Optimal classification scores based on multivariate marker transformations

Pablo Martínez-Camblor[1]*, Sonia Pérez-Fernández[2], Susana Díaz-Coto[2]

[1] Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, NH, USA

[2] Department of Statistics, Oviedo University, Asturies, Spain

Abstract

Modern science frequently involves the study of complex relationships among effects and factors. Flexible statistical tools are commonly used to visualize non-linear associations. When our interest is to study the discrimination capacity of a multivariate marker on a binary outcome, the theoretical transformation leading to the optimal results in terms of sensitivity and specificity has already been settled. It is particularly useful to know this function, not only to allocate items to groups, but also to understand the relationship between the multivariate marker and the outcome. In this paper, we explore the use of the multivariate kernel density estimator in order to approximate such transformation. Large sample properties of the finally derived estimator are outlined while its finite sample behavior is studied via Monte Carlo simulations. We consider six different bivariate and three additional higher dimensional scenarios. The performance of the estimator is studied by using four different tuning parameters computed automatically. Besides a cross-validation algorithm is incorporated with the aim of reducing the potential overfitting. The proposed methodology is applied in order to study the capacity of two molecular characteristics to predict the toxicity of some chemical products. Results suggest that smoothing techniques are promising classical and simple statistical tools which can be used for a better understanding of some current scientific problems. However, the incorporation of additional machine learning techniques such as cross-validation is advisable in order to control the frequently over optimistic results, specially in those cases with small sample size. The function implementing the proposed methodology is provided as supplementary material.

Keywords: Classification problem; Kernel density estimator; Multivariate marker; Optimal transformation; Receiver-operating characteristic (ROC) curve.

*Pablo Martínez-Camblor. 7 Lebanon Street, Suite 309, Hinman Box 7261, Hanover, NH 03755, USA. E-mail: Pablo.Martinez.Camblor@Dartmouth.edu

1 Introduction

Studying associations among some substances and outcomes frequently involves dealing with complex structures, embracing interactions, non-linear and/or non-additive relationships. Examples include the estimation of the effects of multi-pollutant mixtures such as chemical products, air pollution or mixtures of toxic waste on population health [42]. Also the use of microRNAs (evolutionarily conserved small noncoding RNAs that post-transcriptionally regulate gene expression) as diagnostic and prognostic markers for diverse cardiovascular and metabolic disorders. The strong and complex internal structure of those microRNAs makes that standard linear statistical analyses potentially do not capture their optimal classification capacity [9].

Standard parametric statistical techniques do not address these challenges and, in these contexts, the performance of more flexible procedures should be explored. Bobb et al. [1] proposed the use of kernel machine regression for estimating the health effects of multi-pollutant mixtures with a previous hierarchical Bayesian variable selection in the so-called *Bayesian kernel machine regression* (BKMR) algorithm. de Gonzalo-Calvo et al. [10] explored the application of statistical-learning algorithms, particularly classification tree models [3], in order to consider high-order interactions among the microRNAs and traditional clinical markers for identifying risk groups. One of the common underlying features of those procedures is the consideration of more flexible relationships among the markers and the outcome while providing hints about the direction and shape of these relations.

We consider here the case in which we have a continuous multivariate marker and a binary outcome determining whether the subjects have the studied characteristic (frequently a disease) or not and we are interested in studying the ability of such multivariate marker to correctly classify the subjects. There are several machine learning procedures [20] which explore complex relationships among the marker components and the outcome in order to get accurate classification processes. These techniques include support vector machine (SVM) [8], boosting [16] or perceptron [6], among others. The `CARET` [24] package implements in `R` most of these algorithms. The common goal of all those techniques is to get an accurate classification while, in the underlying algorithms, the rules which lead to allocate one item to one particular group are usually relegated to the so-called *black box* step. Besides, some of these procedures report only the predicted group. Therefore, any preference for reducing the false-negative or the false-positive rate has to be previously included in the method.

We focus on the case in which we have a relatively low dimensional marker and we are interested in getting some understanding of the underlying classification rules. That is, those reasons that make an item be more likely to be within a group. With this goal, the overwhelming procedure is to reduce the marker from multivariate to a univariate score based on some linear combination of the original components. Such score is frequently computed via logistic regression although other techniques such as discriminant analysis have also been employed [34] (see Pérez-Fernández et al. [36] for a recent revision of this topic). In order to get the score, we can assume, without loss of generality, that higher values are associated with having a higher probability of being a positive subject (with the characteristic). The pairs formed by the sensitivity and the specificity for all possible derived classification rules are plotted in the so-called receiver-operating characteristic (ROC) curve [17]. The area under this curve (AUC) is frequently used as an index of the overall classification capacity [19]. Reaching an AUC as large as possible has become the final objective (see, for instance, Huang et al. [22] and references therein). McIntosh and Pepe [33] proved that the optimal transformation of a multivariate continuous marker, in terms of getting the optimal binary classification capacity among those reported by any other transformation (and in consequence the largest AUC), is determined by the ratio between the distributions of the marker on the groups defined by the binary outcome. This ratio allows to: i) know the relationship between the marker and the outcome and, ii) know the classification accuracy that one can reach based on such marker. The objective of this paper is to examine the use of multivariate kernel density estimators [41] to find the optimal transformation of multivariate markers in binary diagnostic tasks. With this goal, we use the results reported by McIntosh and Pepe [33] to get the optimal theoretical transformation first, and then smooth statistical techniques [2] for its practical estimation. Notice that, in this sense, our primary objective is not to reach a large AUC but the adequate one. Besides, it is known that the AUC does not identify the ROC curve [26, 27] and that some procedures, such as linear logistic regression, can provide correct AUCs but based on wrong classification rules (see, for instance, Díaz-Coto et al. [12]). Therefore, our focus is to get adequate ROC curve approximations based on a smooth estimator for the optimal transformation. The rest of the paper is organized as follows. In Section 2, we present the theoretical framework including both the uniform consistency and the pointwise asymptotic normality of the resulting estimator for the optimal transformation. In Section 3, we study the finite sample behavior of the proposed procedure via Monte Carlo simulations. We report those results in terms of distances to the real ROC curve. The potential influence of the bandwidth on the obtained estimation is explored by using different estimation procedures [14], while the potential impact of the overfitting is corrected through a standard k-fold cross-validation procedure. Section 4 is devoted to explore the use of molecular characteristics in order to predict the toxicity (defined as high values of LC_{50} 96 hours) for diverse organic molecules towards the fathead minnow (Pimephales promelas). Finally, in Section 5, we summarize our conclusions. The R code used for the implementation of the proposed methodology is provided as online supplementary material. The main R function included, `optimalT`, incorporates a cross-validation procedure which controls the potential overfitting, as well as a flexible choice of the bandwidth computation among those proposed by Duong [14].

2 Theoretical framework

We consider Y a binary variable indicating whether a subject has the characteristic in study ($Y = 1$) or not ($Y = 0$) and a multivariate random variable, $\boldsymbol{X}$, modeling the behavior of the multivariate continuous marker. Using the celebrated Neyman-Pearson lemma, McIntosh and Pepe [33] proved that, in terms of sensitivity and specificity, the optimal classification rules based on the marker $\boldsymbol{X}$ are those derived from the transformation $T(\boldsymbol{X}) = f(\boldsymbol{X})/g(\boldsymbol{X})$ or, equivalently,

$$T(\boldsymbol{X}) = \frac{f(\boldsymbol{X})}{f(\boldsymbol{X}) + g(\boldsymbol{X})}, \tag{1}$$

where $f(\cdot)$ and $g(\cdot)$ are the multivariate density functions of the marker $\boldsymbol{X}$ in the positive ($Y = 1$) and the negative ($Y = 0$) populations, respectively. It should be noted that the decision rules for the marker $f(\boldsymbol{X})/g(\boldsymbol{X})$ are equivalent to those for the score $f(\boldsymbol{X})/(f(\boldsymbol{X})+g(\boldsymbol{X}))$ since the latter is an increasing monotone transformation of $f(\cdot)/g(\cdot)$ and the ROC curve is invariant to this type of transformations.

This *Neyman-Pearson approach* has already been considered in the literature. For instance, Scott and Nowak [40] connected the used terminology with Statistical Learning nomenclature. Qin and Zhang [37] considered the problem of estimating $f(\boldsymbol{u})/g(\boldsymbol{u})$ by assuming

$$\frac{f(\boldsymbol{u})}{g(\boldsymbol{u})} = \exp\{\alpha + \boldsymbol{\beta}^\top \cdot \gamma(\boldsymbol{u})\},$$

where α is a scalar, $\boldsymbol{\beta}$ is a vector and $\gamma(\cdot)$ is a smooth vector function. Chen et al. [5] considered the estimation under the assumption that

$$\frac{f(\boldsymbol{u})}{g(\boldsymbol{u})} = \psi(\boldsymbol{\beta}^\top \cdot \gamma(\boldsymbol{u})),$$

where $\psi(\cdot)$ is an unknown monotonic nondecreasing function. The function $\gamma(\cdot)$ is predetermined in both papers and no hints about its computation are provided.

We consider applying a plug-in method to estimate the function $T(\cdot)$. Particularly, given a generic independent and identically distributed (i.i.d.) random vector, $\{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n\}$, from a d-dimensional random variable, $\boldsymbol{Z}$, with density function $\ell(\cdot)$, the multivariate kernel density estimator [39] is defined by,

$$\hat{\ell}_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{u} - \boldsymbol{Z}_i), \tag{2}$$

where, in $K_{\boldsymbol{H}}(\boldsymbol{u}) = |\boldsymbol{H}|^{-1} \cdot K(\boldsymbol{H}^{-1} \cdot \boldsymbol{u})$, $K(\cdot)$ is a kernel function and $\boldsymbol{H}$ is a $d \times d$ nonsingular matrix containing the tuning parameters or bandwidth. The kernel function, $K(\cdot)$, is assumed to be a multivariate probability density centered in zero and with covariance matrix the identity, that is, $K(\cdot)$ satisfies:

a) $\int_{\mathbb{R}^d} K(\boldsymbol{u})d\boldsymbol{u} = 1$,

b) $\int_{\mathbb{R}^d} \boldsymbol{u} \cdot K(\boldsymbol{u})d\boldsymbol{u} = 0$, and

c) $\int_{\mathbb{R}^d} \boldsymbol{u}\boldsymbol{u}^T \cdot K(\boldsymbol{u})d\boldsymbol{u} = \boldsymbol{I}_d$.

Since the impact of the kernel function on the obtained estimations is not relevant (see, for instance, Silverman [43]), for simplicity sake, hereafter, $K(\cdot)$ is chosen to be the density function of a standard d-dimensional normal variable (thus satisfying the assumptions b) and c)). Devroye and Penrod [11] proved that, for K satisfying a) and b) and if $\boldsymbol{H} = b \cdot \boldsymbol{D}$, where $\boldsymbol{D}$ is a $d \times d$ matrix with $|\boldsymbol{D}| = 1$ and with b satisfying

d) $b = Cte \cdot n^{-1/(d+4)}$ with Cte a positive constant, then, assuming that the real density function, $\ell(\cdot)$, is smooth enough (having at least two continuous and bounded derivatives is required), we have that

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d} |\hat{\ell}_n(\boldsymbol{u}) - \ell(\boldsymbol{u})| \longrightarrow_n 0 \quad a.s. \tag{3}$$

Furthermore, for any fixed $\boldsymbol{u} \in \mathbb{R}^d$, Hall [18] proved that

$$\frac{\sqrt{nb^d}}{\sigma_d(\boldsymbol{u})} \cdot \left\{ [\hat{\ell}_n(\boldsymbol{u}) - \ell(\boldsymbol{u})] + b^2 B_d(\boldsymbol{u}) \right\} \longrightarrow_n \mathscr{N}(0,1), \tag{4}$$

where $\sigma_d^2(\boldsymbol{u}) = \ell(\boldsymbol{u}) \int K(\boldsymbol{u})^2 d\boldsymbol{u}$ and $B_d(\boldsymbol{u}) = -1/2 \cdot tr\{\boldsymbol{D}\boldsymbol{D}^T \nabla^2 \ell(\boldsymbol{u})\}$ are the asymptotic variance and bias of the kernel density estimator, respectively. Condition d) is stronger than it is required; theoretical results allow a wider range for the convergence ratio of the bandwidth. However, $\boldsymbol{H}$ is usually selected to minimize the mean integrated square error and, in this case, the optimal convergence ratio is the one asked in the condition d) [41]. The problem of selecting the optimal bandwidth is reduced to the estimation of the value of Cte and the elements of the matrix $\boldsymbol{D}$. Different procedures have been proposed with this goal (see, for instance, Duong [13]) although, unfortunately, there is not a uniformly best solution.

Given an i.i.d. random vector, $\{(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_N, Y_N)\}$, with $n = \sum_{i=1}^N Y_i$ and $m = N - n$, for each $\boldsymbol{u} \in \mathbb{R}^d$, we propose to estimate the optimal transformation T through its natural smooth estimator, that is

$$\hat{T}_N(\boldsymbol{u}) = \frac{m \cdot \sum_{i=1}^N K_{\boldsymbol{H}_1}(\boldsymbol{u} - \boldsymbol{X}_i) \cdot Y_i}{m \cdot \sum_{i=1}^N K_{\boldsymbol{H}_1}(\boldsymbol{u} - \boldsymbol{X}_i) \cdot Y_i + n \cdot \sum_{i=1}^N K_{\boldsymbol{H}_0}(\boldsymbol{u} - \boldsymbol{X}_i) \cdot (1 - Y_i)}, \tag{5}$$

where $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ are the matrices containing the bandwidths for the negative and the positive population, respectively. If i) $g(\cdot)$ has two continuous and bounded derivatives; ii) $K(\cdot)$ satisfies a), b) and c); iii) both $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ satisfy d); and iv) $n/m \rightarrow_n \tau > 0$, we have that, from Eq. (4), for a fixed $\boldsymbol{u} \in \mathbb{R}^d$,

$$\begin{aligned} \hat{f}_n(\boldsymbol{u}) &= f(\boldsymbol{u}) + o_P(N^{-2/(d+4)}), \\ \hat{g}_m(\boldsymbol{u}) &= g(\boldsymbol{u}) + o_P(N^{-2/(d+4)}). \end{aligned}$$

Therefore,

$$\begin{aligned}\hat{T}_N(\boldsymbol{u}) - T(\boldsymbol{u}) =& \frac{\hat{f}_n(\boldsymbol{u}) \cdot [f(\boldsymbol{u}) + g(\boldsymbol{u})] - f(\boldsymbol{u}) \cdot [\hat{f}_n(\boldsymbol{u}) + \hat{g}_m(\boldsymbol{u})]}{[\hat{f}_n(\boldsymbol{u}) + \hat{f}_n(\boldsymbol{u})] \cdot [f(\boldsymbol{u}) + g(\boldsymbol{u})]} \\ =& \frac{\hat{f}_n(\boldsymbol{u}) \cdot g(\boldsymbol{u}) - f(\boldsymbol{u}) \cdot \hat{g}_m(\boldsymbol{u})}{[f(\boldsymbol{u}) + g(\boldsymbol{u})]^2 + O_P(N^{-2/(d+4)})} \\ =& \frac{g(\boldsymbol{u})[\hat{f}_n(\boldsymbol{u}) - f(\boldsymbol{u})] - f(\boldsymbol{u})[\hat{g}_m(\boldsymbol{u}) - g(\boldsymbol{u})]}{[f(\boldsymbol{u}) + g(\boldsymbol{u})]^2} + o_P(N^{-2/(d+4)}). \qquad (6)\end{aligned}$$

Then, if $f(\cdot)$ has two continuous and bounded derivatives, both the uniform consistency and the pointwise weak convergence of $\hat{T}_N(\cdot)$ can be derived from (3) and (4), respectively.

Large sample properties for the resulting empirical ROC curve estimator [21] are straightforward. Let $\mathcal{R}_T(\cdot)$ be the ROC curve associated with the marker $T(\boldsymbol{X})$ and $\hat{\mathcal{R}}_T(\cdot)$ its empirical estimator (for $T(\cdot)$ fixed). Notice that, with this notation, $\mathcal{R}_{\hat{T}}(\cdot)$ and $\hat{\mathcal{R}}_{\hat{T}}(\cdot)$ denote the real and the empirical estimator for the ROC curve, respectively, associated with the marker $\hat{T}_N(\boldsymbol{X})$ ($\hat{T}_N(\cdot)$ estimated). For each $t \in (0, 1)$, we have that

$$\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_T(t) = \hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_{\hat{T}_N}(t) + \mathcal{R}_{\hat{T}_N}(t) - \mathcal{R}_T(t). \qquad (7)$$

Under assumptions a), b), c) and d) and assuming that both $f(\cdot)$ and $g(\cdot)$ have two continuous and bounded derivatives, ROC curve properties guarantee the uniform consistency for $|\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_{\hat{T}_N}(t)|$, while $\mathcal{R}(\cdot)$ continuity and kernel density estimator consistency guarantee the uniform consistency for $|\mathcal{R}_{\hat{T}_N}(t) - \mathcal{R}_T(t)|$. Hsieh and Turnbull [21] guarantees the asymptotic normality for

$$\sqrt{N} \cdot [\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_{\hat{T}_N}(t)],$$

and, since $\hat{T}_N(\cdot)$ converges slower than $\hat{R}_T(\cdot)$, the regularity conditions satisfied by $f(\cdot)$ and $g(\cdot)$ imply that

$$\sqrt{N} \cdot [\mathcal{R}_{\hat{T}_N}(t) - \mathcal{R}_T(t)] \longrightarrow_N 0,$$

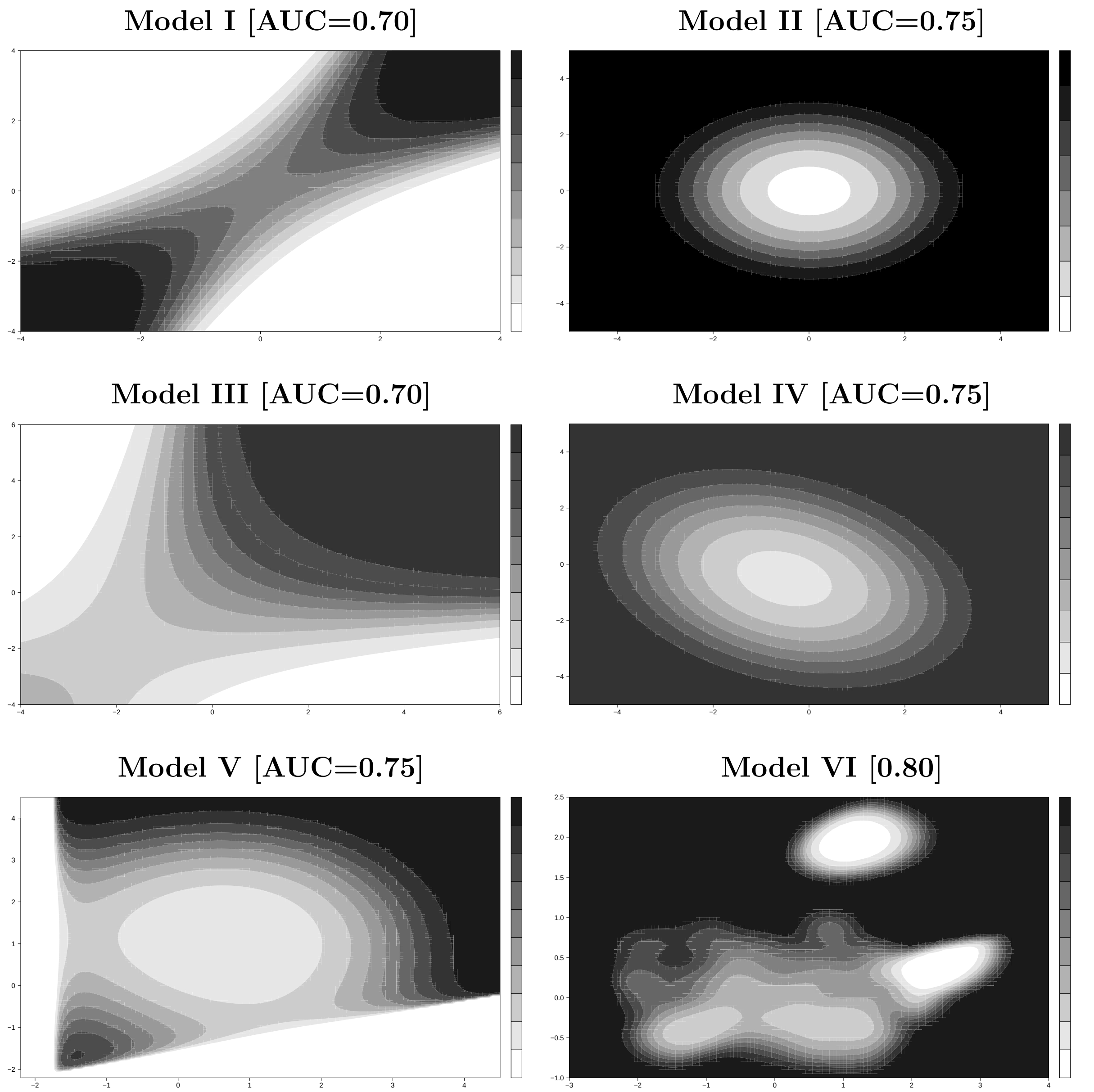
and therefore, we have then the asymptotic normality for $\sqrt{N} \cdot [\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_T(t)]$.

3 Monte Carlo simulation study

The practical behavior of the proposed procedure is studied via Monte Carlo simulations. First, we consider a two-dimensional marker following a bivariate normal distribution with mean vector zero and covariance matrix the identity, $\boldsymbol{I}_2$, for the negative (without the characteristic) population and five different bivariate distributions for the positive (with the characteristic) population. In models I and II, the only difference between the distribution in the positive and the negative populations is in the covariance matrix. In model I, we introduce a positive correlation coefficient, ρ; and in model II, joint with this correlation parameter, the variances of the marginal markers are 3. Different values for ρ are selected in order to obtain different classification accuracy. In models III and IV, the positive distribution is also normal, but we consider some differences in the location parameter. In model III, the covariance matrix has 1 in the main-diagonal and the correlation is 1/4, while the mean vector is (μ_1, μ_1). In model IV, the covariance matrix has 2 in the main-diagonal and the correlation is also 1/4, while the mean vector is (μ_2, μ_2). Values for μ_1 and μ_2 are selected to have different discrimination accuracy. In model V, we consider an asymmetric distribution for the positive population. Particularly, we compute the first component as $Q_1 = (1/\sqrt{8}) \cdot (\chi^2_4[1] - 4) + \mu$, where $\chi^2_4[1]$ represents a chi-2 variable with four degrees of freedom, and the second component by $Q_2 = \rho \cdot (Q_1 - \mu) + ((1 - \rho^2)/8)^{1/2} \cdot (\chi^2_4[2] - 4) + \mu$, with $\chi^2_4[2]$ another chi-2 distributed random variable with four degrees of freedom independently drawn from $\chi^2_4[1]$, $\rho = 0.3$ and μ chosen to obtain different classification accuracy. Finally, in model VI, we study the situation in which the distribution in both the negative and the positive populations follow the structure of the real dataset considered in this document (see Section 4). We first center both the negative and the positive samples and then we compute the densities (considered as real) by using the bivariate kernel density estimator with biased cross-validation bandwidths (BCV). Finally, we run samples from those densities adding the quantity $(-\mu, \mu)$ to the positive sample, where μ is selected to obtain different discrimination capacities.

Figure 3 shows the contour plots for the function $f(\cdot)/(f(\cdot)+g(\cdot))$ where $f(\cdot)$ and $g(\cdot)$ are the density functions in the positive and in the negative populations, respectively. Darker colors indicate a higher probability of being within the positive population. Following, a schematic description of the considered models:

Model I.- Normal: mean zero, variances 1 and correlation ρ (= 0.68, 0.87).
Model II.- Normal: mean zero, variances 3 and correlation ρ (= 0.01, 0.88).



Model I [AUC=0.70]

Model II [AUC=0.75]

Model III [AUC=0.70]

Model IV [AUC=0.75]

Model V [AUC=0.75]

Model VI [0.80]

Figure 1: Contour plots for the function $f(\cdot)/(f(\cdot)+g(\cdot))$ where $f(\cdot)$ and $g(\cdot)$ are the density functions in the positive and in the negative populations. Darker colors indicate a higher probability of being within the positive population.

Model III.- Normal: mean $\mu_1 \cdot (1, 1)$ (= 0.53, 0.88), variances 1 and correlation 1/4.

Model IV.- Normal: mean $\mu_2 \cdot (1, 1)$ (= 0.78, 1.33), variances 2 and correlation 1/4.

Model V.- Asymmetric distribution based on translated chi-2: $\mu = -0.30, -0.91$ and correlation 0.3.

Model VI.- Real problem (Section 4) based samples: μ = 0.49, 0.61.

Table 1 shows the mean of the integrate absolute error (Integ. absolute error) between the real ROC curve, $\mathcal{R}_T(\cdot)$, and its estimation, $\hat{\mathcal{R}}_{\hat{T}_N}(\cdot)$ ($\int_0^1 |\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_T(t)|dt$), and the mean of the estimated AUCs based on 2,000 Monte Carlo simulations drawn from the six previously described models. Additionally, the case in which both the positive and the negative populations follow a standard bivariate normal distribution (model 0) is also shown. We consider different sample sizes for the positive, n, and the negative, m, populations and different real AUCs ($\mathcal{A}$). Besides, four different procedures for estimating the multivariate bandwidth are considered: smooth cross-validation (SCV), plug-in (PI), normal scale (NS) and biased cross-validation (BCV). A fully description of all these procedures can be found in Duong [13]. The estimation procedure includes a 2-fold cross-validation algorithm: the data are randomly split in two halves and the value of the function for each subject is based on the estimation obtained from the half in which such subject is not included. As a reference method, we compute the ROC curve (and its AUC) for the predictive model based on the linear combination of the two components resulting from a binary logistic regression (RL), applying a similar 2-fold cross-validation procedure.

As it was expected, logistic regression-based procedure does not detect complex classification rules. It fails to find any classification capacity of the marker when this is based on differences in the variance (model I and II). It performs better when these differences are based on location parameters (model III) and does similar when they are mainly based on the location parameter but also include other components (model IV). In the last two studied models, the structure of data is more complex. In model V, the logistic regression approach does not capture all the differences and the proposed methodology gets better results, but still slightly far away from the real ROC curves. In model VI, all the procedures, including logistic regression, report good results. It is worth mentioning that the proposed smooth estimator always under-estimates the real discrimination accuracy. The implemented 2-fold algorithm solves the common overfitting problem, even when the marker does not distinguish between the two populations (model 0) or for smaller sample sizes (see Table S2 in the supplementary material). Remark that the proposed estimator without this 2-fold algorithm always overestimates the results (see Table S1 and S3). The overfitting problem is not surprising and does not seem to be very serious when the real capacity of the marker to classify is medium-large; however, it could be problematic for

Table 1: Means for the integrate absolute error (Integ. absolute error) between the real ROC curve, $\mathcal{R}_T(\cdot)$, and its estimation, $\hat{\mathcal{R}}_{\hat{T}_N}(\cdot)$ ($\int_0^1 |\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_T(t)|dt$) and for the AUC from 2000 Monte Carlo simulations for the six considered models by using 2-fold cross-validation procedure. Sample sizes were n and m for positive and negative groups, respectively, $\mathcal{A}$ is the real AUC. Considered bandwidths were smooth cross-validation (SCV), plug-in (PI), normal scale (NS) and biased cross-validation (BCV). RL stands for model based on standard binary logistic regression.

			AUC					Integ. absolute error				
n	m	$\mathcal{A}$	**SCV**	**PI**	**NS**	**BCV**	**RL**	**SCV**	**PI**	**NS**	**BCV**	**RL**
Model 0												
400	400	0.50	**0.501**	**0.501**	0.502	0.502	0.493	0.028	0.028	0.028	0.028	**0.026**
	600	0.50	**0.501**	**0.501**	**0.501**	**0.501**	0.492	0.024	0.024	0.024	0.024	**0.023**
Model I												
400	400	0.70	**0.681**	0.678	0.679	0.673	0.491	**0.029**	0.030	0.030	0.034	0.210
	600	0.70	**0.682**	0.680	0.681	0.674	0.491	**0.026**	0.028	0.027	0.032	0.210
400	400	0.80	**0.790**	0.789	0.789	0.779	0.491	**0.021**	0.022	0.022	0.028	0.311
	600	0.80	**0.791**	0.790	0.790	0.781	0.492	**0.019**	0.020	0.020	0.025	0.311
Model II												
400	400	0.75	0.735	0.733	0.734	**0.736**	0.494	**0.024**	0.025	0.025	**0.024**	0.255
	600	0.75	0.735	0.734	0.734	**0.736**	0.492	0.024	0.024	0.024	**0.023**	0.257
400	400	0.80	**0.788**	0.787	0.787	0.769	0.492	**0.022**	**0.022**	**0.022**	0.034	0.306
	600	0.80	**0.789**	0.787	0.788	0.769	0.492	**0.020**	**0.020**	**0.020**	0.034	0.307
Model III												
400	400	0.70	0.677	0.674	0.675	0.677	**0.684**	0.030	0.032	0.031	0.030	**0.025**
	600	0.70	0.681	0.679	0.679	0.682	**0.687**	0.026	0.027	0.026	0.025	**0.022**
400	400	0.80	0.786	0.785	0.785	0.787	**0.792**	0.022	0.023	0.023	0.022	**0.020**
	600	0.80	0.790	0.789	0.789	0.791	**0.794**	0.020	0.020	0.020	0.019	**0.018**
Model IV												
400	400	0.75	0.732	0.730	0.730	**0.733**	0.716	0.028	0.029	0.029	**0.027**	0.038
	600	0.75	0.735	0.734	0.734	**0.736**	0.719	0.025	0.026	0.025	**0.024**	0.034
400	400	0.85	0.839	0.837	0.838	**0.840**	0.839	0.021	0.021	0.021	**0.020**	0.020
	600	0.85	0.842	0.841	0.841	**0.843**	0.842	0.019	0.019	0.019	0.019	**0.018**
Model V												
400	400	0.75	0.707	**0.708**	0.705	0.701	0.627	**0.044**	**0.044**	0.046	0.050	0.121
	600	0.75	**0.706**	**0.706**	0.703	0.699	0.630	**0.045**	**0.045**	0.047	0.051	0.119
400	400	0.85	0.825	0.824	0.825	**0.825**	0.810	0.028	0.029	0.028	**0.027**	0.040
	600	0.85	0.825	0.824	**0.825**	**0.825**	0.811	0.028	0.028	**0.027**	**0.027**	0.038
Model VI												
400	400	0.80	0.783	0.782	0.784	**0.785**	**0.785**	0.025	0.026	0.025	0.024	**0.023**
	600	0.80	0.786	0.785	0.786	**0.788**	0.787	0.023	0.023	0.023	0.022	**0.021**
400	400	0.85	0.836	0.836	0.837	0.838	**0.840**	0.021	0.021	0.021	0.020	**0.018**
	600	0.85	0.839	0.838	0.839	0.840	**0.842**	0.019	0.019	0.018	0.018	**0.016**

markers with small diagnostic capacity.

In addition, it should be highlighted that the results observed by using the proposed estimator are quite stable with respect to the different bandwidth matrices considered. Those results can vary slightly with the bandwidth selection, but it is relevant that, with reasonable and standard automatic selections, the results are quite similar.

Finally, in order to check the behavior of the procedure in higher dimensional markers, we consider $d = 4$, 6 and 8 dimensional situations. The marker in the negative populations follows a normal distribution with mean vector zero and covariance matrix the identity, $\boldsymbol{I}_d$ ($d = 4$, 6, 8). In models 4D-0, 6D-0 and 8D-0, the positive subjects follow the same distribution as the negative ones ($\mathcal{A} = 1/2$). In model 4D-I, they follow the distribution described in models I and II with the parameters adjusted to have an AUC of 0.80. Same structure for models 6D-I and 8D-I, where the positive population follows the distribution described in models I, II and III (and IV for $d = 8$) such that $\mathcal{A} = 0.80$. The procedure described by Su and Liu [44] (SL) is included as a reference method.

Table 2 is equivalent to Table 1 for higher dimensional models, considering just three different bandwidths and the model proposed by Su and Liu [44] (SL) as a reference. The proposed algorithm always performs adequately when the marker may have some ability to discriminate between the populations. Even though it reports slightly underestimated values, it always reaches better results than those reported by SL. The well-known difficulties estimating a high dimensional density with a relatively small sample size should be remarked.

4 Detecting toxicity of chemical products

New regulations [38] oblige both manufacturers and importers to empirically test the safety of their products for human health and environment. In this context, the necessity of generating new data to support this statement arises. *In silico* and *in vitro* methodologies allow to study the toxicity of particular chemical compounds based on theoretical or experimental variables called molecular descriptors. We consider the assessment of toxicity towards the fathead minnow (Pimephales promelas) of 908 chemicals, based on two univariate variables: the *2D matrix-based descriptors* (*2D descriptors*) and the *Information indices*. We define the toxicity as those values of LC_{50} 96 hours above

Table 2: Means for the integrate absolute error (Integ. absolute error) between the real ROC curve, $\mathcal{R}_T(\cdot)$, and its estimation, $\hat{\mathcal{R}}_{\hat{T}_N}(\cdot)$ ($\int_0^1 |\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_T(t)|dt$) and for the AUC from 2,000 Monte Carlo simulations for the six considered models by using 2-fold cross-validation procedure. Sample sizes were n and m for positive and negative groups, respectively, $\mathcal{A}$ is the real AUC. Considered bandwidths were smooth cross-validation (SCV), plug-in (PI) and normal scale (NS). SL stands for model based on Su and Liu (1993) optimal linear transformation.

			AUC				Integ. absolute error			
$\boldsymbol{n}$	$\boldsymbol{m}$	$\boldsymbol{\mathcal{A}}$	**SCV**	**PI**	**NS**	**SL**	**SCV**	**PI**	**NS**	**SL**
Model 4D-0										
400	400	0.50	0.499	0.499	0.499	0.502	0.027	0.027	0.027	0.026
	600	0.50	0.498	0.499	0.498	0.501	0.025	0.025	0.025	0.025
Model 4D-I										
400	400	0.80	0.754	0.723	0.753	0.502	0.047	0.077	0.048	0.297
	600	0.80	0.759	0.726	0.758	0.502	0.042	0.074	0.043	0.297
Model 6D-0										
400	400	0.50	0.499	0.499	0.499	0.495	0.025	0.025	0.025	0.028
	600	0.50	0.499	0.500	0.500	0.500	0.023	0.023	0.023	0.023
Model 6D-I										
400	400	0.80	0.718	0.692	0.719	0.628	0.081	0.106	0.080	0.170
	600	0.80	0.725	0.696	0.726	0.632	0.074	0.103	0.073	0.166
Model 8D-0										
400	400	0.50	0.501	0.501	0.501	0.499	0.027	0.026	0.027	0.029
	600	0.50	0.502	0.502	0.502	0.502	0.022	0.023	0.023	0.024
Model 8D-I										
400	400	0.80	0.704	0.691	0.707	0.507	0.096	0.109	0.093	0.292
	600	0.80	0.709	0.692	0.712	0.507	0.091	0.108	0.088	0.292

4.0 (median value for the considered sample). The used dataset is freely available at `http://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity#`. More information about this problem can be found in Cassotti et al. [4].

In the positive samples, *Information indices* ranges between 0.67 and 5.93 with a mean±standard deviation (sd) of 3.08±0.80 (median value of 3.17); in the negative samples, it ranges from 0.96 to 4.81 with mean±sd of 2.72±0.66 (median of 2.75). Figure 2, at top-right, depicts the boxplot for this variable. On the other hand, *2D descriptors* mean±standard deviation are 0.37±0.31 and 0.78±0.47 in the negative and the positive samples, respectively. It ranges from 0 to 2.17 with a median of 0.41 in the negative samples

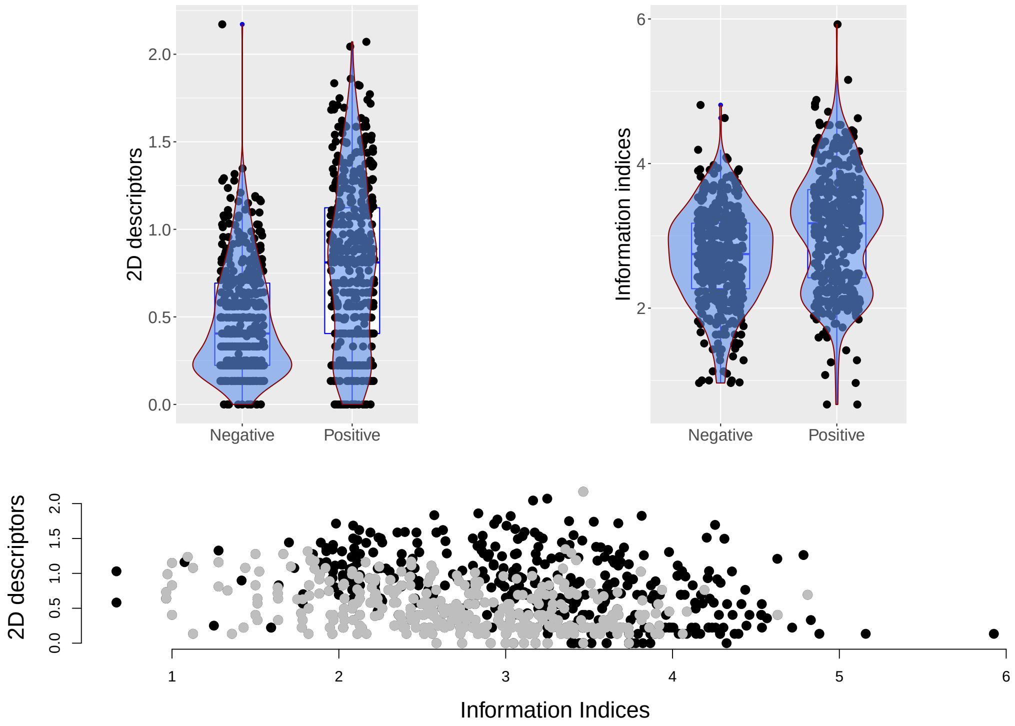


Figure 2: Boxplots for the *2D descriptors*, al left, and the *Information Indices*, at right. At bottom, scatter plot for those variables: in gray, the negative samples, in black, the positive ones.

and from 0 to 2.07 with a median of 0.81 in the positive ones. Figure 2, at top-left, depicts the corresponding boxplot. At bottom, the scatter plot for those variables is shown: in gray, the negative samples, in black, the positive ones.

Individually, markers show a moderate overall classification accuracy. Classifications based on *2D matrix-based descriptors* achieve an AUC of 0.67 (95% confidence interval of (0.64-0.71)), while for those based on the *Information indices* it is 0.63 (0.59-0.67) (the complete ROC curves are shown in Figure 5, dashed and continuous black-thin lines, respectively). When we combine both markers in a logistic regression and use the resulting score ($1.28 \cdot$ *Information indices* $+ 2.90 \cdot$ *2D descriptors*), the overall classification capacity achieves an AUC of 0.80 (0.77-0.82) (the ROC curve is displayed in Figure 5, gray-thick line). When we estimate the optimal transformation from $\hat{T}_N(\cdot)$ with the biased cross-validation (BCV) bandwidth, the AUC obtained is 0.84 (0.81-0.86) (ROC curve in Figure 5, black-thick line). Notice that the confidence intervals for the multivariate model do not consider the additional variability due to the model estimation. When we consider this

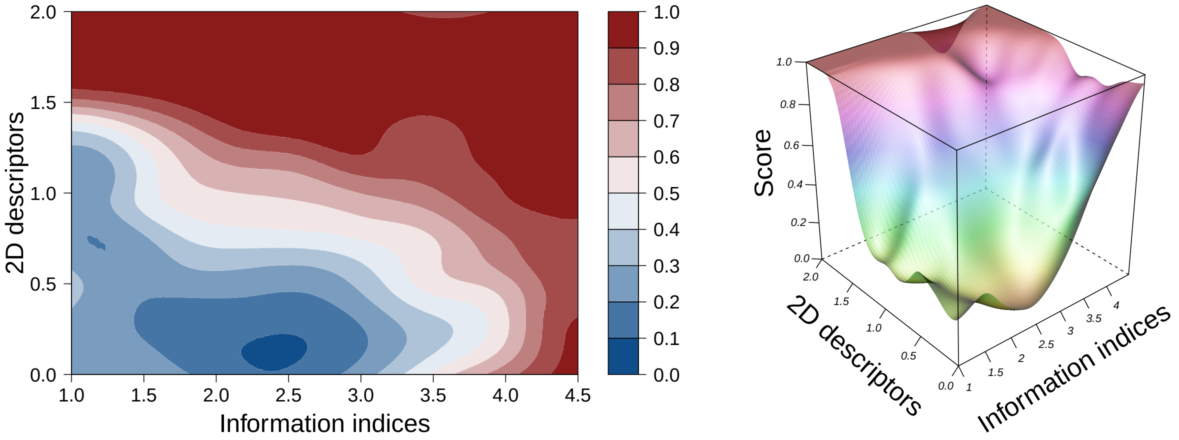


Figure 3: At left, contour plot for $\hat{f}(\cdot)/(\hat{f}(\cdot) + \hat{g}(\cdot))$ where $\hat{f}(\cdot)$ and $\hat{g}(\cdot)$ are the smooth kernel density function estimations for the positive and negative populations, respectively. At right, the function $\hat{f}(\cdot)/(\hat{f}(\cdot) + \hat{g}(\cdot)) : [1, 4.5] \times [0, 2] \longrightarrow [0, 1]$. In both estimations we used the biased cross-validation (BCV) procedure for computing the bandwidth matrices.

variability by using a standard bootstrap procedure (5,000 iterations), the 95% confidence intervals are the same in both the optimal transformation and the logistic regression-based models. Results are also similar when different criteria for computing the bandwidth are used. When we include a k-fold cross-validation algorithm for controlling the potential overfitting, the two-fold procedure ($k = 2$) reports an average (based on 500 iterations) AUC of 0.81 (0.80-0.82). The logistic regression-based procedure gets similar AUCs, around 0.79 (0.78-0.80).

Figure 3 represents the contour plot (left) and the 3D function (right) for the function $\hat{f}(\cdot)/(\hat{f}(\cdot) + \hat{g}(\cdot))$ with $\hat{f}(\cdot)$ and $\hat{g}(\cdot)$ being the multivariate kernel density estimates for the positive and the negative population, respectively, obtained by considering the BCV bandwidth. Both plots suggest that higher values of *2D matrix-based descriptors* are associated with a higher likelihood of being a positive subjects, but the role of *Information indices* changes with the value of *2D matrix-based descriptors*. Figure 4 shows the average effect of *Information indices* at different levels of *2D matrix-based descriptors* (particularly at its quartiles) determined by both the proposed estimator and the linear logistic regression. Its impact on the score $\hat{f}(\cdot)/(\hat{f}(\cdot) + \hat{g}(\cdot))$ is scarce for higher values of *2D matrix-based descriptors*, where the values of the score are already high, but clearly relevant for the rest.

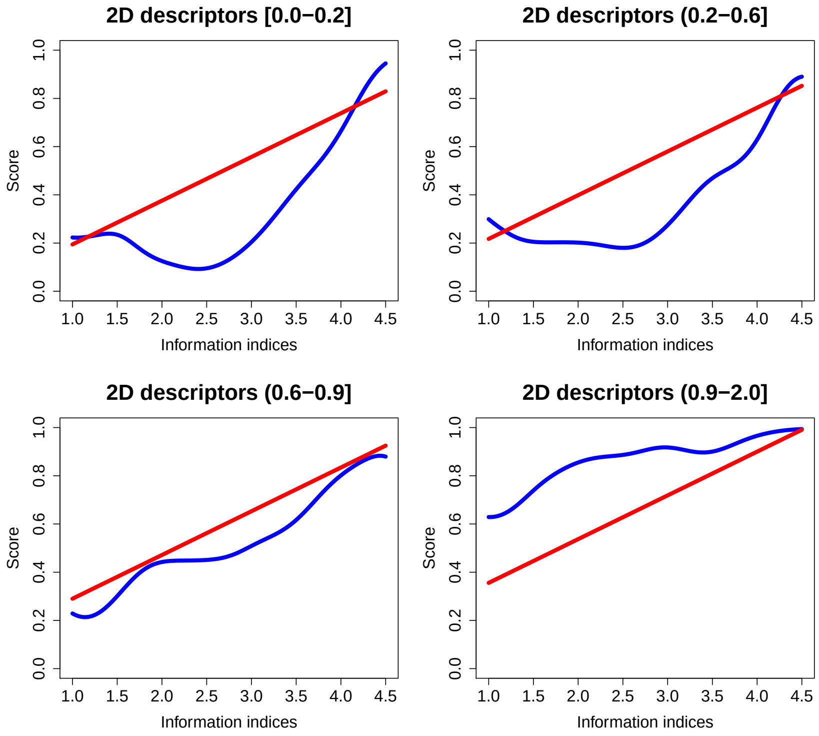


Figure 4: Average effect of *Information indices* (at different levels of *2D matrix-based descriptors*) on the score $\hat{f}(\cdot)/(\hat{f}(\cdot)+\hat{g}(\cdot))$ (in blue) and on the punctuation derived from the logistic regression (red).

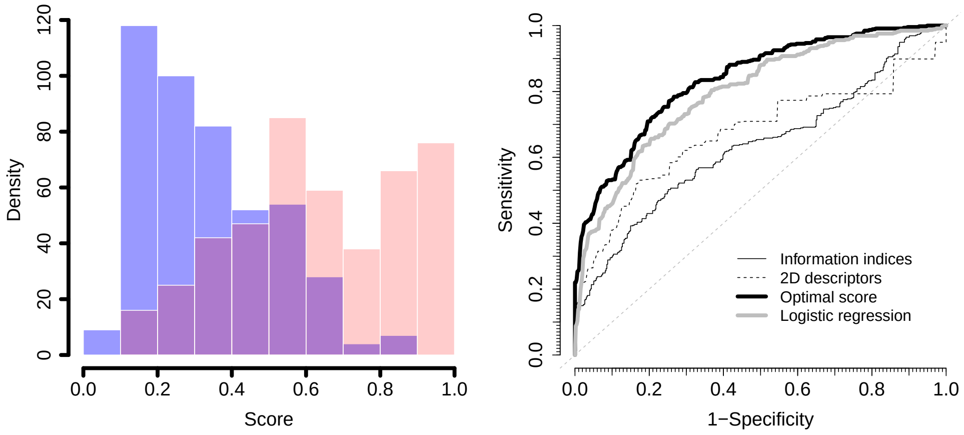


Figure 5: At left, histograms for the score, defined as $\hat{f}(\cdot)/(\hat{f}(\cdot)+\hat{g}(\cdot))$, distribution in both the negative (blue) and the positive (red) groups. At right, ROC curves for the individual markers (thin lines) and for the bivariate models based on $\hat{T}_N(\cdot)$ (black-thick line) and on standard logistic regression (gray-thick line).

This trend is confirmed in the left panel of Figure 3, where for higher levels of the second component (*2D matrix-based descriptors*), the subjects are mostly classified as positive, whereas for lower levels, the decision rule clearly depends on the value of the first component (*Information indices*). This reflects the importance of the visualization in having a clear knowledge of the problem under study, highlighting the classification rules over the domain of the bivariate marker; that is, studying both markers simultaneously and thus considering the associations between them. Notice that, in spite of the fact that the logistic regression reaches similar overall classification capacity, the information derived from this procedure does not lead to the same conclusions and does not allow us to understand the interaction between the different components of the marker neither the different impact of these values on the likelihood of being in the positive group. In this case, the linearity restriction leads to similar overall diagnostic capacity but it misunderstands the relationship between the outcome and the marker.

5 Discussion

Making binary classifications based on indirect information (other than the gold standard) implies the definition of binary decision rules. When the indirect information is given in terms of continuous measures, there is *a continuous number* of those decision rules. The receiver-operating characteristic, ROC, curve graphically represents the classification capacity of the underlying decision rules in terms of their sensitivity and specificity. Standard univariate ROC curves assume that the classification subsets (those classifying a subject in the positive group) are intervals of the form (c, ∞) with $c \in \mathbb{R}$. In more general cases, the usual practice is to look for an appealing transformation of the marker which improves its discrimination ability. In this respect, Martínez-Camblor et al. [32] compared the overall classification performance of the so-called gROC curve [28, 30], which keeps the interpretability of the underlying classification rules versus a quite free transformation of the marker. Results indicated that, in terms of overall diagnostic accuracy measured through the area under the curve, the marker free transformation was only at a small advantage over the gROC curve in the considered case. With the same philosophy, for multivariate markers, a usual practice is to look for an adequate or simple transformation to get proper univariate decision rules. The optimal linear transformation have been deeply considered in the specialized literature [23, 45, 36].

While in some practical problems it is of interest to keep some rationality (simplicity) behind those decisions, when we consider multivariate markers with potentially complex relationship among their components, the focus is frequently to develop a valuable univariate score and to obtain information about the nature of the relationship between the marker and the outcome. McIntosh and Pepe [33] found the theoretical multivariate transformation leading to the optimal classification rules; in this paper, we have proposed making use of smooth techniques to estimate this transformation. We have outlined the large sample properties of both the transformation and the resulting ROC curve estimates when the plug-in method and the multivariate kernel density estimator are employed.

Selecting the smoothing parameter or bandwidth is one of the main handicaps against the use of kernel techniques. Monte Carlo simulations results suggest that the estimation is quite stable when we consider a reasonable (automatic) criterion to estimate the optimal bandwidth. In our simulations, we have considered four different criteria based on minimizing the mean integrated squared error (MISE) in the estimation of the density functions. Unfortunately, it is well-known that the optimal bandwidth strongly depends on the particular problem we are dealing with [29] and that there is not an easy or general solution for this issue. Looking for an optimal bandwidth in the current context would involve complex theoretical developments which are far from the goals of this paper. However, simulation results suggest that the outcomes obtained through any of the proposed automatic bandwidth selections are similar and close to the underlying real model. A second handicap against the use of the estimator $\hat{T}_N(\cdot)$ is the presence of a density estimate in the denominator. This could be a source of instability when the value of the real density is close to zero [25]. There is not an easy solution for that but to operate with the equivalent transformation $f(\cdot)/(f(\cdot) + g(\cdot))$ dilutes the problem. The third relevant issue is the potential overfitting. In this respect, both the Monte Carlo simulations (see Table S1 and S2 in the online supplementary material) and the real-world application suggest that, with enough sample size and moderate-good classification performances, the overfitting problem is not too serious. Nevertheless, it is worth to remark that this kind of procedures provide really over-optimistic results when the associations are poor and/or the sample size is small. It is not new that using flexible data analysis techniques is a risk when the sample size is not large enough and that, in those cases, introducing some parametric restrictions is advisable.

That is, smooth techniques detect, at least, weak classification capacities, even when there is not any (see Copas and Corbett [7] for an overview of this problem in logistic regression). The use of machine learning techniques such as cross-validation or bootstrapping may be integrated to improve the results obtained (see, for instance, Martínez-Camblor and Pardo-Fernández [31]), achieving adequate results even for small sample sizes (see Table S2). Remark that, in our Monte Carlo study, these results under-estimate the real discrimination ability.

In the considered real-world problem, the behavior of the two variables in the negative and the positive populations differs in both mean and shape. The area under the ROC curves based on the individual markers suggest a moderate classification capacity. When we combine both markers in a simple logistic regression model, the AUC is almost 0.80 (0.79 after the 2-fold cross-validation correction). Optimal decision rules improve this number a little bit (AUC of 0.81 after applying the 2-fold cross-validation procedure) but, perhaps, the most relevant discovery is the interaction between the two variables in order to get the optimal classification based on those markers. *Information indices* seems to have a small effect on toxicity for moderate-high values of *2D matrix-based descriptors*, but it plays a relevant role when the latter take small values. These relationships are missing when we perform simple linear logistic regression. The differences produced by using different bandwidths criteria and the impact of the transformation estimation on the variability of the overall discrimination capacity are negligible.

In summary, the proposed procedure, which includes the estimation of two multivariate density functions by using bandwidths computed through an automatic selection criterion, reports good results when the sample sizes are high enough and the classification criteria accuracy obtained is moderate-high. Otherwise, we should double-check for potential overfitting and, in this case, the inclusion of a cross-validation procedure is advisable. With certainty, different and complex simulations can help to have a better knowledge of the estimator behavior. As supplementary material, we provide all the codes implemented for doing these simulations.

Supplementary Material

As supplementary material of this paper we provide the `R` code implemented to compute plots and models reported herein. The main provided function, `optimalT`, incorporates a general k-fold cross-validation procedure to control the potential overfitting. `R` packages `nsROC` (developed by Pérez-Fernández et al. [35]) and `ks` (developed by Duong [15]) are required. The used dataset is freely available at `http://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity#`. Results from additional simulations are provided in Tables S1, S2 and S3.

Acknowledgment

This work is supported by the Grants MTM2015-63971-P and MTM2014-55966-P from the Spanish Ministerio of Economía y Competitividad; FC-GRUPIN-IDI/2018/000132 and Severo Ochoa Grant BP16118 (this one for S. Pérez-Fernández) from the Asturias Government (Gobierno del Principado de Asturias).

References

[1] Bobb, J. F., L. Valeri, H. B. Claus, D. C. Christiani, R. O. Wright, M. Mazumdar, J. J. Godleski, and B. A. Coull (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics 16*(3), 493–508.

[2] Bowman, A. W. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Oxford: Oxford Science Publications.

[3] Breiman, L. (2017). *Classification and Regression Trees.* CRC Press.

[4] Cassotti, M., D. Ballabio, R. Todeschini, and V. Consonni (2015). A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (pimephales promelas). *SAR and QSAR in Environmental Research 26*(3), 217–243.

[5] Chen, B., P. Li, J. Qin, and T. Yu (2016). Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests. *Journal of the American Statistical Association 111*(514), 861–874.

[6] Collins, M. (2002). Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 1–8. Association for Computational Linguistics.

[7] Copas, J. B. and P. Corbett (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika 89*(2), 315–331.

[8] Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning 20*(3), 273–297.

[9] de Gonzalo-Calvo, D., D. Vilades, P. Martínez-Camblor, A. Vea, A. Ferrero-Gregori, L. Nasarre, O. Bornachea, J. Sánchez Vega, R. Leta, N. Puig, S. Benítez, J. L. Sánchez-Quesada, F. Carreras, and V. Llorente-Cortés (2019). Plasma microRNA profiling reveals novel biomarkers of epicardial adipose tissue: A multidetector computed tomography study. *Journal of Clinical Medicine 8*(6).

[10] de Gonzalo-Calvo, D., D. Vilades, P. Martínez-Camblor, A. Vea, L. Nasarre, J. Sánchez Vega, R. Leta, F. Carreras, and V. Llorente-Cortés (2019). Circulating microRNAs in suspected stable coronary artery disease: A coronary computed tomography angiography study. *Journal of Internal Medicine 286*(3), 341–355.

[11] Devroye, L. and C. Penrod (1986). The strong uniform convergence of multivariate variable kernel estimates. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique 14*(3), 211–219.

[12] Díaz-Coto, S., N. Corral-Blanco, and P. Martínez-Camblor (2020). Two-stage receiver operating-characteristic curve estimator for cohort studies. *The International Journal of Biostatistics* (0), 1–22.

[13] Duong, T. (2004). Bandwidth matrices for multivariate kernel density estimation. *Ph.D. Thesis, University of Western Australia.*

[14] Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software 21*(7), 1–16.

[15] Duong, T. (2019). *ks: Kernel Smoothing.* R package version 1.11.5.

[16] Freund, Y. and R. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computational Systems Science 55*(1), 119–139.

[17] Green, D. and J. Swets (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

[18] Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis 14*(1), 1–16.

[19] Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*, 29–36.

[20] Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

[21] Hsieh, F. and B. W. Turnbull (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics 24*(1), 25–40.

[22] Huang, X., G. Qin, and Y. Fang (2011). Optimal combinations of diagnostic tests based on AUC. *Biometrics 67*(2), 568–576.

[23] Kang, L., C. Xiong, P. Crane, and L. Tian (2013). Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories. *Statistics in Medicine 32*(4), 631–643.

[24] Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, Articles 28*(5), 1–26.

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

[26] Martínez-Camblor, P., C. Carleos, and N. Corral (2011). Powerful nonparametric statistics to compare k independent ROC curves. *Journal of Applied Statistics 38*(7), 1317–1332.

[27] Martínez-Camblor, P., C. Carleos, and N. Corral (2013). General nonparametric ROC curve comparison. *Journal of the Korean Statistical Society 42*(1), 71 – 81.

[28] Martínez-Camblor, P., N. Corral, C. Rey, J. Pascual, and E. Cernuda-Morollón (2017). Receiver operating characteristic curve generalization for non-monotone relationships. *Statistical Methods in Medical Research 26*(1), 113–123.

[29] Martínez-Camblor, P. and J. de Uña-Álvarez (2013). Studying the bandwidth in k-sample smooth tests. *Computational Statistics 28*(2), 875–892.

[30] Martínez-Camblor, P. and J. Pardo-Fernández (2019a). Parametric estimates for the receiver operating characteristic curve generalization for non-monotone relationships. *Statistical Methods in Medical Research 28*(7), 2032–2048.

[31] Martínez-Camblor, P. and J. Pardo-Fernández (2019b). The Youden index in the generalized receiver operating characteristic curve context. *The International Journal of Biostatistics 15*(1), 1–28.

[32] Martínez-Camblor, P., S. Pérez-Fernández, and S. Díaz-Coto (2019). Improving the biomarker diagnostic capacity via functional transformations. *Journal of Applied Statistics 46*(9), 1550–1566.

[33] McIntosh, M. W. and M. S. Pepe (2002). Combining several screening tests: Optimality of the risk score. *Biometrics 58*(3), 657–664.

[34] Pepe, M. S. and M. L. Thompson (2000). Combining diagnostic test results to increase accuracy. *Biostatistics 1*(2), 123–140.

[35] Pérez-Fernández, S., P. Martínez-Camblor, P. Filzmoser, and N. Corral (2018). *nsROC: An R package for Non-Standard ROC Curve Analysis.*

[36] Pérez-Fernández, S., P. Martínez-Camblor, P. Filzmoser, and N. Corral (2020). Visualizing the decision rules behind the ROC curves: understanding the classification process. *AStA Advances in Statistical Analysis (In press).*

[37] Qin, J. and B. Zhang (2010). Best combination of multiple diagnostic tests for screening purposes. *Statistics in Medicine 29*(28), 2905–2919.

[38] Regulation (EC) (2006). No 1907/2006. pp. 1–849.

[39] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics 27*(3), 832–837.

[40] Scott, C. and R. Nowak (2005). A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory 51*(11), 3806–3819.

[41] Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley Series in Probability and Statistics. Wiley.

[42] Signes-Pastor, A. J., B. T. Doherty, M. E. Romano, K. M. Gleason, J. Gui, E. Baker, and M. R. Karagas (2019). Prenatal exposure to metal mixture and sex-specific birth outcomes in the New Hampshire birth cohort study. *Environmental Epidemiology 3*(5), 1–8.

[43] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics & Applied Probability. London: Chapman & Hall.

[44] Su, J. and J. Liu (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association 88*(424), 1350–1355.

[45] Yan, Q., L. E. Bantis, J. L. Stanford, and Z. Feng (2018). Combining multiple biomarkers linearly to maximize the partial area under the ROC curve. *Statistics in Medicine 37*(4), 627–642.