to 0.678. For a fixed specificity of 0.9, the rule is to class as counterfeit those banknotes whose *kurtosis* value is in the region $(-\infty, -4.7) \cup (-1.33, -1.29) \cup (5.57, \infty)$ (bottom figure in Fig. 5).

On the other hand, considering more general classification regions over the original marker, particularly of the form $(-\infty, x^U] \cup (x^L, \infty)$, the resulting gAUC differs only 0.011 from eAUC. Imposing the restriction (C) to those subsets, the gAUC is reduced to 0.647. The regions to classify a banknote as counterfeit regarding its *kurtosis* with a specificity of 0.9 are $(-\infty, -4.38] \cup (5.59, \infty)$ and $(-\infty, -3.41] \cup (8.83, \infty)$ without and with restriction, respectively (Fig. 5).

When the *skewness* is also taken (multiplied by -1 to assume that higher values of the marker are related to a higher likelihood of being a forged banknote), the optimal linear combination $\mathcal{L}_\beta(\cdot) = \beta_1 \cdot skewness' + \beta_2 \cdot kurtosis'$ (where $'$ means standardized) resulting from each method in Sect. 3.1.1 has been estimated. The ROC curves and classification regions for specificity 0.9 are displayed in Fig. 6. The AUCs are around 0.815 except for Liu et al. (2011) method, whose AUC (0.692) is even lower than the result considering only *skewness* (0.749).

Regarding linear combinations with dynamic parameter $\beta(t)$, the ROC curves estimated by the two procedures in Sect. 3.1.2 are shown in Fig. 7, reporting AUCs 0.835 and 0.858 for Meisner et al. (2017) and proposed empirical algorithm, respectively. In the video linked in Fig. 7, the construction of the first ROC curve, not monotone over t , is illustrated. Besides, considering nonlinear combinations by means of a quadratic logistic regression, the AUC is slightly higher (0.860).

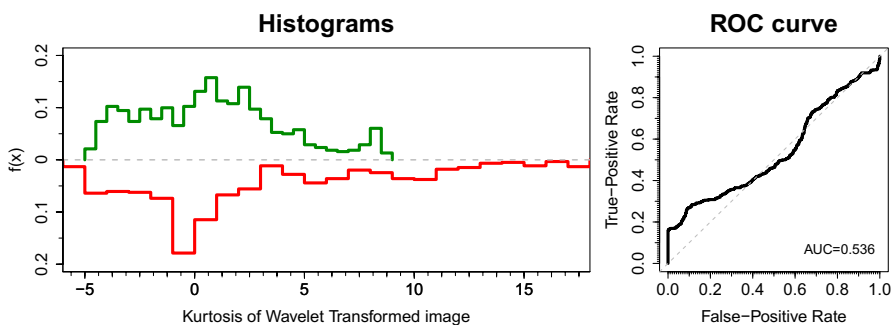


Fig. 4 Standard ROC curve. Left, histograms of kurtosis of wavelet transformed image in genuine (green) and forged (red) banknotes. Right, the ROC curve

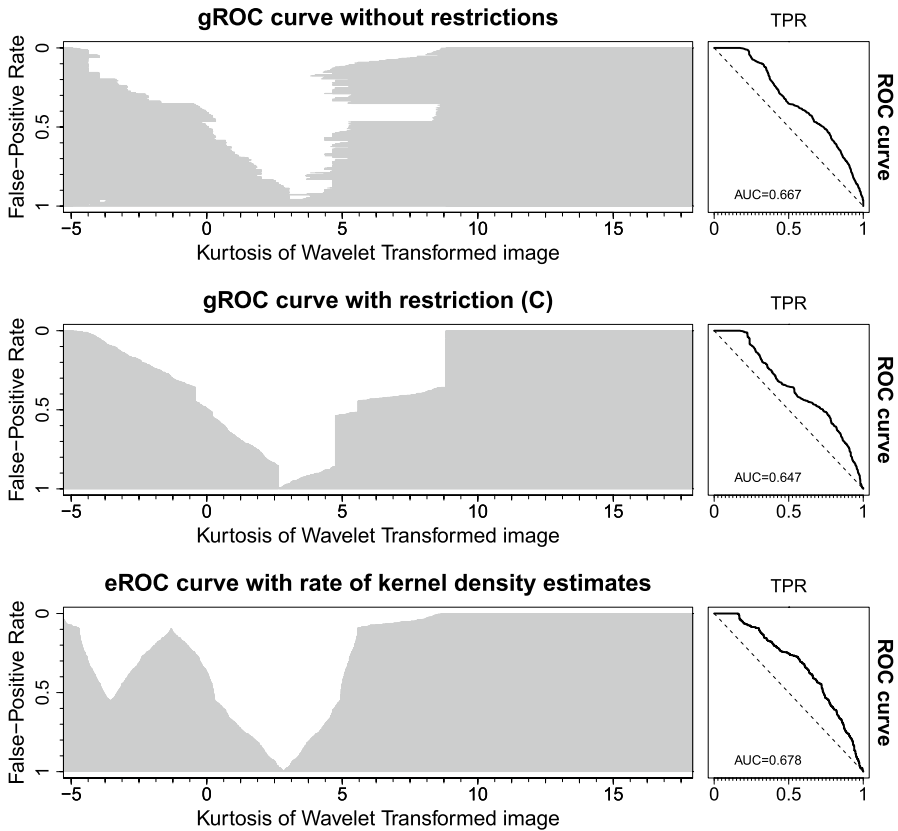


Fig. 5 General and efficient ROC curves. Classification subsets of *kurtosis* of wavelet transformed image and resulting ROC curves. From top to bottom: gROC curve without restrictions, gROC curve with restriction (*C*) and eROC curve

If the *entropy* of the image is also considered (multiplied by -1), the parameters of the optimal combination $\mathcal{L}_{\beta}(\cdot) = \beta_1 \cdot \textit{skewness}' + \beta_2 \cdot \textit{kurtosis}' + \beta_3 \cdot \textit{entropy}'$ were estimated using the four techniques in Sect. 3.2: Su and Liu (1993), Pepe and Thompson (2000) with the extension proposed by the authors (here $p = 3$), linear logistic regression and Liu et al. (2011) method. The AUCs are approximately 0.880 except for Liu et al. (2011) method, with an AUC of 0.711 (Fig. 8). In the video linked in Fig. 8, the construction of the ROC curve from Pepe and Thompson (2000) procedure is illustrated: The projection plane is constituted by the variables *kurtosis* and *entropy* (standardized). The decision rule for each t is displayed highlighting the subjects (points) classed as positive.

Finally, Meisner et al. (2017) approach was considered to estimate the linear combination $\mathcal{L}_{\beta(t)}(\cdot)$ with dynamic parameters reporting the maximum sensitivity for each specificity. The AUC is slightly higher (0.897), but it can be seen that the

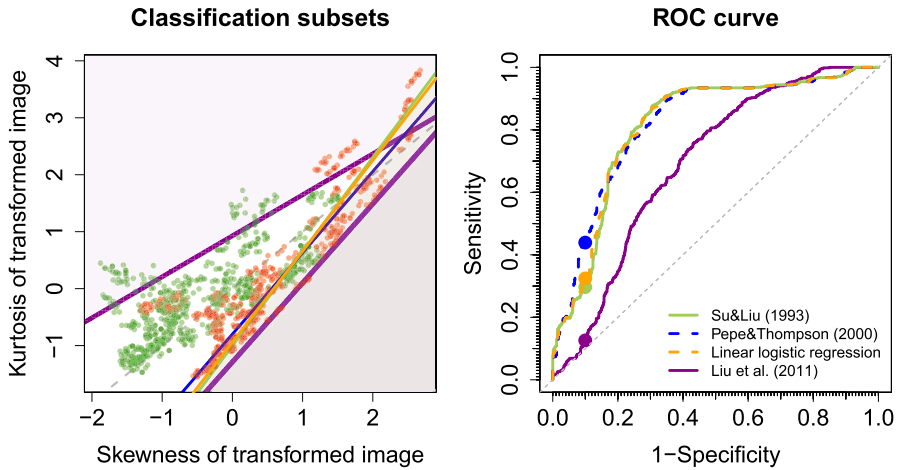


Fig. 6 Bivariate ROC curves. Classification subsets of the standardized bivariate marker (*skewness, kurtosis*) for FPR = 0.1 and resulting ROC curves using different methods to find a linear combination with fixed parameters

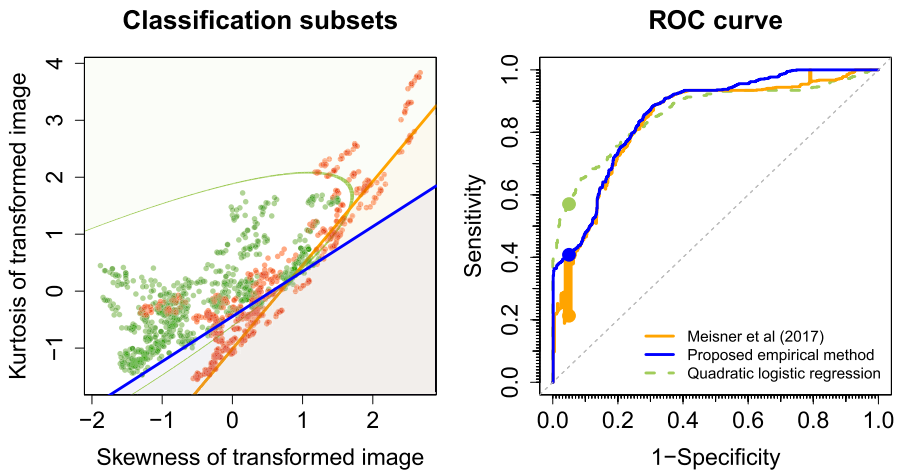


Fig. 7 Bivariate ROC curves. Classification subsets of the standardized bivariate marker (*skewness, kurtosis*) for FPR = 0.05 and resulting ROC curves using different methods to find a linear combination with dynamic parameters (solid lines) and quadratic combination from a logistic regression model (dashed line). Video: <https://doi.org/10.6084/m9.figshare.9891833>

ROC curve estimate is not monotone. Considering a quadratic logistic regression over the three considered features, the final AUC is 0.907 (Fig. 9).

To sum up, Table 1 shows the results in terms of the estimated AUCs and the improvement in each method and each variable included in this dataset. It should be noted that, even if the AUC of *entropy* as a univariate marker is very small

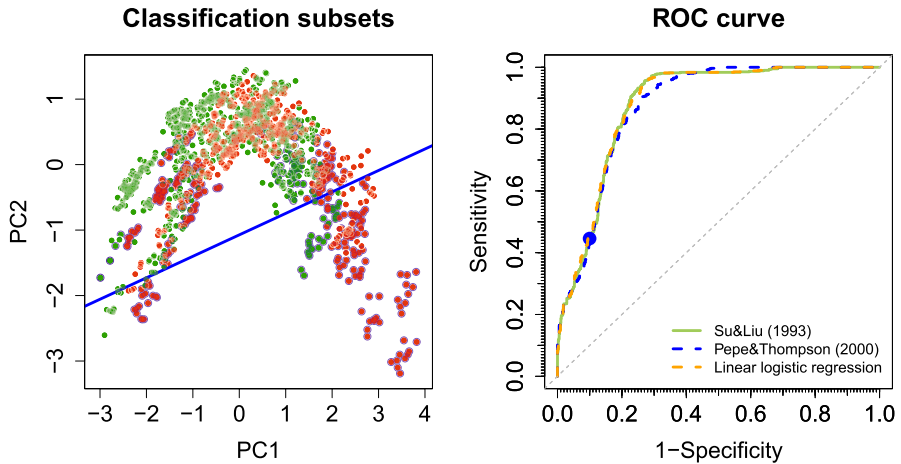


Fig. 8 Multivariate ROC curves. Left, projection of the marker (*skewness*, *kurtosis*, *entropy*) over the two principal components from PCA; the subjects classed as positive for FPR = 0.1 using Pepe and Thompson (2000) method are highlighted in blue. Right, the ROC curves for different methods to find a linear combination with fixed parameters. Video: <https://doi.org/10.6084/m9.figshare.9891935>

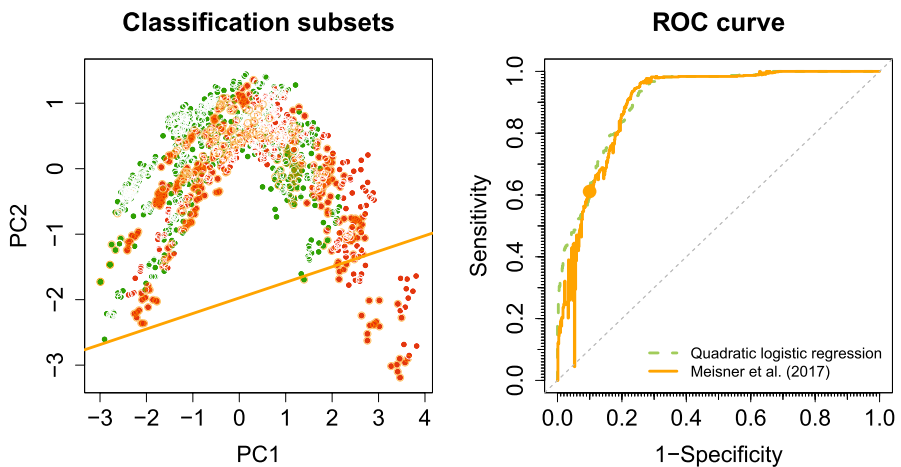


Fig. 9 Multivariate ROC curves. Left, projection of the standardized marker (*skewness*, *kurtosis*, *entropy*) over the two principal components from a PCA; subjects classed as positive for FPR = 0.1 by Meisner et al. (2017) method are highlighted in orange. Right, the ROC curves from Meisner et al. (2017) procedure (orange) and fitting a quadratic logistic regression model (green)

(0.519), when it is combined with the other two features (*skewness* and *kurtosis*), the resulting AUCs increase. The quadratic transformation is more general, and it reports the highest AUCs, although the procedures to obtain linear combinations with dynamic parameters are comparable to it.

Table 1 Area under the ROC curve estimate for different methods and variables

		<i>skewness</i>	<i>kurtosis</i>	<i>entropy</i>
Univariate	Standard ROC curve	0.749	0.536	0.519
Multivariate (linear combination fixed parameters)	Su and Liu (1993)	0.814		0.880
	Pepe and Thompson (2000)	0.819		0.875
	Linear logistic regression	0.816		0.882
Multivariate (linear combination dynamic parameters)	Liu et al. (2011)	0.692		0.711
	Meisner et al. (2017)	0.835		0.897
Multivariate (quadratic combination)	Empirical dynamic	0.858		–
	Quadratic logistic regression	0.860		0.907

The last column of multivariate techniques refers to the consideration of the three features (*skewness*, *kurtosis*, *entropy*), while the second column refers to the bivariate marker (*skewness*, *kurtosis*)

5 Discussion

In this article, we have proposed alternatives to visualize the classification regions leading to the construction of the ROC curve, on the basis of either a univariate or a multivariate marker. The solutions include both static and dynamic graphical representations which help the practitioner to understand the classification problem under study. In addition, a graphical tool including the different options of visualization has been implemented in the user-friendly R package `movieROC`.

In the literature, there exist different approaches for increasing the classification accuracy of a marker. Particularly, dealing with univariate markers, two options may be considered: to make the underlying classification regions more flexible (one of those generalizations defines the so-called *gROC curve*) or to transform the marker by means of a functional transformation (resulting in the *eROC curve* if the transformation considered is the one maximizing the sensitivity for each specificity). In both cases, it becomes very relevant to become aware of the classification subsets which derive every pair of specificity and sensitivity. An algorithm to estimate the *gROC curve* under a particular restriction over the regions has been proposed in this paper.

On the other hand, handling multivariate markers becomes a challenge, not only in terms of visualization but also for definition of the ROC curve in this context. A theoretical framework is provided in the present paper for generalizing the definition, keeping the interpretation of the underlying classification problem. The methods collected are mainly focused on finding hyperplanes in the original space \mathbb{R}^p which optimize the AUC (*linear combinations with fixed parameters*) or the sensitivity over each specificity in the unit interval (*linear combinations with dynamic parameters*), but also different shapes of classification regions can be examined by fitting logistic regression models.

It should be noted that the ROC curve definition in (2) extends the working area in two senses: First, it does not only accept univariate markers but also multivariate markers (Fisher 1936; Zhang 1998) or even other kind of data also common

in practice such as functional data (Hall et al. 2001; Biau et al. 2005) or compositional data (Aitchison and Egozcue 2005); and second, that generalization allows us to accommodate flexible shapes of regions by means of the novel concept of *family of eligible classification subsets*, $\mathcal{I}(t)$. For instance, extending the gROC curve approach for bivariate markers, $\mathcal{I}(t)$ can be defined as the set of regions in the original space which lie between two parallel lines. Mathematically,

$$\mathcal{I}_{g_2}(t) = \{s \subset \mathbb{R}^2 \text{ such that } \beta^T y \leq x^L \cup \beta^T y \geq x^U \quad \forall y \in s \text{ and } \mathcal{P}\{\chi \in s\} \leq t\}.$$

Making use of the framework discussed and particularly focusing on the extension of the definition of the ROC curve in (2), some other considerations of families of classification subsets and their visualization may be further studied.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

Appendix

Theoretical result about the existence of a transformation $h(\cdot)$ of the marker which reports a standard ROC curve equivalent to the gROC curve for the original marker

Proposition 1 *Not for any set of values of (x_t^L, x_t^U) with $t \in [0, 1]$ giving rise to the gROC curve one can find a transformation $h(\cdot)$ of the marker such that the classification regions $s_t = (x_t^L, x_t^U]$ (or $s_t = (-\infty, x_t^L] \cup (x_t^U, \infty)$) without loss of generality can be expressed as $\mathcal{C}_t = \{x \in \mathbb{R} \text{ such that } h(x) \geq x_t^*\}$ for some x_t^* for all $t \in [0, 1]$. In other words, in some scenarios there is no transformation $h(\cdot)$ of the marker such that the resulting standard ROC curve is the same as the gROC curve for the original marker.*

Proof

- (1) Suppose that there exists a function $h : \mathcal{D}(h) \subseteq \mathbb{R} \rightarrow \mathcal{R}(h) \subseteq \mathbb{R}$ (where $\mathcal{D}(h)$ and $\mathcal{R}(h)$ denote the domain and codomain of the function $h(\cdot)$, respectively) such that, for every false-positive rate $t \in [0, 1]$, there exists a $x_t^* \in \mathbb{R}$ such that the classification subset \mathcal{C}_t defined as

$$\mathcal{C}_t = \{x \in \mathcal{D}(h) \text{ such that } h(x) \geq x_t^*\}$$

is equivalent to the classification subset $s_t = (x_t^L, x_t^U]$ resulting from the classification process used in the definition of the gROC curve.

- (2) By definition, $t = \mathcal{P}\{\chi \in \mathcal{C}_t\}$ and hence $x_t^* \in \mathbb{R}$ is such that $\mathcal{P}\{h(\chi) \geq x_t^*\} = t$. Therefore, given two false-positive rates $t_1, t_2 \in [0, 1]$ such that $t_1 > t_2$, then $\mathcal{P}\{h(\chi) \geq x_{t_1}^*\} > \mathcal{P}\{h(\chi) \geq x_{t_2}^*\}$ and thus $x_{t_1}^* < x_{t_2}^*$.

(3) But for any function $h : \mathcal{D}(h) \subseteq \mathbb{R} \rightarrow \mathcal{R}(h) \subseteq \mathbb{R}$, we know that the following subsets content relationship is fulfilled:

$$\{x \in \mathcal{D}(h) \text{ such that } h(x) \geq a_2\} \subseteq \{x \in \mathcal{D}(h) \text{ such that } h(x) \geq a_1\}$$

for every pair $a_1, a_2 \in \mathcal{R}(h)$ such that $a_1 < a_2$. This is due to the fact that, every $x \in \{x \in \mathcal{D}(h) \text{ such that } h(x) \geq a_2\}$, by definition fulfills that $h(x) \geq a_2$ so, particularly, $h(x) \geq a_1$ for every $a_1 < a_2$, i.e., $x \in \{x \in \mathcal{D}(h) \text{ such that } h(x) \geq a_1\}$.

Joining the results above, we have that given two false-positive rates $t_1, t_2 \in [0, 1]$ such that $t_1 > t_2$, then $x_{t_1}^* < x_{t_2}^*$ [by (2)] and thus $\mathcal{C}_{t_2} \subseteq \mathcal{C}_{t_1}$ [by (3)]. But $\mathcal{C}_t = s_t$ for every $t \in [0, 1]$ by (1), therefore $s_{t_2} \subseteq s_{t_1}$.

In summary, if there exists a transformation $h(\cdot)$ of the marker such that the resulting ROC curve coincides with the gROC curve for the initial marker, then the corresponding classification regions, denoted by $s_t \subset \mathbb{R}$, fulfill that

$$s_{t_2} \subseteq s_{t_1} \text{ for every pair of false-positive rates } t_1, t_2 \in [0, 1] \text{ such that } t_1 > t_2.$$

In the scenario where the marker in the negative population is normally distributed as $\mathcal{N}(0, 0.75)$ and in the positive population follows the mixture of normal distributions $\Delta \times \mathcal{N}(-0.5, 0.25) + \Delta \times \mathcal{N}(0.75, 0.25)$ where Δ is a Bernoulli random variable with success probability $\pi = 0.5$, the classification subsets underlying the gROC curve are those reported in Fig. 10.

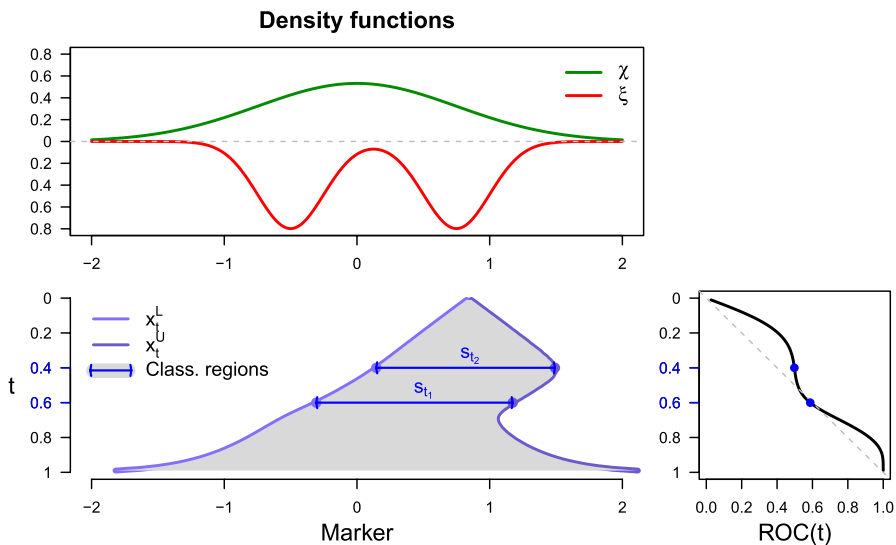


Fig. 10 Top, the density functions for the scenario $\chi \sim \mathcal{N}(0, 0.75)$ and $\xi \sim \Delta \times \mathcal{N}(-0.5, 0.25) + \Delta \times \mathcal{N}(0.75, 0.25)$ where Δ is a Bernoulli random variable with success probability $\pi = 0.5$. Bottom, the classification subsets (left) underlying the gROC curve (right). The classification subsets are of the form $(x^L, x^U] \subset \mathbb{R}$ and are colored in gray for every false-positive rate $t \in [0, 1]$. In particular, those corresponding to $t_1 = 0.6$ and $t_2 = 0.4$ are highlighted in blue

The classification subsets are of the form $(x^L, x^U] \subset \mathbb{R}$ and, by definition, are those reporting the maximum sensitivity, $Se(x^L, x^U)$ for each fixed specificity $Sp(x^L, x^U) = 1 - t$, where

$$Sp(x^L, x^U) = 1 - \mathcal{P}\{x^L < \chi \leq x^U\} = 1 - \Phi\left(\frac{x^U}{0.75}\right) + \Phi\left(\frac{x^L}{0.75}\right) \tag{12}$$

$$\begin{aligned} Se(x^L, x^U) = \mathcal{P}\{x^L < \xi \leq x^U\} &= 0.5 \cdot \left[\Phi\left(\frac{x^U + 0.5}{0.25}\right) + \Phi\left(\frac{x^U - 0.75}{0.25}\right) \right] \\ &- 0.5 \cdot \left[\Phi\left(\frac{x^L + 0.5}{0.25}\right) + \Phi\left(\frac{x^L - 0.75}{0.25}\right) \right] \end{aligned} \tag{13}$$

with $\Phi(\cdot)$ denoting the cumulative distribution function of a standard normal, $\mathcal{N}(0, 1)$

Therefore,

$$\mathcal{R}_g(t) = \sup_{s \in \mathcal{I}_g(t)} \mathcal{P}\{\xi \in s\}$$

where $\mathcal{I}_g(t) = \{s = (x^L, x^U] \subset \mathbb{R} \text{ such that } \mathcal{P}\{\chi \in s\} \leq t\}$ is equivalent to

$$\mathcal{R}_g(t) = 0.5 \cdot \sup_{x^L \in \mathcal{I}^g(t)} H(x^L)$$

where $\mathcal{I}^g(t) = (-\infty, 0.75 \cdot \Phi^{-1}(1 - t)]$, with $\Phi^{-1}(\cdot)$ is the quantile distribution function of a standard normal, and

$$\begin{aligned} H(x^L) &= \Phi\left(\frac{0.75 \cdot \Phi^{-1}\left(t + \Phi\left(\frac{x^L}{0.75}\right)\right) + 0.5}{0.25}\right) - \Phi\left(\frac{x^L + 0.5}{0.25}\right) \\ &+ \Phi\left(\frac{0.75 \cdot \Phi^{-1}\left(t + \Phi\left(\frac{x^L}{0.75}\right)\right) - 0.75}{0.25}\right) - \Phi\left(\frac{x^L - 0.75}{0.25}\right) \end{aligned}$$

by using the substitution $x^U = 0.75 \cdot \Phi^{-1}\left(t + \Phi\left(\frac{x^L}{0.75}\right)\right)$ from $t = 1 - Sp(x^L, x^U) = \mathcal{P}\{\chi \in (x^L, x^U]\}$.

Going back to the beginning of the proof, if there exists a transformation $h(\cdot)$ of the marker such that the resulting ROC curve classification subsets over the original space coincide with those subsets underlying the gROC curve, then the highlighted classification regions in Fig. 10, s_{t_1} and s_{t_2} corresponding to $t_1 = 0.6$ and $t_2 = 0.4$, respectively, should fulfill that $s_{t_2} \subseteq s_{t_1}$ since $t_1 > t_2$. However, as it can be seen,

$$s_{t_2} = (0.145, 1.492] \not\subseteq (-0.307, 1.174] = s_{t_1}.$$

That is, we have found a scenario where there is no transformation $h(\cdot)$ of the marker with the ROC curve for such transformation being the same as the gROC curve for the original marker. □

Simulation study about the influence of imposing the restriction (C) on the classification subsets underlying the gROC curve (Sect. 2.2.1)

In order to explore the influence of the restriction (C) on the resulting gROC curves and the impact of the selection of the initial point in Step 2 of the algorithm proposed in Sect. 2.2.1 on the classification subsets, a simulation study was carried out. An analysis of the change on the area under the gROC curve imposing the restriction (C) departing from different FPRs was conducted for different scenarios and sample sizes. Particularly:

Scenario 1. $\chi \sim \mathcal{N}(0, 1)$ and $\xi \sim \mathcal{M}(a, b)$.

Scenario 2. $\chi \sim \mathcal{U}(a, b)$ and $\xi \sim \Delta \times \mathcal{M}(-2, 1) + (1 - \Delta) \times \mathcal{M}(3, 0.5)$ where Δ is a Bernoulli random variable with success probability $\pi = 0.5$.

The parameters a and b were taken for obtaining gAUCs without restrictions of 0.75 and 0.85. The classification subsets and gROC curves without and with restriction (C) are shown in Figure S1 (Supplementary Material). The results are based on $B = 500$ simulations, and these are displayed in Fig 11. The numerical results have also been collected in Table S1 in Supplementary Material.

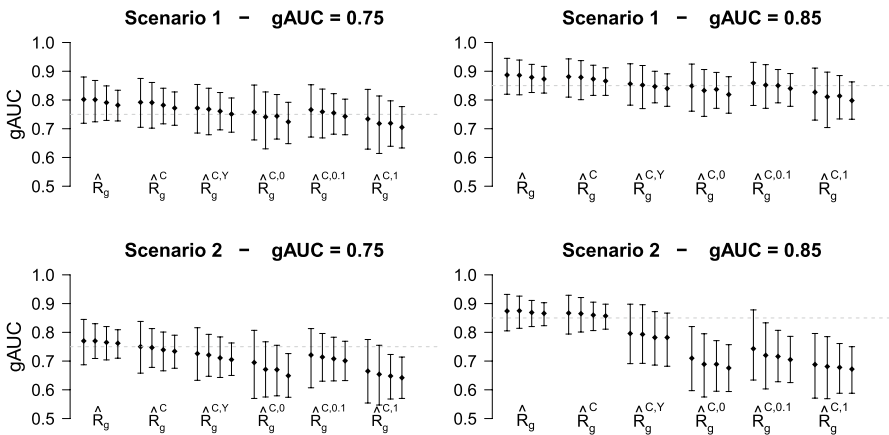


Fig. 11 Results of 500 simulations of Scenario 1 and 2 for different sample sizes and theoretical gAUC. The mean (95% C.I.) for every estimator is displayed. For each one, the four vertical lines correspond to $n = m = 50, 2n_c = m = 100, n = m = 100, 2n = m = 200$. \hat{R}_g denotes the estimated gROC curve without restrictions, \hat{R}_g^C the optimal estimated gROC curve with restriction (C), and $\hat{R}_g^{C,Y}, \hat{R}_g^{C,0}, \hat{R}_g^{C,0.1}$ and $\hat{R}_g^{C,1}$ the estimated gROC curve with restriction (C) considering as starting point in Step 2 of the algorithm the FPR related to the Youden index, FPR = 0, FPR = 0.1 and FPR = 1, respectively

From the results, it can be seen that the area under the optimal gROC curve with the restriction (C) is similar to the gAUC without restrictions. Only a small decreasing is observed between $\hat{\mathcal{R}}_g$ and $\hat{\mathcal{R}}_g^C$, being 0.028 the maximum difference in gAUC means for Scenario 2. This scenario has been designed to be pathological regarding the non-compliance of restriction (C) (Figure S1).

The estimation of the optimal restricted gROC curve is computationally time-consuming for high sample sizes. In those cases, the suggestion is to use the FPR reported by the Youden index as the initial point in Step 2, because it results in higher AUCs for all the scenarios considered, compared to other initial points 0, 0.1 and 1. Despite its superiority, in the Scenario 2 with gAUC= 0.85, $\hat{\mathcal{R}}_{g,C}^F$ slightly underestimates the area (with a maximum difference in means of 0.07), but the optimal $\hat{\mathcal{R}}_g^C$ remains giving adequate results.

References

- Aitchison, J., Egozcue, J.J.: Compositional data analysis: where are we and where should we be heading? *Math. Geol.* **37**(7), 829–850 (2005). <https://doi.org/10.1007/s11004-005-7383-7>
- Bamber, D.: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* **12**(4), 387–415 (1975). [https://doi.org/10.1016/0022-2496\(75\)90001-2](https://doi.org/10.1016/0022-2496(75)90001-2)
- Biau, G., Bunea, F., Wegkamp, M.H.: Functional classification in Hilbert spaces. *IEEE Trans. Inf. Theory* **51**(6), 2163–2172 (2005). <https://doi.org/10.1109/TIT.2005.847705>
- Chen, B., Li, P., Qin, J., Yu, T.: Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests. *J. Am. Stat. Assoc.* **111**(514), 861–874 (2016). <https://doi.org/10.1080/01621459.2015.1066681>
- Fisher, R.A.: The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188 (1936). <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>
- Floege, J., Kim, J., Ireland, E., Chazot, C., Drucke, T., de Francisco, A., Kronenberg, F., Marcelli, D., Passlick-Deetjen, J., Schernthaner, G., Fouqueray, B., Wheeler, D.C., Investigators, A.: Serum iPTH, calcium and phosphate, and the risk of mortality in a European haemodialysis population. *Nephrol. Dial. Transplant.* **26**(6), 1948–1955 (2011). <https://doi.org/10.1093/ndt/gfq219>
- Fluss, R., Faraggi, D., Reiser, B.: Estimation of the youden index and its associated cutoff point. *Biom J* **47**(4), 458–472 (2005). <https://doi.org/10.1002/bimj.200410135>
- Gardner, J.G., Bhamidipati, D.R., Rueda, A.M., Graviss, E., Nguyen, D., Musher, D.M.: The white blood cell count and prognosis in pneumococcal pneumonia. *Open Forum Infect. Dis.* (2016). <https://doi.org/10.1093/ofid/ofw172.948>
- Green, D.M., Swets, J.A.: *Signal Detection Theory and Psychophysics*. Wiley, New York (1966)
- Hall, P., Poskitt, D.S., Presnell, B.: A functional data-analytic approach to signal discrimination. *Technometrics* **43**(1), 1–9 (2001). <https://doi.org/10.1198/00401700152404273>
- Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**(1), 29–36 (1982). <https://doi.org/10.1148/radiology.143.1.7063747>
- Kang, L., Liu, A., Tian, L.: Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Stat. Methods Med. Res.* **25**(4), 1359–1380 (2016). <https://doi.org/10.1177/0962280213481053>
- Kauppi, H.: The Generalized Receiver Operating Characteristic Curve. Discussion paper 114, Aboa Centre for Economics (2016)
- Liu, C., Liu, A., Halabi, S.: A min–max combination of biomarkers to improve diagnostic accuracy. *Stat. Med.* **30**(16), 2005–2014 (2011). <https://doi.org/10.1002/sim.4238>
- López-Ratón, M.: Optimal cutoff points for classification in diagnostic studies: new contributions and software development. Ph.D. Thesis, Universidade de Santiago de Compostela, (2015). <http://hdl.handle.net/10347/14593>

- Martínez-Camblor, P., Pardo-Fernández, J.C.: Parametric estimates for the receiver operating characteristic curve generalization for non-monotone relationships. *Stat. Methods Med. Res.* **28**(7), 2032–2048 (2019). <https://doi.org/10.1177/0962280217747009>
- Martínez-Camblor, P., Corral, N., Rey, C., Pascual, J., Cernuda-Morollón, E.: Receiver operating characteristic curve generalization for non-monotone relationships. *Stat. Methods Med. Res.* **26**(1), 113–123 (2017). <https://doi.org/10.1177/0962280214541095>
- Martínez-Camblor, P., Pérez-Fernández, S., Díaz-Coto, S.: Improving the biomarker diagnostic capacity via functional transformations. *J. Appl. Stat.* **46**(9), 1550–1566 (2019). <https://doi.org/10.1080/02664763.2018.1554628>
- McClish, D.K., Powell, S.H.: How well can physicians estimate mortality in a medical intensive care unit? *Med. Decis. Mak.* **9**(2), 125–132 (1989). <https://doi.org/10.1177/0272989X8900900207>
- McIntosh, M.W., Pepe, M.S.: Combining several screening tests: optimality of the risk score. *Biometrics* **58**(3), 657–664 (2002). <https://doi.org/10.1111/j.0006-341x.2002.00657.x>
- Meisner, A., Carone, M., Pepe, M. S., Kerr, K. F.: Combining biomarkers by maximizing the true positive rate for a fixed false positive rate. UW Biostatistics Working Paper Series (Working Paper 420) (2017)
- Nielsen, J.D., Rumí, R., Salmerón, A.: Supervised classification using probabilistic decision graphs. *Comput. Stat. Data Anal.* **53**(4), 1299–1311 (2009). <https://doi.org/10.1016/j.csda.2008.11.003>
- Pepe, M.S.: *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford (2003)
- Pepe, M.S., Thompson, M.L.: Combining diagnostic test results to increase accuracy. *Biostatistics* **1**(2), 123–140 (2000). <https://doi.org/10.1093/biostatistics/1.2.123>
- Pepe, M.S., Cai, T., Longton, G.: Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62**(1), 221–229 (2006). <https://doi.org/10.1111/j.1541-0420.2005.00420.x>
- Pérez-Fernández, S., Martínez-Camblor, P., Filzmoser, P., Corral, N.: nsROC: an R package for non-standard ROC curve analysis. *R J.* **10**(2), 55–77 (2018). <https://doi.org/10.32614/RJ-2018-043>
- Su, J.Q., Liu, J.S.: Linear combinations of multiple diagnostic markers. *J. Am. Stat. Assoc.* **88**(424), 1350–1355 (1993). <https://doi.org/10.2307/2291276>
- Xu, T., Fang, Y., Rong, A., Wang, J.: Flexible combination of multiple diagnostic biomarkers to improve diagnostic accuracy. *BMC Med. Res. Methodol.* **15**(1), 94 (2015). <https://doi.org/10.1186/s12874-015-0085-z>
- Yan, L., Tian, L., Liu, S.: Combining large number of weak biomarkers based on AUC. *Stat. Med.* **34**(29), 3811–3830 (2015). <https://doi.org/10.1002/sim.6600>
- Yin, J., Tian, L.: Optimal linear combinations of multiple diagnostic biomarkers based on Youden index. *Stat. Med.* **33**(8), 1426–1440 (2014). <https://doi.org/10.1002/sim.6046>
- Zhang, H.: Classification trees for multiple binary responses. *J. Am. Stat. Assoc.* **93**(441), 180–193 (1998). <https://doi.org/10.2307/2669615>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Sonia Pérez-Fernández¹  · Pablo Martínez-Camblor²  · Peter Filzmoser³  · Norberto Corral¹ 

¹ Department of Statistics and O.R. and M.D., University of Oviedo, Oviedo, Spain

² Geisel School of Medicine at Dartmouth, Hanover, NH, USA

³ Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria