Bootstrap-based procedures for inference in nonparametric receiver-operating characteristic curve regression analysis*

María Xosé Rodríguez-Álvarez[1,2,3,*], Javier Roca-Pardiñas [1], Carmen Cadarso-Suárez[4] and Pablo G. Tahoces [5]

[1] Department of Statistics and Operations Research and Biomedical Research Centre (CINBIO), University of Vigo, Vigo, Spain
[2] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[3] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[4] Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), University of Santiago de Compostela, Santiago de Compostela, Spain
[5] Centro Singular de Investigación en Tecnologías de la Información (CITIUS), University of Santiago de Compostela, Santiago de Compostela, Spain
* Corresponding author: María Xosé Rodríguez-Álvarez, mxrodriguez@bcamath.org

Abstract

Prior to using a diagnostic test in a routine clinical setting, the rigorous evaluation of its diagnostic accuracy is essential. The receiver operating characteristic (ROC) curve is the measure of accuracy most widely used for continuous diagnostic tests. However, the possible impact of extra information about the patient (or even the environment) on diagnostic accuracy also needs to be assessed. In this paper, we focus on an estimator for the covariate-specific ROC curve based on direct regression modelling and nonparametric smoothing techniques. This approach defines the class of generalised additive models for the ROC curve (ROC-GAM). The main aim of the paper is to offer new inferential procedures for testing the effect of covariates on the conditional ROC curve within the ROC-GAM context. Specifically, two different bootstrap-based tests are suggested to check (a) the possible effect of continuous covariates on the ROC curve; and (b) the presence of factor-by-curve interaction terms. The validity of the proposed bootstrap-based procedures is supported by simulations. To facilitate the application of these new procedures in practice, an `R`-package, known as `npROCRegression`, is provided and briefly described. Finally, data derived from a computer-aided diagnostic (CAD) system for the automatic detection of tumour masses in breast cancer is analysed.

*This paper has been published in *Statistical Methods in Medical Research*. DOI: 10.1177/0962280217742542

Keywords: ROC curve; generalised additive models; bootstrap; computer-aided diagnosis.

1 Introduction

In many biometrical applications, the classification of individuals or observations based on covariate information is one of the most important goals of a statistical analysis. For example, the classification of patients as healthy or diseased (to consider only the most simple classification task) on the basis of demographic information and individual disease history would be the point of departure for subsequent treatment. In this context, a diagnostic test can be any diagnostic procedure conducted to differentiate between different types of patients, e.g. healthy versus diseased, or patients in different stages of disease progression. However, classification of an individual's status based on the result of a diagnostic test is usually not error-free and some individuals will be misclassified. Accordingly, before the routine application of a diagnostic test in clinical practice, any errors of classification must be quantified in order to check a diagnostic test's validity or invalidity, i.e. diagnostic accuracy, or ability to discriminate between alternative health states must be measured.

In the case of binary or dichotomous tests, diagnostic accuracy is often summarised by means of the true positive fraction (TPF) and the false positive fraction (FPF). Let Y denote the result of the diagnostic test ($Y = 1$ for diseased and $Y = 0$ for healthy), and D the dummy variable that indicates the true disease status ($D = 1$ for presence and $D = 0$ for absence of the disease). The TPF or 'sensitivity', then, is the probability of an individual being correctly classified as diseased, i.e. $TPF = P(Y = 1 \mid D = 1)$, whereas the FPF, or '1−specificity', is the probability of a healthy individual being falsely classified as diseased, i.e. $FPF = P(Y = 1 \mid D = 0)$.

For tests with continuous or ordinal results, the most widely used measure of diagnostic accuracy is the receiver operating characteristic (ROC) curve (Krzanowski and Hand, 2009; Pepe, 2003; Zou et al., 2002). The ROC curve extends the concepts of sensitivity and specificity to the continuous/ordinal case by depicting these quantities for all possible cut-off values or decision thresholds c applied to the test result. In other words, the ROC curve relies on all possible transformations of a continuous/ordinal test to a binary test. More specifically, the ROC curve is defined as the set of all TPF-FPF pairs $\{(TPF(c), FPF(c)), c \in (-\infty, \infty)\}$ that can be obtained by varying the cut-off value c, where $TPF(c) = P(Y \geq c \mid D = 1)$ and $FPF(c) = P(Y \geq c \mid D = 0)$. When the diagnostic test Y is continuous, the ROC curve is usually represented as $ROC(p) = S_D\left(S_{\bar{D}}^{-1}(p)\right)$ for $0 \leq p \leq 1$, where $S_D(c) = P(Y \geq c \mid D = 1)$ and $S_{\bar{D}}(c) = P(Y \geq c \mid D = 0)$.

It is well known, however, that in many situations the discriminatory capacity or accuracy of a diagnostic test can be affected by covariates (see Pepe, 2003, pp 48-49, for examples). In such cases, failure to incorporate information furnished by covariates in the ROC analysis may lead to erroneous conclusions (Janes and Pepe, 2008, 2009; Pardo-

Fernández et al., 2014). Denoting as $\boldsymbol{X}$ the d-variate vector of covariates we are interested in, the conditional or covariate-specific ROC curve, given a covariate value $\boldsymbol{x}$, is defined as

$$ROC_{\boldsymbol{x}}(p) = S_D\left(S_{\bar{D}}^{-1}(p \mid \boldsymbol{x}) \mid \boldsymbol{x}\right),\ 0 \leq p \leq 1, \tag{1}$$

where, by a slight abuse of notation, $S_D(c \mid \boldsymbol{x}) = P(Y \geq c \mid D = 1, \boldsymbol{X} = \boldsymbol{x})$ and $S_{\bar{D}}(c \mid \boldsymbol{x}) = P(Y \geq c \mid D = 0, \boldsymbol{X} = \boldsymbol{x})$. Note that a continuum of different ROC curves (and therefore, a continuum of different diagnostic accuracies) is obtained by varying the value $\boldsymbol{x}$ in the range of $\boldsymbol{X}$. As a consequence, the conditional ROC curve can be viewed as a tool which helps to identify those patients' strata (or subpopulations) that may benefit from the application of the diagnostic test, as well as those for which the test does not provide valuable information.

Estimation of the conditional ROC curve has been explored in the statistical literature from (semi) parametric and nonparametric perspectives, and within frequentist and Bayesian paradigms. A detailed review and comparison of (semi) parametric frequentist proposals can be found in Rodríguez-Álvarez et al. (2011), whereas the paper by Pardo-Fernández et al. (2014) mainly focuses on the nonparametric counterparts. In brief, there are two main strategies for approaching estimation: one based on estimating the conditional cumulative survival functions involved in the definition given in (1) (López-de-Ullibarri et al., 2008; Inácio de Carvalho et al., 2013); and the other based on modelling (and estimating) the effect of covariates on the ROC curve through regression models. In the latter case, the literature on ROC regression techniques has led to two different methodologies: 'induced' and 'direct' (Pepe, 2003). Induced methodology is based on inducing the expression of the conditional ROC curve through regressing the diagnostic test on the available covariates separately in healthy and diseased individuals (Faraggi, 2003; González-Manteiga et al., 2011; Pepe, 1998; Rodríguez-Álvarez et al., 2011b; Rodríguez and Martínez, 2013; Zheng and Heagerty, 2004; Yao et al., 2010). On the other hand, direct methodology directly regresses the ROC curve. This methodology has yielded (1) the general class of ROC-GLM regression models (Alonzo and Pepe, 2002; Cai, 2004; Cai and Pepe, 2002; Pepe and Cai, 2004), due to its similarity to generalised linear models (GLM, McCullagh and Nelder, 1989); and (2) its extension to a more flexible regression setting, the ROC-GAM class (Rodríguez-Álvarez et al., 2011a), along the line of the generalised additive model (GAM, Hastie and Tibshirani, 1990).

The aim of this paper is twofold. Firstly, as in any regression context, in the ROC regression framework it is important to have formal procedures for testing model assumptions and/or effects of covariates. In spite of its importance, to the best of our knowledge this topic has received little attention in the statistical literature, especially in the nonparametric framework. Interesting contributions to the topic can be found in the paper by Cai and Zheng (2007), where several model-checking procedures for (semi) parametric approaches are presented, and in the paper by Rodríguez-Álvarez et al. (2011a), where a bootstrap-based test to check for the effect of a continuous covariate is proposed. This paper thus

focuses on presenting new inferential procedures for testing the effect of covariates over the conditional ROC curve. Specifically, we present two different bootstrap-based tests to check (a) the possible effect of continuous covariates on the ROC curve; and (b) the presence of factor-by-curve interaction terms. Both tests are proposed within the ROC-GAM context.

Secondly, nowadays there is an undeniable need for software development of new statistical methods. In fact, the implementation of new methodological advances in user-friendly software has dramatically increased in the last few years. This tendency has had an important impact on shortening the time from the development of new statistical advances to their application. Therefore, accompanying this paper we provide a free `R`-package (R Core Team, 2017) called, `npROCRegression`. The package allows for the practical application of several nonparametric approaches to the inclusion of covariates in the ROC curve. More precisely, `npROCRegression` implements the nonparametric induced and direct proposals as presented in Rodríguez-Álvarez et al. (2011b,a), as well as the inferential procedures described in this paper. The package is freely available from CRAN at `https://cran.r-project.org/package=npROCRegression`. We hope that the existence of easy-to-use software will encourage the use of these techniques in clinical research.

The remainder of the paper is structured as follows: Section 2 briefly discusses the statistical literature on the inclusion of covariate information in the ROC regression framework. Section 3 presents in more detail the class of ROC-GAM regression models, and in Section 4 the proposed bootstrap-based procedures are introduced. The performance of these procedures have been evaluated by means of simulations, and results are presented in Section 5. Additional results have been added as online Supplementary Material. Section 6 describes the `npROCRegression` `R`-package. We illustrate our approach and the usage of the package in Section 7 using data from a computer-aided diagnostic (CAD) system. The Discussion closes the paper. Some technical details are made available in an Appendix.

2 Modelling covariate effects on the ROC curve

This section reviews the literature on ROC curves in the presence of covariate information. It is beyond the scope of this paper to present an exhaustive review, and we refer the readers to Pepe (2003), Rodríguez-Álvarez et al. (2011) and Pardo-Fernández et al. (2014) for a more in-depth survey. However, with this section we aim to put different modelling strategies for the incorporation of covariates on the ROC curve into context. More precisely, we focus here on those that are considered to be within the general framework of regression, namely 'induced' and 'direct' methodologies. A qualitative comparison of both approaches is also presented. The section ends with a presentation of several summary statistics of the conditional ROC curve.

2.1 ROC regression approaches

2.1.1 Induced ROC regression methodology.

This approach is based on firstly modelling the effect of covariates on the diagnostic test, and then compounding the conditional ROC curve. In its most general specification, a location-scale regression model is assumed for the classification variable Y in each population separately

$$Y_{\bar{D}} = (Y \mid D = 0) = \mu_{\bar{D}}(\boldsymbol{X}) + \sigma_{\bar{D}}(\boldsymbol{X})\varepsilon_{\bar{D}},$$
$$Y_D = (Y \mid D = 1) = \mu_D(\boldsymbol{X}) + \sigma_D(\boldsymbol{X})\varepsilon_D,$$

where, for $j \in \{\bar{D}, D\}$, $\mu_j(\boldsymbol{x}) = E(Y_j|\boldsymbol{X} = \boldsymbol{x})$ and $\sigma_j^2(\boldsymbol{x}) = Var(Y_j|\boldsymbol{X} = \boldsymbol{x})$ are the conditional mean and the conditional variance of Y_j given $\boldsymbol{X} = \boldsymbol{x}$, respectively. The error ε_j is assumed independent of the covariate $\boldsymbol{X}$, with zero mean, unit variance and cumulative survival function G_j, i.e., $G_j(c) = P(\varepsilon_j \geq c)$. With this configuration, and given the independence of the errors and the covariates, it follows that

$$ROC_{\boldsymbol{x}}(p) = G_D\left(\frac{\mu_{\bar{D}}(\boldsymbol{x}) - \mu_D(\boldsymbol{x})}{\sigma_D(\boldsymbol{x})} + \frac{\sigma_{\bar{D}}(\boldsymbol{x})}{\sigma_D(\boldsymbol{x})} G_{\bar{D}}^{-1}(p)\right).$$

Note that, under this approach, the effect of the covariates on the ROC curve is expressed in terms of their effects on the mean and variance of the diagnostic test in healthy and diseased subjects.

In a parametric or semiparametric framework, important references for the estimation of induced methodology include Pepe (1998), Faraggi (2003) and Zheng and Heagerty (2004). All these papers propose modelling the covariate effects on the result of the diagnostic test parametrically. Nonparametric specifications of the conditional means and variances have been considered in Yao et al. (2010), González-Manteiga et al. (2011), Rodríguez-Álvarez et al. (2011b) and Rodríguez and Martínez (2013). The first three papers propose fully nonparametric estimators based on kernel-type regression techniques (Fan and Gijbels, 1996). We should note that these proposals are restricted to one-dimensional covariates. The proposal by Rodríguez and Martínez, framed in a Bayesian setting, allows for incorporating multidimensional continuous covariates, but the authors assume that the error terms are distributed according to a Student's t distribution.

2.1.2 Direct ROC regression methodology.

In contrast to the induced method, direct methodology *directly* models the effects of covariates on the ROC curve. In this approach, the general form of the conditional ROC curve is given by the following regression model

$$ROC_{\boldsymbol{x}}(p) = g(\mu(\boldsymbol{x}) + h_0(p)), \tag{2}$$

where function μ collects the effects of the covariates on the ROC curve, h_0 is a monotonically increasing baseline function of the FPF, p (responsible for modelling the shape of the ROC curve), and g is the function linking the covariates and FPF with the conditional TPFs (i.e., the ROC curve).

In the (semi) parametric framework, different proposals have been suggested in the literature, which mainly differ in the assumptions made about the function of the FPF, $h_0(\cdot)$ (see Alonzo and Pepe, 2002; Cai, 2004; Cai and Pepe, 2002; Pepe and Cai, 2004). In all these approaches the effect of covariates $\boldsymbol{X}$ on the ROC curve is incorporated parametrically, i.e., $\mu(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$. Thus, models such as (2) define the so-called class of ROC-GLMs (Pepe, 2003), due to their resemblance to generalised linear models (GLMs). To the extent of our knowledge, to date, only the paper by Rodríguez-Álvarez et al. (2011a) addresses the inclusion of covariate information on direct modelling nonparametrically. In that paper the authors propose to extend the class of ROC-GLM regression models by assuming an ROC-GAM regression model for the ROC curve. In this case, rather than assuming a parametric form for the effect of the continuous covariates, it is solely assumed that these effects can be represented by arbitrary smooth functions. We discuss this approach in more detail in Section 3.

2.1.3 Qualitative comparison.

As mentioned, there are two different regression methodologies for the incorporation of covariates into the ROC analysis. From an applied point of view, the natural question arising is: which one should be used in practice? Unfortunately, there is no simple answer. Both methodologies present appealing features and may provide valuable information. We would therefore suggest using both, whenever it is possible.

Regarding induced methodology, its main advantage is that it models covariate effects on the result of the diagnostic test. Even though, on the one hand, it means that the modelling of covariate effects on the ROC curve is indirect, on the other hand (a) it allows for the use of more "standard" regression techniques and model checking procedures than direct methodology, especially in the parametric framework; and (b) it relates the effect of the covariates on the mean and variance of the diagnostic test to their effect on the ROC curve, which can help to understand and explain the covariate impact on the accuracy of the test. For ease of reading, throughout all of our presentation, covariates affecting the test results in healthy and diseased populations are assumed to be the same, although this is not necessarily the case in practice (e.g., disease stage). Induced methodology allows for the incorporation of specific covariates of healthy or diseased populations, or even both. Finally, by modelling covariate effects on the variance of the diagnostic test, the shape of the ROC curve is allowed to vary with the covariates. In regression terminology, this would be equivalent to including the interaction between the covariates and the FPF, p. However, in the case that the covariate vector $\boldsymbol{X}$ is multidimensional (and no restrictions are imposed), heteroskedasticity would also mean that the interaction among all covariates

(and possibly the FPF) is implicitly incorporated into the "model" for the conditional ROC curve. This can make the interpretation and visualisation of results considerably difficult, especially in the presence of several continuous covariates. Moreover, testing for covariate effects on the ROC curve would become a complex task.

As far as direct methodology is concerned, its obvious advantage is that it directly evaluates covariate effect on the measure of interest, the ROC curve. As a consequence, it enables the accuracy of different diagnostic tests to be compared (Pepe, 2003). In addition, inclusion of multidimensional covariates is straightforward, and interactions among covariates can be accommodated in a more natural way than through induced methodology. Moreover, although it has not been considered here, it is possible to incorporate the interaction between covariates and the FPF. However, to the best of our knowledge, none of the approaches presented in the statistical literature ensures that the resulting interaction estimates are monotonic in the FPF direction as required by theoretical properties of the ROC curve. This is undoubtedly an interesting topic of research. Finally, direct methodology also allows the incorporation of disease-specific covariates. It does not, however, permit health-related information.

2.2 Conditional summary statistics

It is common to summarise the information of the ROC curve by means of single indexes. We list here those which are most commonly used, and present a summary measure that is meaningful in the conditional case.

2.2.1 Area under the conditional ROC curve.

The area under the ROC curve (AUC) is possibly the most widely used summary measure of discriminatory performance. In the conditional case, the AUC is defined as

$$AUC_{\boldsymbol{x}} = \int_0^1 ROC_{\boldsymbol{x}}(p)\, dp. \tag{3}$$

The $AUC_{\boldsymbol{x}}$ ranges from 0.5 to 1, taking the value of 0.5 in the case of an uninformative test and 1 in a perfect test.

The most obvious way to estimate the conditional AUC is to simply plug-in an estimate for the conditional ROC curve in (3), and approximate the integral using numerical integration methods. However, this approach might not be the most efficient way, and several methods to *directly* estimate $AUC_{\boldsymbol{x}}$ have been proposed in the literature. For instance, Faraggi (2003) discusses a fully parametric estimation approach based on induced modelling. In Dodd and Pepe (2003b,a), and Cai and Dodd (2008), a semiparametric regression model for the conditional (partial) AUC is proposed, similar in spirit to the direct ROC regression methodology. In a fully nonparametric setting, Yao et al. (2010) present a "conditional" Mann-Whitney estimator for $AUC_{\boldsymbol{x}}$ estimation. This method has been

recently generalised to functional covariates by Inácio de Carvalho et al. (2016). In that paper the authors propose a functional conditional partial area under the specificity-ROC curve. The estimator by Yao et al. (2010) is a particular case when the sensitivity is not restricted to a specific interval.

2.2.2 Conditional Youden index.

Another common summary index is the Youden index (Youden, 1950), which, in the conditional case, can be defined as

$$\begin{aligned} YI_{\boldsymbol{x}} &= \max_{c_{\boldsymbol{x}}} \left\{ TPF\left(c_{\boldsymbol{x}} \mid \boldsymbol{x}\right) - FPF\left(c_{\boldsymbol{x}} \mid \boldsymbol{x}\right) \right\} \\ &= \max_{c_{\boldsymbol{x}}} \left\{ S_D\left(c_{\boldsymbol{x}} \mid \boldsymbol{x}\right) - S_{\bar{D}}\left(c_{\boldsymbol{x}} \mid \boldsymbol{x}\right) \right\} \qquad (4) \\ &= \max_{p_{\boldsymbol{x}}} \left\{ ROC_{\boldsymbol{x}}\left(p_{\boldsymbol{x}}\right) - p_{\boldsymbol{x}} \right\}, \qquad (5) \end{aligned}$$

where we use the notation $c_{\boldsymbol{x}}$ and $p_{\boldsymbol{x}}$ to emphasise that these values depend on covariate $\boldsymbol{x}$. The $YI_{\boldsymbol{x}}$ takes values between 0 and 1.0, in the case of an uninformative test and a perfect test, respectively. The value $c^*_{\boldsymbol{x}}$, which maximises (4), is frequently used in practice as a threshold value to separate diseased from healthy status (in those individuals with covariate value $\boldsymbol{x}$).

Parametric and nonparametric approaches to the estimation of the conditional Youden index (and associated threshold value) can be found in Faraggi (2003) and Xu et al. (2014), among others.

2.2.3 Covariate adjusted ROC curve.

All measures discussed above depict the accuracy of a diagnostic test for specific covariate values. However, it would be undoubtedly interesting to have a global summary that also takes covariate information into account. To that aim, Janes and Pepe (2009) propose the covariate-adjusted ROC curve (AROC), defined as

$$AROC\left(p\right) = \int ROC_{\boldsymbol{x}}\left(p\right) dH_D\left(\boldsymbol{x}\right), \qquad (6)$$

where $H_D\left(\boldsymbol{x}\right) = P\left(\boldsymbol{X} \leq \boldsymbol{x} \mid D = 1\right)$. Thus, the AROC curve is an average of conditional ROC curves, weighted according to the distribution of the covariates in the diseased population. It should be noted that when a diagnostic test's discriminatory capacity is not affected by covariates, this does not necessarily mean that the conditional ROC curve (which in this case is common to all covariate values) coincides with the ROC curve obtained when pooling the data without regard to the values of the covariates. It does coincide, however, with the AROC curve (see Janes and Pepe, 2009; Pardo-Fernández et al., 2014, for more details). Consequently, even in those situations where the accuracy of a test does not vary

with the covariates, inferences based on the pooled ROC curve might be biased, and thus meaningless. In such cases the AROC curve should be used instead.

To the best of our knowledge, estimation of the AROC curve has been only discussed in Janes and Pepe (2009), from both (semi) parametric and nonparametric perspectives, and in Rodríguez-Álvarez et al. (2011b), in the context of nonparametric induced modelling approaches.

3 The ROC-GAM regression model

As discussed before, the ROC-GAM regression model extends the ROC-GLM by allowing the incorporation of arbitrary nonparametric functions for (some) continuous covariates, along the line of the generalised additive model. Specifically, the ROC-GAM regression model is expressed as

$$ROC_{\boldsymbol{x}}(p) = g\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_u + \sum_{k=1}^{V} f_k(x_{vk}) + h_0(p)\right), \tag{7}$$

where $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$ denote subsets of the covariate vector $\boldsymbol{x}$. Here β_0 and $\boldsymbol{\beta}$ are unknown regression coefficients (modelling parametric effects of continuous covariates and, by a slight abuse of notation, categorical covariates) and f_k are unknown nonparametric functions of continuous covariates. Under this approach, h_0 is also assumed to be nonparametric (and unknown). For identifiability reasons (see Hastie and Tibshirani, 1990), a constant β_0 is introduced into the model, and it is required that $E(f_k(X_{kv})) = 0$ $(k = 1, \ldots, V)$ and $\int_0^1 h_0(p)\, dp = 0$.

In many situations the effect of a continuous covariate on the ROC curve may vary across groups defined by levels of a categorical covariate. A generalisation of the 'pure' ROC-GAM in (7) is the ROC-GAM with factor-by-curve interactions. Without loss of generality, let us assume that $\boldsymbol{X}$ is a two-dimensional covariate, with X_v being a continuous covariate, and X_u a factor with M levels $\{1, \ldots, M\}$. The factor-by-curve ROC-GAM takes the form

$$ROC_{\boldsymbol{x}}(p) = g\left(\beta_0 + \sum_{l=1}^{M} \beta_l I(x_u = l) + f(x_v) + \sum_{l=1}^{M} f^l(x_v)\, I(x_u = l) + h_0(p)\right), \tag{8}$$

where β_0 and $\{\beta_l\}_{l=1}^{M}$ are unknown regression coefficients, and h_0, f_1 and f^l are unknown nonparametric functions. $I(A)$ denotes the indicator function of event A. In much the same way as for model (7), the following conditions are required for identifiability

$$E(f(X_v)) = 0, \quad E\left(f^l(X_v)\right) = 0 \ (l = 1, \ldots, M) \quad \text{and} \quad \int_0^1 h_0(p)\, dp = 0,$$

jointly with

$$\sum_{l=1}^{M} \beta_l = 0 \quad \text{and} \quad \sum_{l=1}^{M} f^l\left(x_v\right) = 0.$$

Note that, given the previous constraints, model (8) has been parametrised so it is *hierarchical*. As a consequence, f is the smooth main effect of covariate X_v, and thus f^l $(l = 1, \ldots, M)$ represent deviations from that main effect for each level of X_u.

Appendix A presents the main steps of the estimation process of the ROC-GAMs (7) and (8), and we refer the interested reader to Rodríguez-Álvarez et al. (2011a) for a more detailed description. However, for a better understanding of the procedures to be presented in Section 4, we should note that the proposed algorithm requires the estimation of the conditional cumulative survival function of the diagnostic test in healthy subjects, $S_{\bar{D}}\left(\cdot \mid \boldsymbol{x}\right)$ (Step 2). For that purpose, Rodríguez-Álvarez et al. (2011a) suggest modelling the effect of covariates on $Y_{\bar{D}}$ by a nonparametric location-scale regression model

$$\begin{aligned} Y_{\bar{D}} &= \mu_{\bar{D}}(\boldsymbol{X}) + \sigma_{\bar{D}}(\boldsymbol{X})\varepsilon_{\bar{D}} \\ &= \beta_{\bar{D}0} + \beta_{\bar{D}}^T \boldsymbol{X}_u + \sum_{k=1}^{V} f_{\bar{D}k}\left(X_{kv}\right) + \exp\left(\alpha_{\bar{D}0} + \alpha_{\bar{D}}^T \boldsymbol{X}_u + \sum_{k=1}^{V} g_{\bar{D}k}\left(X_{kv}\right)\right)\varepsilon_{\bar{D}}. \end{aligned} \tag{9}$$

For ease of notation, we assume that the sets of covariates whose effects are to be modelled parametrically and nonparametrically are the same for the conditional mean, the conditional variance, and the conditional ROC curve. Obviously, this might not be necessarily so. In addition, factor-by-curve interaction terms can also be included. Note that under (9), it follows that

$$S_{\bar{D}}\left(c \mid \boldsymbol{x}\right) = G_{\bar{D}}\left(\frac{c - \mu_{\bar{D}}(\boldsymbol{x})}{\sigma_{\bar{D}}(\boldsymbol{x})}\right).$$

4 Testing for effects in ROC-GAM regression models

This section introduces the bootstrap-based procedures proposed to test for: (a) continuous covariate effect on the ROC-GAM regression model specified in (7); and (b) factor-by-curve interaction terms in model (8).

Specifically, for model (7) we focus on testing for the effect of those covariates modelled nonparametrically. Accordingly, for each continuous covariate X_{vr} in (7), we consider the null hypothesis

$$H_0^r : \ f_r(x_{vr}) = 0.$$

That is to say, the ROC curve, and therefore the accuracy of the test, is not affected by covariate X_{vr}.

For model (8) our interest is focused on the null hypothesis

$$H_0 : \ f^1\left(x_v\right) = \ldots = f^M\left(x_v\right) = 0,$$

namely, that the effect of continuous covariate X_v on the ROC curve does not depend on the levels of factor X_u.

In both cases we propose the use of various tests based on the estimates of the partial functions f_r, and on the estimates of the interaction curves f^l $(l = 1, \ldots, M)$.

From now on, let us assume that we have two independent samples of independently and identically distributed (i.i.d.) observations $(\boldsymbol{x}_1^{\bar{D}}, y_1^{\bar{D}}), \ldots, (\boldsymbol{x}_{n_{\bar{D}}}^{\bar{D}}, y_{n_{\bar{D}}}^{\bar{D}})$ from population $(\boldsymbol{X}_{\bar{D}}, Y_{\bar{D}})$, and $(\boldsymbol{x}_1^{D}, y_1^{D}), \ldots, (\boldsymbol{x}_{n_D}^{D}, y_{n_D}^{D})$ from population $(\boldsymbol{X}_D, Y_D)$.

4.1 Testing for continuous covariate effect

The test for the null hypothesis

$$H_0^r : ROC_{\boldsymbol{x}}(p) = g\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_u + \sum_{k=1}^{r-1} f_k(x_{vk}) + \sum_{k=r+1}^{V} f_k(x_{vk}) + h_0(p)\right), \tag{10}$$

versus the general hypothesis

$$H_1^r : ROC_{\boldsymbol{x}}(p) = g\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_u + \sum_{k=1}^{V} f_k(x_{vk}) + h_0(p)\right),$$

is based on the estimate $\hat{f}_r$. To that aim, L_1 and L_2 norms are considered, yielding the following test statistics

$$T_{||} = \sum_{j=1}^{n_D} \left|\hat{f}_r\left(x_{jvr}^D\right)\right| \quad \text{and} \quad T_2 = \sum_{j=1}^{n_D} \hat{f}_r\left(x_{jvr}^D\right)^2.$$

Note that the proposed statistics are measures of deviations from the estimated nonparametric function $\hat{f}_r$ to its mean $n_D^{-1}\sum_{j=1}^{n_D}\hat{f}_r\left(x_{jvr}^D\right)$ which is forced to be zero during estimation in order to avoid identifiability problems.

It must be remarked that, if the null hypothesis is verified, then T ($T_{||}$ or T_2) should be close to zero but will be positive. Thus, for an observed value of the test statistic, T^o, the null hypothesis in (10) is rejected if the p-value $P(T > T^o \mid H_0) < \alpha$, where α is a specified level of significance.

To approximate the distributions of the test statistics under the null hypothesis a general bootstrap procedure is proposed, which consists of the following steps:

Step 1. Estimate $\mu_{\bar{D}}(\cdot)$, $\sigma_{\bar{D}}(\cdot)$, and $G_{\bar{D}}(\cdot)$ in (9) from $\left\{\left(\boldsymbol{x}_i^{\bar{D}}, y_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$ as explained in Appendix A. Let $\widehat{\mu}_{\bar{D}}(\cdot)$, $\widehat{\sigma}_{\bar{D}}(\cdot)$, and $\widehat{G}_{\bar{D}}(\cdot)$ be these estimates.

Step 2. Estimate the null ROC-GAM regression model (10) from $\left\{\left(\boldsymbol{x}_j^{D}, y_j^{D}\right)\right\}_{j=1}^{n_D}$ as explained in Appendix A, and obtain the bootstrap pilot estimates $\widehat{ROC}^0_{\boldsymbol{x}_j^D}(p)$, $1 \le j \le n_D$.

For $b = 1, \ldots, B$

Step 3. Generate bootstrap resamples $\left\{\left(\boldsymbol{x}_i^{\bar{D}}, y_{i,b}^{\bar{D}*}\right)\right\}_{i=1}^{n_{\bar{D}}}$ and $\left\{\left(\boldsymbol{x}_j^{D}, y_{j,b}^{D*}\right)\right\}_{j=1}^{n_D}$ as follows

$$y_{i,b}^{\bar{D}*} = \widehat{\mu}_{\bar{D}}\left(\boldsymbol{x}_i^{\bar{D}}\right) + \widehat{\sigma}_{\bar{D}}\left(\boldsymbol{x}_i^{\bar{D}}\right)\varepsilon_{i,b}^{\bar{D}*}, \tag{11}$$

$$y_{j,b}^{D*} = \widehat{\mu}_{\bar{D}}\left(\boldsymbol{x}_j^{D}\right) + \widehat{\sigma}_{\bar{D}}\left(\boldsymbol{x}_j^{D}\right)\widehat{G}_{\bar{D}}^{-1}\left(\left(\widehat{ROC}_{\boldsymbol{x}_j^{D}}^{0}\right)^{-1}\left(u_{j,b}^{*}\right)\right), \tag{12}$$

where $\left\{\varepsilon_{i,b}^{\bar{D}*}\right\}_{i=1}^{n_{\bar{D}}}$ is a sample of i.i.d. observations from distribution $\widehat{G}_{\bar{D}}$, and $\left\{u_{j,b}^{*}\right\}_{j=1}^{n_D}$ is a sample of i.i.d. observations from a uniform distribution on the interval $[0, 1]$.

Step 4. From $\left\{\left(\boldsymbol{x}_i^{\bar{D}}, y_{i,b}^{\bar{D}*}\right)\right\}_{i=1}^{n_{\bar{D}}}$ and $\left\{\left(\boldsymbol{x}_j^{D}, y_{j,b}^{D*}\right)\right\}_{j=1}^{n_D}$ obtain T^b ($T_{||}^b$ or T_2^b).

In Section 4.3 we prove that the resamples obtained as explained in Step 3 above verify the null hypothesis. Accordingly, the previous procedure approximates the distribution of the test statistic T ($T_{||}$ or T_2) under H_0. Thus, the test rule based on T ($T_{||}$ or T_2) consists of rejecting the null hypothesis if $T^o > T_\alpha^B$, where T_α^B is the empirical (1-α)-percentile of the values of $T^1, \ldots, T^B$ obtained in Step 4.

4.2 Testing for factor-by-curve interaction

The test for the null hypothesis

$$H_0 : ROC_{\boldsymbol{x}}(p) = g\left(\beta_0 + \sum_{l=1}^{M}\beta_l I\left(x_u = l\right) + f\left(x_v\right) + h_0\left(p\right)\right), \tag{13}$$

versus the general hypothesis

$$H_1 : ROC_{\boldsymbol{x}}(p) = g\left(\beta_0 + \sum_{l=1}^{M}\beta_l I\left(x_u = l\right) + f\left(x_v\right) + \sum_{l=1}^{M} f^l\left(x_v\right) I\left(x_u = l\right) + h_0\left(p\right)\right),$$

is based on the estimates of the interaction curves f^l $(l = 1, \ldots, M)$. As before, L_1 and L_2 norms are considered, yielding the test statistics

$$S_{||} = \sum_{j=1}^{n_D}\sum_{l=1}^{M}\left|\hat{f}^l\left(x_{jv}^D\right) I\left(x_{ju}^D = l\right)\right| \quad \text{and} \quad S_2 = \sum_{j=1}^{n_D}\sum_{l=1}^{M}\left(\hat{f}^l\left(x_{jv}^D\right) I\left(x_{ju}^D = l\right)\right)^2.$$

The proposed statistics are measures of deviations from the estimated nonparametric interaction curves $\hat{f}^l$ $(l = 1, \ldots, M)$ to their means $n_D^{-1}\sum_{j=1}^{n_D}\hat{f}^l\left(x_{jv}^D\right) I\left(x_{ju}^D = l\right)$. We note that, as before, the means are forced to be zero during estimation.

The bootstrap-based testing procedure in this case is the same as that presented above to test for the effect of continuous covariates on the ROC curve. The only difference is Step 2 of the algorithm, which now must be

Step 2. Estimate the null ROC-GAM regression model (13) from $\left\{\left(\boldsymbol{x}_j^D, y_j^D\right)\right\}_{j=1}^{n_D}$, and obtain the bootstrap pilot estimates $\widehat{ROC}^0_{\boldsymbol{x}_j^D}(p)$, $1 \leq j \leq n_D$.

4.3 Resampling under the null hypothesis

As previously discussed, a crucial point when applying the procedures presented above is to obtain bootstrap resamples verifying the null hypothesis. In this section we show that the resampling mechanism explained in Section 4.1 meets this requirement.

First, let us re-express the conditional ROC curve given in (1) as follows:

$$\begin{aligned} ROC_{\boldsymbol{x}}(p) &= S_D\left(S_{\bar{D}}^{-1}\left(p \mid \boldsymbol{x}\right) \mid \boldsymbol{x}\right) \\ &= P\left(Y_D \geq S_{\bar{D}}^{-1}\left(p \mid \boldsymbol{x}\right) \mid \boldsymbol{X} = \boldsymbol{x}\right) \\ &= P\left(S_{\bar{D}}\left(Y_D \mid \boldsymbol{x}\right) \leq p \mid \boldsymbol{X} = \boldsymbol{x}\right). \end{aligned}$$

Thus, the conditional ROC curve may be seen as the conditional cumulative distribution function of the random variable $S_{\bar{D}}\left(Y_D \mid \boldsymbol{x}\right)$. This equivalence, in conjunction with the location-scale regression model assumed for $Y_{\bar{D}}$ (see (9)), implies that

$$Y_D^* = \mu_{\bar{D}}\left(\boldsymbol{X}\right) + \sigma_{\bar{D}}\left(\boldsymbol{X}\right) G_{\bar{D}}^{-1}\left(ROC_{\boldsymbol{X}}^{-1}\left(U\right)\right), \tag{14}$$

with $U \sim U[0,1]$, is a random variable with conditional cumulative survival function

$$\begin{aligned} S_D^*\left(c \mid \boldsymbol{x}\right) &= P\left(Y_D^* \geq c \mid \boldsymbol{X} = \boldsymbol{x}\right) \\ &= P\left(ROC_{\boldsymbol{x}}^{-1}\left(U\right) \leq G_D\left(\frac{c - \mu_{\bar{D}}\left(\boldsymbol{x}\right)}{\sigma_{\bar{D}}\left(\boldsymbol{x}\right)}\right) \middle| \boldsymbol{X} = \boldsymbol{x}\right) \\ &= P\left(ROC_{\boldsymbol{x}}^{-1}\left(U\right) \leq S_D\left(c \mid \boldsymbol{x}\right) \mid \boldsymbol{X} = \boldsymbol{x}\right) \\ &= S_D\left(S_{\bar{D}}^{-1}\left(S_{\bar{D}}\left(c \mid \boldsymbol{x}\right) \mid \boldsymbol{x}\right) \mid \boldsymbol{x}\right) \\ &= S_D\left(c \mid \boldsymbol{x}\right). \end{aligned}$$

Therefore, given $\boldsymbol{X} = \boldsymbol{x}$, the conditional ROC curve related to $Y_{\bar{D}}$ and Y_D^* is the same as the one associated with $Y_{\bar{D}}$ and Y_D, i.e., $ROC_{\boldsymbol{x}}\left(\cdot\right)$. Note that in the previous result it is assumed that $S_{\bar{D}}\left(\cdot \mid \boldsymbol{x}\right)$ is a monotonically strictly decreasing function (which also implies that the conditional ROC curve is continuous).

Finally, it is worth emphasising that the (conditional) ROC curve provides a description of the separation between the (conditional) distributions of the diagnostic test in healthy and diseased populations, regardless of the specific location of both distributions. This property of the ROC curve, jointly with result (14), thus suggests the resampling plan discussed in Section 4.1:

- The healthy population is kept fixed, and a bootstrap of residuals is used to obtain the sample in the healthy population (see eqn. (11)).
- Result (14) is used to obtain the bootstrap sample in the diseased population, where the theoretical quantities are replaced by their respective estimates (see eqn. (12)). In order to "mimic" the null hypothesis when resampling, the conditional ROC curve under the null hypothesis (see models (10) and (13)) is substituted for $ROC_{\boldsymbol{x}}(\cdot)$ in (14).

5 Simulation study

In this section we report on a simulation study designed to assess the validity of the bootstrap-based tests described in Section 4 above. Extra simulation results can be found in the Supplementary Material available online.

Data are simulated from four different scenarios, namely,

- Scenario I

$$Y_{\bar{D}} = \sin(\pi X_{v1}) - a0.3X_{v1}^3 + \sqrt{0.2 + 0.5\exp(X_{v1})}\varepsilon_{\bar{D}},$$
$$Y_D = \sin(\pi X_{v1}) + \sqrt{0.2 + 0.5\exp(X_{v1})} + \sqrt{0.2 + 0.5\exp(X_{v1})}\varepsilon_D.$$

- Scenario II

$$Y_{\bar{D}} = -2X_{v1}^2 + 0.5\exp(X_{v2}) + 0.5\varepsilon_{\bar{D}},$$
$$Y_D = aX_{v1}^2 - 2X_{v1}^2 + 0.5\sin(\pi(X_{v2}+1)) + 0.5\exp(X_{v2}) + 0.5\varepsilon_D.$$

- Scenario III

$$Y_{\bar{D}} = -0.25X_{v1}^3 + 0.5X_{v1}^2 + 0.5X_{v1}^2X_{u1} - 0.5X_{v1}^2(1 - X_{u1}) + 0.5\varepsilon_{\bar{D}},$$
$$Y_D = 0.25X_{v1}^3 + (a+1)\left(0.5X_{v1}^2 + 0.5X_{v1}^2X_{u1} - 0.5X_{v1}^2(1 - X_{u1})\right) + 0.5\varepsilon_D.$$

- Scenario IV

$$Y_{\bar{D}} = \sin(2X_{v1}) + (1-a)\left(\frac{1}{1+\exp(-10X_{v1})}(1 - X_{u1}) - \frac{1}{1+\exp(-10X_{v1})}X_{u1}\right) + \varepsilon_{\bar{D}},$$
$$Y_D = X_{u1} + \sin(X_{v1}) + \frac{1}{1+\exp(-10X_{v1})}(1 - X_{u1}) - \frac{1}{1+\exp(-10X_{v1})}X_{u1} + \varepsilon_D.$$

In all cases, a is a real constant, X_{v1} is simulated from a uniform distribution on $[-1, 1]$ and $\varepsilon_{\bar{D}}$ and $\varepsilon_D \sim N(0, 1)$. In Scenario II, X_{v2} is a continuous covariate which is simulated from a uniform distribution on $[-1, 1]$, and Scenario III and IV represent the factor-by-curve case. Here $X_{u1} \sim Bernoulli(0.5)$.

With the above configurations, the corresponding conditional ROC curves are respectively

- Scenario I

$$ROC_{\boldsymbol{x}}(p) = \Phi\left(1 + \frac{a0.3x_{v1}^3}{\sqrt{0.2 + 0.5\exp\left(x_{v1}\right)}} + \Phi^{-1}\left(p\right)\right),$$

- Scenario II

$$ROC_{\boldsymbol{x}}(p) = \Phi\left(2ax_{v1}^2 + \sin(\pi(x_{v2}+1)) + \Phi^{-1}(p)\right),$$

- Scenario III

$$ROC_{\boldsymbol{x}}(p) = \Phi\left(x_{v1}^3 + a\left(x_{v1}^2 + x_{v1}^2 x_{u1} - x_{v1}^2(1-x_{u1})\right) + \Phi^{-1}\left(p\right)\right),$$

- Scenario IV

$$ROC_{\boldsymbol{x}}(p) = \Phi\left(x_{u1} + a\left(\frac{1}{1+\exp(-10x_{v1})}(1-x_{u1}) - \frac{1}{1+\exp(-10x_{v1})}x_{u1}\right) + \Phi^{-1}\left(p\right)\right),$$

where Φ denotes the cumulative distribution function of a standard normal random variable. We note that Scenario I was also considered in Rodríguez-Álvarez et al. (2011b) and that Scenario IV was designed to mimic the CAD data discussed in Section 7 below.

In order to fit a ROC-GAM regression model, several choices need to be made. In all results shown below, the probit function, namely $g^{-1} = \Phi^{-1}$, is taken as the link function. With respect to the set of FPFs – needed in Step 1 of the algorithm presented in Appendix A – $n_P = 50$ and equally-spaced values are considered. Our implementation makes use of binning type acceleration techniques (Fan and Marron, 1994) to reduce computational time. In this study, we use 30 equally-spaced binning points along the range of each of the continuous covariates. The bandwidths involved in the local-linear kernel smoothers are selected using the standard procedure of leave-one-out cross validation, and recomputed for each bootstrap resample.

To study the size and power of the tests, different values are considered for a. Note that a controls the deviation from the null hypothesis. In Scenarios I and II, $a = 0$ corresponds to the hypothesis of no effect of covariate X_{v1} on the ROC curve, and the more the constant a shifts towards zero, the greater the effect of the covariate on the ROC curve. For Scenarios III and IV, the value $a = 0$ corresponds to the hypothesis of no interaction between X_{v1} and X_{u1}, and as the value of a rises, so does the degree of interaction. These behaviours are illustrated in Figure 1. Note that the y-scale is different in the four plots. Thus, for a specific value of a (excluding $a = 0$), the largest deviation from the null hypothesis would be for Scenario II, and the lowest for Scenarios I and IV.

The bootstrap procedure described in Section 4.1 is applied to Scenarios I and II, and the one presented in Section 4.2 to Scenarios III and IV. In all cases, critical values and p-values are determined using $B = 400$ bootstrap samples. Type I errors and powers are

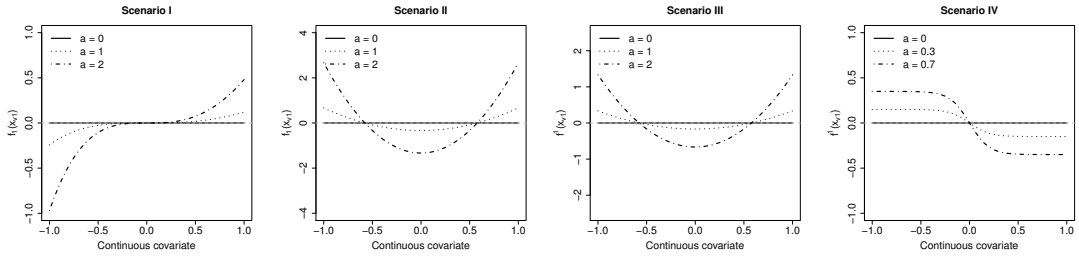


Figure 1: For Scenarios I and II: centred nonparametric function of X_{v1}. For Scenarios III and IV: centred interaction curve of X_{v1} for $X_{u1} = 1$. In all cases, the partial functions are shown for different values of a. The dotted grey line represents the null hypothesis (no effect/interaction), which also corresponds to $a = 0$.

calculated as the proportion of rejections of H_0 in 1000 runs. Bearing in mind the sample sizes, the following situations are considered: (1) the same sample size for both healthy and diseased subjects, with $n_D = n_{\bar{D}} = 50, 100, 200, 500, 1000$; and (2) very unbalanced sample sizes – consistent with the CAD data – with $n_D = 32$ and $n_{\bar{D}} = 200$, $n_D = 64$ and $n_{\bar{D}} = 400$, $n_D = 128$ and $n_{\bar{D}} = 800$, and $n_D = 256$ and $n_{\bar{D}} = 1600$. For the sake of brevity, only the results for $n_D = n_{\bar{D}} = 50$ and 1000, $n_D = 32$ and $n_{\bar{D}} = 200$, and $n_D = 256$ and $n_{\bar{D}} = 1600$ are shown below. Results for the remaining sample sizes are consistent with those presented here. Since p-values should be uniformly distributed under the null hypothesis, the Kolmogorov-Smirnov (KS) test for uniformity of the resulting p-values is also performed.

Table 1 shows the type I errors registered by the proposed tests for different significance levels and sample sizes. The p-values of the KS-test are also shown in this table. Figure 2 depicts quantile-quantile plots of the expected p-values (under the uniform distribution) and the observed p-values for all Scenarios and tests considered in this paper. As can be seen, the tests perform well in general, with type I errors proving to be relatively close to nominal errors (Table 1), and p-value distributions close to the uniform one (Figure 2). We note that there are some situations where the KS test rejects the null hypothesis of a uniform distribution, but mainly at low sample sizes. We are especially concerned about the result for Scenario III with such a large sample size, $n_D = n_{\bar{D}} = 1000$. Accordingly, we evaluate the behaviour of the tests (under the same conditions) for a sample size of $n_D = n_{\bar{D}} = 2000$. In this case (results not shown), the KS test gives p-values of 0.223 and 0.471 for $S_{||}$ and S_2 tests, respectively.

Power as a function of constant a, at different significance levels, is shown in Tables 2 - 5, and Figure 3 shows the power curves at 0.05 significance level. Both tests register satisfactory power curves, with the probability of rejection rising in response to any increase in the value of the constant a and/or the sample size. In general, both tests have very similar power. However, especially for Scenario IV, the test based on the L_1 norm seems

to be more powerful. Finally, note that the power curves depict the expected behaviour according to the plots shown in Figure 1.

Additional simulation studies are provided as online Supplementary material. Shown there are the results when assuming (a) different distributions for $\varepsilon_{\bar{D}}$ and ε_D (Student's t distributions and mixture of Gaussian distributions); and (b) that covariates only affect the result of the diagnostic test in the diseased population. The results are consistent with those discussed here. In brief, the tests produce type I errors close to nominal levels. As expected, the probability of rejection rises as the sample size increases. In addition, the results also highlight that the L_1-norm test is slightly more powerful.

6 Software implementation: the `npROCRegression` package

This section contains a brief description of the `R`-package we developed to accompany this paper. The package can be freely downloaded from https://cran.r-project.org/package=npROCRegression, where a more detailed explanation of its use can be found. To facilitate the use of the package by the biomedical community, `npROCRegression` has been designed in a similar fashion to other regression functions/packages in `R`. The main functions of the package are `DNPROCreg()` and `INPROCReg`, which estimate the conditional ROC curve based on, respectively, the nonparametric direct (Rodríguez-Álvarez et al., 2011a) and induced (Rodríguez-Álvarez et al., 2011b) regression approaches. Numerical and graphical summaries of the fitted models can be obtained by calling the functions `print()`, `summary()` and `plot()`.

6.1 `DNPROCreg()` function

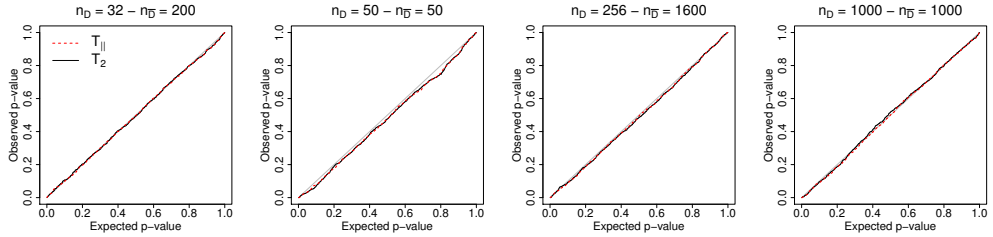
The function `DNPROCreg()` estimates the conditional ROC curve in the presence of multi-dimensional covariates by means of the ROC-GAM regression approach presented earlier. Usage is as follows:

```
DNPROCreg(marker, formula.h = ~ 1, formula.ROC = ~ 1,
group, tag.healthy, data,
ci.fit = FALSE
test.partial = NULL,
newdata = NULL,
control = controlDNPROCreg(),
weights = NULL)
```

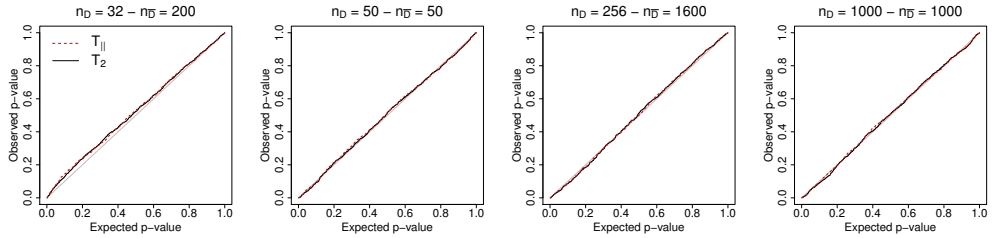
The diagnostic test variable is indicated by the argument `marker`. The nonparametric location-scale regression model for the healthy population (see (9)) is specified by `formula.h`. This argument should be a vector (of length 2) of right-hand formulas (atomic values are also valid, because they are recycled). The first right-hand formula is the model for the regression function, $\mu_{\bar{D}}(\boldsymbol{x})$, and the second one is the model for the (logarithm)

Table 1: For Scenarios I, II, III and IV: estimated type I error registered by the proposed tests under the null hypothesis, for different significance levels and sample sizes. The last column presents the p-values of the Kolmogorov-Smirnov test for uniformity of the observed p-values.

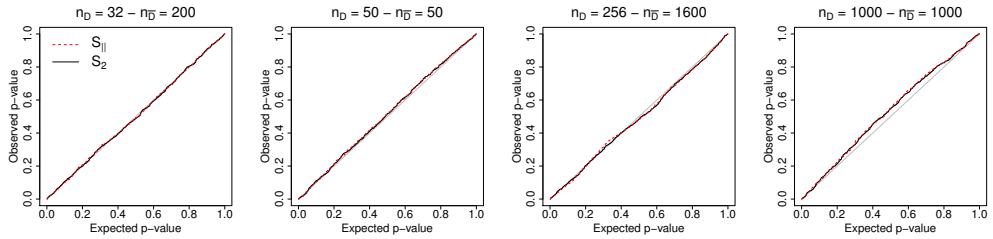
	Sample size			Level					
	n_D	$n_{\bar{D}}$	Test	0.01	0.05	0.10	0.15	0.20	KS p-value
Scenario I	32	200	$T_{\|\|}$	0.007	0.039	0.102	0.159	0.204	0.901
			T_2	0.009	0.045	0.092	0.146	0.196	0.948
	50	50	$T_{\|\|}$	0.011	0.070	0.126	0.173	0.220	0.020
			T_2	0.013	0.077	0.127	0.174	0.210	0.004
	256	1600	$T_{\|\|}$	0.011	0.053	0.113	0.165	0.210	0.695
			T_2	0.011	0.047	0.112	0.162	0.211	0.663
	1000	1000	$T_{\|\|}$	0.020	0.058	0.110	0.171	0.217	0.746
			T_2	0.016	0.056	0.108	0.161	0.217	0.744
Scenario II	32	200	$T_{\|\|}$	0.008	0.031	0.067	0.108	0.157	0.009
			T_2	0.007	0.033	0.074	0.121	0.165	0.042
	50	50	$T_{\|\|}$	0.015	0.052	0.107	0.145	0.184	0.582
			T_2	0.016	0.060	0.112	0.145	0.192	0.709
	256	1600	$T_{\|\|}$	0.012	0.060	0.111	0.163	0.212	0.611
			T_2	0.015	0.063	0.119	0.164	0.219	0.732
	1000	1000	$T_{\|\|}$	0.019	0.053	0.109	0.152	0.205	0.558
			T_2	0.019	0.055	0.111	0.168	0.203	0.657
Scenario III	32	200	$S_{\|\|}$	0.008	0.054	0.102	0.150	0.190	0.805
			S_2	0.008	0.051	0.095	0.149	0.203	0.892
	50	50	$S_{\|\|}$	0.015	0.053	0.097	0.143	0.191	0.412
			S_2	0.011	0.053	0.091	0.152	0.193	0.359
	256	1600	$S_{\|\|}$	0.009	0.056	0.120	0.162	0.205	0.358
			S_2	0.011	0.047	0.104	0.163	0.197	0.221
	1000	1000	$S_{\|\|}$	0.017	0.044	0.091	0.136	0.177	0.001
			S_2	0.013	0.053	0.089	0.134	0.188	0.003
Scenario IV	32	200	$S_{\|\|}$	0.013	0.055	0.110	0.153	0.211	0.080
			S_2	0.012	0.056	0.118	0.161	0.213	0.075
	50	50	$S_{\|\|}$	0.013	0.045	0.093	0.141	0.187	0.006
			S_2	0.014	0.043	0.095	0.134	0.185	0.001
	256	1600	$S_{\|\|}$	0.018	0.056	0.109	0.156	0.218	0.533
			S_2	0.017	0.053	0.104	0.157	0.221	0.691
	1000	1000	$S_{\|\|}$	0.027	0.071	0.121	0.159	0.227	0.412
			S_2	0.023	0.062	0.120	0.174	0.217	0.481



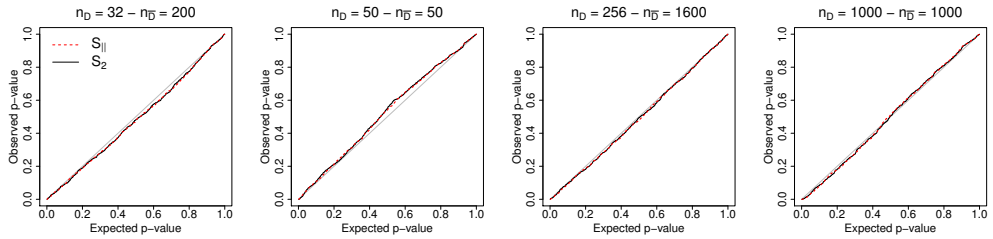
(a) Scenario I



(b) Scenario II



(c) Scenario III



(d) Scenario IV

Figure 2: For Scenarios I, II, III and IV: Quantile-quantile plot for the observed p-values vs. the expected p-values when the null hypothesis is correct.

Table 2: For Scenario I: estimated rejection probabilities registered by the proposed tests under the alternative hypothesis, for different values of a, significance levels and sample sizes.

		Sample size			Level				
		n_D	$n_{\bar{D}}$	Test	0.01	0.05	0.10	0.15	0.20
Scenario I	$a = 0.5$	32	200	$T_{\|\|}$	0.011	0.055	0.107	0.164	0.221
				T_2	0.010	0.057	0.101	0.156	0.206
		50	50	$T_{\|\|}$	0.023	0.068	0.133	0.184	0.227
				T_2	0.021	0.076	0.131	0.180	0.223
		256	1600	$T_{\|\|}$	0.046	0.120	0.200	0.268	0.330
				T_2	0.039	0.112	0.192	0.258	0.329
		1000	1000	$T_{\|\|}$	0.097	0.222	0.318	0.392	0.475
				T_2	0.091	0.210	0.304	0.396	0.467
	$a = 1.0$	32	200	$T_{\|\|}$	0.015	0.061	0.124	0.188	0.236
				T_2	0.015	0.063	0.122	0.181	0.237
		50	50	$T_{\|\|}$	0.027	0.091	0.151	0.208	0.269
				T_2	0.034	0.092	0.158	0.209	0.265
		256	1600	$T_{\|\|}$	0.133	0.309	0.417	0.500	0.557
				T_2	0.142	0.316	0.422	0.488	0.551
		1000	1000	$T_{\|\|}$	0.417	0.652	0.764	0.832	0.874
				T_2	0.468	0.673	0.779	0.834	0.875
	$a = 2.0$	32	200	$T_{\|\|}$	0.048	0.115	0.201	0.288	0.349
				T_2	0.048	0.112	0.212	0.287	0.338
		50	50	$T_{\|\|}$	0.060	0.159	0.241	0.311	0.382
				T_2	0.074	0.161	0.247	0.318	0.381
		256	1600	$T_{\|\|}$	0.681	0.846	0.911	0.937	0.958
				T_2	0.712	0.859	0.917	0.945	0.967
		1000	1000	$T_{\|\|}$	0.990	0.999	1.000	1.000	1.000
				T_2	0.997	0.999	1.000	1.000	1.000

Table 3: For Scenario II: estimated rejection probabilities registered by the proposed tests under the alternative hypothesis, for different values of a, significance levels and sample sizes.

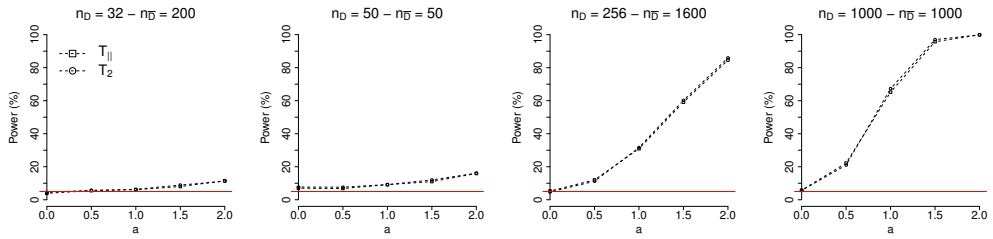
		Sample size			Level				
		n_D	$n_{\bar{D}}$	Test	0.01	0.05	0.10	0.15	0.20
Scenario II	$a = 0.5$	32	200	$T_{\|\|}$	0.031	0.096	0.165	0.211	0.271
				T_2	0.024	0.087	0.160	0.216	0.268
		50	50	$T_{\|\|}$	0.037	0.097	0.176	0.226	0.268
				T_2	0.032	0.097	0.164	0.225	0.280
		256	1600	$T_{\|\|}$	0.648	0.818	0.889	0.912	0.935
				T_2	0.600	0.794	0.872	0.901	0.926
		1000	1000	$T_{\|\|}$	0.993	0.996	0.998	0.998	0.999
				T_2	0.990	0.996	0.998	0.998	0.999
	$a = 1.0$	32	200	$T_{\|\|}$	0.155	0.325	0.421	0.519	0.593
				T_2	0.137	0.299	0.414	0.505	0.575
		50	50	$T_{\|\|}$	0.142	0.305	0.398	0.473	0.540
				T_2	0.126	0.293	0.390	0.456	0.512
		256	1600	$T_{\|\|}$	1.000	1.000	1.000	1.000	1.000
				T_2	1.000	1.000	1.000	1.000	1.000
		1000	1000	$T_{\|\|}$	1.000	1.000	1.000	1.000	1.000
				T_2	1.000	1.000	1.000	1.000	1.000
	$a = 2.0$	32	200	$T_{\|\|}$	0.608	0.833	0.896	0.931	0.953
				T_2	0.572	0.809	0.886	0.921	0.950
		50	50	$T_{\|\|}$	0.574	0.783	0.867	0.920	0.942
				T_2	0.534	0.751	0.841	0.904	0.933
		256	1600	$T_{\|\|}$	1.000	1.000	1.000	1.000	1.000
				T_2	1.000	1.000	1.000	1.000	1.000
		1000	1000	$T_{\|\|}$	1.000	1.000	1.000	1.000	1.000
				T_2	1.000	1.000	1.000	1.000	1.000

Table 4: For Scenario III: estimated rejection probabilities registered by the proposed tests under the alternative hypothesis, for different values of a, significance levels and sample sizes.

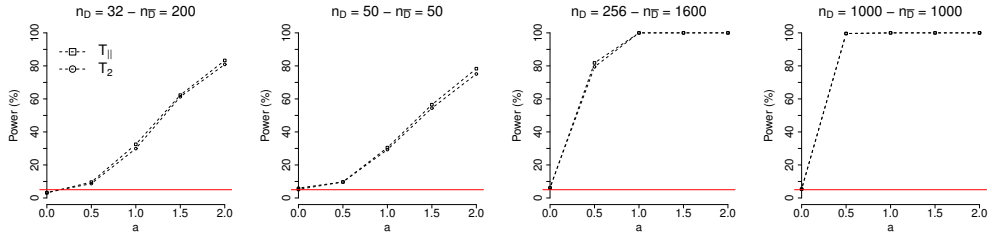
		Sample size			Level				
		n_D	$n_{\bar{D}}$	Test	0.01	0.05	0.10	0.15	0.20
Scenario III	$a = 0.5$	32	200	$S_{\|\|}$	0.016	0.060	0.123	0.174	0.239
				S_2	0.015	0.065	0.122	0.181	0.237
		50	50	$S_{\|\|}$	0.020	0.081	0.146	0.207	0.257
				S_2	0.020	0.083	0.143	0.203	0.257
		256	1600	$S_{\|\|}$	0.172	0.321	0.429	0.509	0.579
				S_2	0.160	0.310	0.423	0.501	0.587
		1000	1000	$S_{\|\|}$	0.567	0.769	0.853	0.902	0.929
				S_2	0.527	0.753	0.848	0.890	0.917
	$a = 1.0$	32	200	$S_{\|\|}$	0.024	0.092	0.177	0.262	0.332
				S_2	0.027	0.086	0.175	0.253	0.319
		50	50	$S_{\|\|}$	0.052	0.130	0.210	0.284	0.358
				S_2	0.050	0.127	0.208	0.276	0.347
		256	1600	$S_{\|\|}$	0.734	0.875	0.926	0.950	0.966
				S_2	0.690	0.853	0.918	0.942	0.963
		1000	1000	$S_{\|\|}$	0.998	1.000	1.000	1.000	1.000
				S_2	0.998	0.999	1.000	1.000	1.000
	$a = 2.0$	32	200	$S_{\|\|}$	0.073	0.197	0.308	0.417	0.487
				S_2	0.056	0.185	0.297	0.383	0.471
		50	50	$S_{\|\|}$	0.111	0.270	0.372	0.464	0.527
				S_2	0.095	0.235	0.361	0.438	0.518
		256	1600	$S_{\|\|}$	1.000	1.000	1.000	1.000	1.000
				S_2	1.000	1.000	1.000	1.000	1.000
		1000	1000	$S_{\|\|}$	1.000	1.000	1.000	1.000	1.000
				S_2	1.000	1.000	1.000	1.000	1.000

Table 5: For Scenario IV: estimated rejection probabilities registered by the proposed tests under the alternative hypothesis, for different values of a, significance levels and sample sizes.

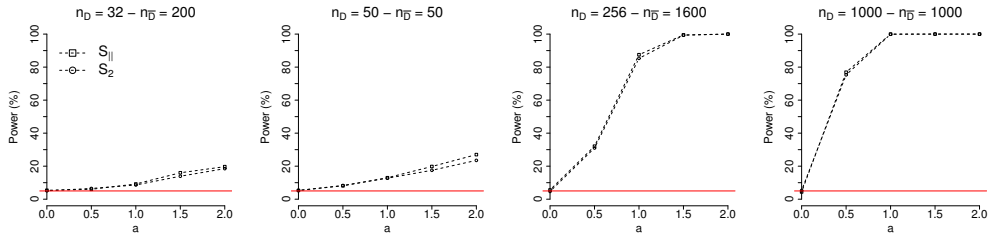
		Sample size			Level				
		n_D	$n_{\bar{D}}$	Test	0.01	0.05	0.10	0.15	0.20
Scenario IV	$a = 0.3$	32	200	$S_{\|\|}$	0.016	0.066	0.131	0.189	0.246
				S_2	0.016	0.062	0.130	0.180	0.240
		50	50	$S_{\|\|}$	0.011	0.068	0.122	0.175	0.218
				S_2	0.013	0.063	0.117	0.171	0.222
		256	1600	$S_{\|\|}$	0.164	0.351	0.476	0.571	0.641
				S_2	0.121	0.299	0.420	0.521	0.599
		1000	1000	$S_{\|\|}$	0.569	0.786	0.866	0.906	0.921
				S_2	0.477	0.694	0.815	0.862	0.903
	$a = 0.5$	32	200	$S_{\|\|}$	0.021	0.085	0.155	0.219	0.290
				S_2	0.023	0.096	0.171	0.236	0.297
		50	50	$S_{\|\|}$	0.027	0.087	0.149	0.211	0.268
				S_2	0.020	0.068	0.134	0.199	0.271
		256	1600	$S_{\|\|}$	0.563	0.787	0.856	0.900	0.922
				S_2	0.469	0.705	0.817	0.872	0.911
		1000	1000	$S_{\|\|}$	0.982	0.995	0.998	0.998	0.999
				S_2	0.959	0.989	0.994	0.998	0.998
	$a = 0.7$	32	200	$S_{\|\|}$	0.046	0.130	0.227	0.299	0.365
				S_2	0.039	0.115	0.196	0.277	0.339
		50	50	$S_{\|\|}$	0.039	0.130	0.204	0.276	0.340
				S_2	0.032	0.116	0.187	0.248	0.315
		256	1600	$S_{\|\|}$	0.908	0.973	0.983	0.989	0.993
				S_2	0.855	0.949	0.978	0.984	0.992
		1000	1000	$S_{\|\|}$	1.000	1.000	1.000	1.000	1.000
				S_2	1.000	1.000	1.000	1.000	1.000



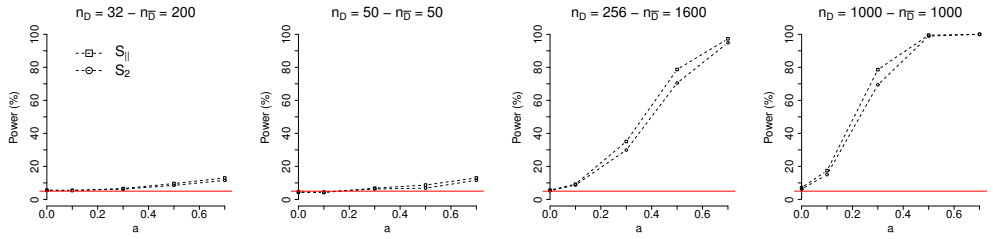
(a) Scenario I



(b) Scenario II



(c) Scenario III



(d) Scenario IV

Figure 3: For Scenarios I, II, III and IV: estimated rejection probabilities registered by the proposed tests, as a function of the parameter a, for different sample sizes and at a 0.05 significance level (red line).

of the variance function, $\sigma^2_{\bar{D}}(\boldsymbol{x})$. These formulas are similar to that used for the `glm()` function, except that nonparametric functions can be added to the additive predictor by means of function `s()`. For instance, specification $\sim$ `x1 + s(x2)` would assume a linear effect of `x1` and a nonparametric effect of `x2`. Categorical variables (factors) can be also incorporated, as well as factor-by-curve interaction terms as was discussed in Section 3. For example, to include the interaction between *age* and *gender* we need to specify $\sim$ `gender + s(age) + s(age, by = gender)`. Note that, for identifiability purposes, the "main" effects of the continuous and categorical covariates need to be included in the formula. These considerations also apply to the argument `formula.ROC`, where the ROC-GAM regression model (see (7) and (8)) is specified.

The name of the variable that distinguishes healthy from diseased individuals is represented by argument `group`, and the value codifying the healthy individuals in `group` is specified by `tag.healthy`. The `data` argument is a data frame containing the data and all needed variables.

Pointwise bootstrap confidence intervals for each component of the additive predictor of the ROC-GAM, as well as for the conditional AUCs (with the integral approximated by numerical integration methods), are obtained by setting the argument `ci.fit` to `TRUE`.

The components of the ROC-GAM to be tested for their possible effect are indicated in `test.partial`. In this argument, we pass the position of the components as specified in the `formula.ROC` argument.

An optional data frame containing the covariate values at which predictions are required can be specified in argument `newdata`. If missing, an adequate set of points from the dataset used in the fit is selected. To that end, the function `DNPROCregdata()` is used.

Argument `control` allows us to modify some default parameters that control the fitting process. For instance, the cardinality of the set of FPF used in the estimation process (see Appendix A, Step 1) can be specified using this argument (by default $n_P = 50$), as can the link function, the number of bootstrap resamples, and the significance level used for the construction of the confidence intervals.

6.2 `INPROCreg()` function

The function `INPROCreg()` estimates the conditional ROC curve in the presence of a one-dimensional continuous covariate, using the induced nonparametric ROC regression approach as presented in Rodríguez-Álvarez et al. (2011b). The call to the function is as follows:

```
INPROCreg (marker, covariate,
group, tag.healthy, data,
ci.fit = FALSE, test = FALSE,
accuracy = NULL, accuracy.cal = c("ROC","AROC"),
newdata = NULL, control = controlINPROCreg(),
weights = NULL)
```

Through `marker` and `covariate` arguments, users indicate the diagnostic test variable and the continuous covariate of interest, respectively.

In `group` and `tag.healthy` arguments, we indicate respectively the name of the variable that distinguishes healthy from diseased individuals, and the value codifying healthy individuals in that variable. The `data` argument is a data frame containing the data and all needed variables.

Bootstrap confidence intervals for the regression and variance functions, as well as for several accuracy measures, are obtained by setting the argument `ci.fit` to `TRUE`. Argument `test` should be set to `TRUE` in order to evaluate the effect of the continuous covariate on the ROC curve via the test presented in Rodríguez-Álvarez et al. (2011b).

By default, the `INPROCreg()` function returns the estimated regression and variance functions both in healthy and diseased populations. As far as accuracy measures are concerned, the function provides the estimated conditional ROC curve, the associated conditional AUCs (with the integral approximated by numerical integration methods), and the covariate-adjusted ROC curve, AROC (see (6)). In addition, it is also possible to obtain the conditional Youden index ("YI"), the covariate-specific values for which the TPF and the TNF coincide ("EQ"), and/or the covariate-specific optimal thresholds ("TH") based on the previous two criteria (argument `accuracy`). Both the YI and the EQ values (and thus the optimal thresholds) can be calculated based on the conditional ROC curve or the AROC curve (argument `accuracy.cal`) (see, e.g., eqn (5)). We recommend the use of the AROC curve in those situations where the accuracy of the test does not vary along with the covariate. Note that, even in this case, covariate-specific thresholds could be obtained (Rodríguez-Álvarez et al., 2011b).

An optional data frame containing the values of the covariate at which predictions are required can be specified in argument `newdata`. If this dataset is not specified, an adequate set of points from the data used in the fit is selected. A finer control of the fitting process can be achieved by the argument `control`. This argument can be used to select the number of binning points, for instance, or the order of the polynomial associated to the kernel smoothers.

7 Application to a CAD system

Computer-aided diagnosis (CAD) has been defined as the diagnosis made by a radiologist who takes into account the results of quantitative computer analysis of a medical diagnosis (Doi, 2007). These kind of systems have demonstrated their usefulness in situations where the radiologists have to discriminate positive cases among hundreds or thousands of normal cases. Screening programs present a challenge for physicians, and the presence of a second reader in the form of a computer algorithm has been demonstrated to be useful (Nishikawa, 2007). The foundation of a CAD system is a computer vision algorithm which extracts some features from an image. The parameters that represent these features are fed to a

classifier which is specifically trained to discriminate between normal and abnormal cases.

CAD schemes have been developed for screening programs related to the detection of cancer in the breast, chest, colon, etc. The main issue for breast cancer detection is the identification of masses and microcalcifications (Sickles, 1984). This paper focuses on a CAD scheme specially designed for breast mass detection (see Varela et al., 2007, for a detailed description). The system is designed to extract several image features, such as maximum and minimum, average, size, eccentricity, contrast, coarseness, etc. for each suspicious region. Among these features, those related to the iris filter have a special importance in terms of the performance of the whole system. The iris filter is an algorithm specially designed for highlighting rounded and brilliant structures within an image. Since masses have such an appearance on a mammogram, features related to this property are of special interest in the development of CAD systems.

Another important consideration in developing a CAD scheme is the impact that a particular feature has on the performance of the whole system. For breast cancer, the main problem for mass detection is the presence of several structures related to glandular tissue, which have an attenuation coefficient similar to that of masses. In some cases (dense breast), the presence of such structures is abundant, therefore hiding the presence of masses. By contrast, when the presence of glandular tissue is negligible (fatty breast), mass detection becomes relatively easy. In addition, the volume of the breast and its composition differ from breast to breast. Thus, the contrast and even the average grey level of the pixels of the final image could be quite different, despite the use of automatic exposure control systems for image acquisition. The consequence is that for humans and machines, the task of detection of possible cancerous masses becomes more difficult.

7.1 Data set

The database contains 580 mammograms, with a total of 190 images classified as abnormal (lesion present), and the remaining 390 as normal (no lesion present). From the 580 original mammograms, the computer detected (in a first step) a total of 2796 regions suspicious of being a malignant mass. Of these, 384 corresponded to true masses, and the remainder, a total of 2412, corresponded to false detections. Table 6 shows summary statistics of the iris filter for fatty and dense tissue types, as well as for several average grey level strata.

7.2 Data analysis

The main purpose of this study on CAD systems is to statistically assess the possible effect of the average grey level (AGL) of the pixels forming the suspicious region and the breast tissue type (TIS) on the accuracy of the iris filter (IRIS) when discriminating between real malignant masses (D) and false detections ($\bar{D}$). To evaluate such effects, Rodríguez-Álvarez et al. (2011a) suggested the use of (semi) parametric ROC regression techniques combined with B-splines, to model the nonlinear effect of AGL on the iris filter, which in turn may

Table 6: Median (interquartile range) of the iris filter for the global sample, for dense and fatty tissues, and for four average grey level strata, based on quartiles.

	True masses	**False detections**
Global sample	0.666 (0.654, 0.676)	0.654 (0.643, 0.667)
Tissue type		
Dense	0.668 (0.654, 0.679)	0.664 (0.674, 0.652)
Fatty	0.664 (0.653, 0.674)	0.639 (0.647, 0.658)
Average grey level		
≤ 0.764	0.659 (0.654, 0.672)	0.649 (0.641, 0.659)
$(0.764, 0.804]$	0.672 (0.662, 0.679)	0.651 (0.642, 0.665)
$(0.804, 0.840]$	0.670 (0.656, 0.680)	0.660 (0.649, 0.671)
> 0.840	0.662 (0.649, 0.673)	0.658 (0.640, 0.670)

vary among tissue types. In this section we re-analyse the CAD data, now using the fully nonparametric ROC-GAM regression approach described in Section 3. This approach allows for the nonparametric specification of the effect of AGL on the ROC curve. Also, the bootstrap-based tests we suggest in Section 4 are used to formally check the possible effect of the covariate AGL and the tissue-by-AGL interaction on the ROC curve.

Before proceeding with the discussion of the results, we should note that all analyses are done with the suspicious region as the unit of analysis. We are aware that the possible correlation induced by the fact a mammogram may contain more that one suspicious region should be taken into account. A limitation of the methodology presented in this paper is that it does not allow to deal with correlated data. As a consequence, the analyses and results discussed here are only presented for the sake of illustrating the proposed methods and the usage of the R-package. Conclusions should, therefore, be analysed with caution.

As a first step of the analysis, the discriminatory capacity of the iris filter is evaluated without taking into account the effect of the covariates. The AUC (95% confidence interval) corresponding to the pooled ROC curve is 0.69 $(0.67, 0.72)$. ROC analysis is also performed for dense and fatty tissues separately, yielding pooled AUCs of 0.64 $(0.59, 0.68)$ and 0.75 $(0.72, 0.79)$, respectively. Additionally, we estimate the AGL-adjusted ROC curves (AROC), both for dense and fatty tissues. The areas under these AROC curves are, in this case 0.60 $(0.55, 0.64)$ and 0.73 $(0.69, 0.78)$. Note that they are slightly lower than the pooled AUCs, possibly indicating that the pooled analysis 'incorporates' the portion of discrimination attributable to AGL (Pardo-Fernández et al., 2014). In any case, these results suggest that the discriminatory capacity of the iris filter is larger for fatty tissue than for dense tissue (see also Table 6).

In order to explore the possible effect of the continuous covariate AGL on the iris filter (and thus on its accuracy), we first consider the induced ROC regression methodology discussed in Section 2.1.1. Specifically, we assume the following nonparametric location-scale regression models for false detection and true masses (separate analyses are conducted

on dense and fatty tissues)

$$\begin{aligned} \text{IRIS}_{\bar{D}} &= \mu_{\bar{D}}(\text{AGL}) + \sigma_{\bar{D}}(\text{AGL})\varepsilon_{\bar{D}}, \\ \text{IRIS}_{D} &= \mu_{D}(\text{AGL}) + \sigma_{D}(\text{AGL})\varepsilon_{D}, \end{aligned} \tag{15}$$

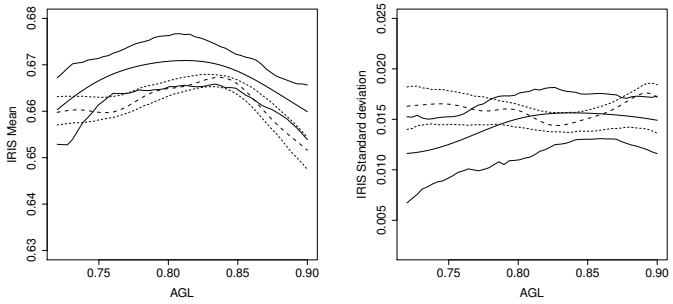
Figure 4 depicts the estimated effect of AGL both on the mean and standard deviation of the filter output, according to breast tissue type, along with 95% pointwise confidence intervals. For masses, in both fatty and dense tissues, mean values rise to a peak approximately midway through the interval and fall thereafter. As a feature that measures the gradual variation in the region's grey level value, filter output tends to rise to a maximum in these intermediate areas, since it is here that such variation could register its most extreme values. For false detections, the pattern is more homogeneous, owing to the fact that, ideally, grey level values display no gradual variation and are instead homogeneously distributed.

The results shown in Figure 4 provide very useful information. First, they suggest the presence of nonlinear effects of AGL on IRIS, which we may expect to be reflected in the diagnostic accuracy. Second, discrimination based on the iris output is much more complex in dense tissue than in fatty tissue. For dense tissue, mean iris filter values are quite similar both for true masses and false detections, and this behaviour is shared by all AGL values. Finally, these results seem to indicate the existence of a possible interaction between average grey levels and tissue type. Accordingly, we fitted the following ROC-GAM regression model including such an interaction

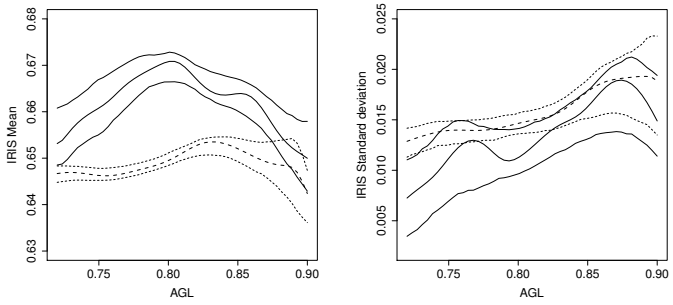
$$ROC_{\text{AGL,TIS}}(p) = \Phi\left(\beta_0 + \sum_{l=1}^{2}\beta_l I\left(TIS = l\right) + f(\text{AGL}) + \sum_{l=1}^{2} f^l(\text{AGL}) I\left(TIS = l\right) + h_0\left(p\right)\right). \tag{16}$$

Here TIS is a binary variable taking a value of 1 in the case of dense tissue and 2 in the case of fatty tissue.

Figure 5 shows the estimated partial functions f (global effect of AGL), f^1 (deviation for dense tissue) and f^2 (deviation for fatty tissue), together with the corresponding 95% pointwise bootstrap confidence intervals. In Figure 6 the estimated conditional AUCs based on model (16) are shown. The estimated AUCs obtained using the induced approach are quite similar to those depicted in Figure 6. However, they are not shown here, for purposes of clarity. As can be seen in Figure 6, the iris filter achieves better results for fatty breasts, as we might expect. Moreover, its performance drops as the AGL increases. This is consistent with the fact that on average, pixel values for fatty breasts are relatively low. When the average pixel value rises it is probably because the overall contrast of the whole breast decreases, due to the size of the breast, the energy of the x-ray beam, or both. In any case, the quality of the image gets worse, and the results achieved by the filter are not as good. On the contrary, for dense breasts, results are almost similar along the entire range of the pixel values. The presence of structures related to the glandular tissue, in



(a) Dense tissue



(b) Fatty tissue

Figure 4: Nonparametric estimates of IRIS by AGL in dense and fatty tissue populations, along with 95% pointwise bootstrap confidence intervals. Solid line: true masses. Dashed line: false detection. Left: nonparametric mean functions. Right: nonparametric variance functions.

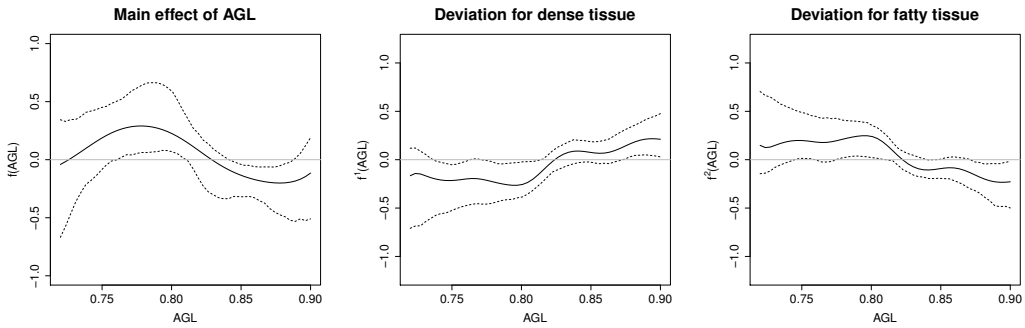


Figure 5: Estimated main effect of AGL in IRIS's accuracy, and deviation for dense and fatty tissue, together with 95% pointwise bootstrap confidence intervals.

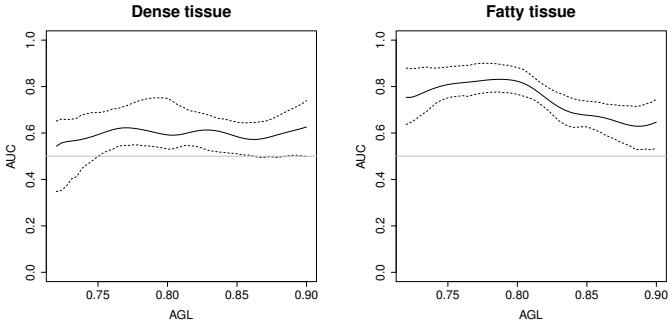


Figure 6: Estimated conditional AUC for the CAD system, according to AGL and type of tissue. The dashed lines represent the 95% pointwise bootstrap confidence intervals.

this case, makes the iris filter not work properly, and as a consequence the enhancement of the mass with respect to other structures is not as pronounced, degrading the detection capabilities of the whole system.

All these results suggest that the presence of an interaction between the type of tissue and the average grey levels is plausible. The tests presented in Section 4.2 are to statistically verify this hypothesis. The resulting p-values are 0.02 and 0.10 for $S_{||}$ and S_2 respectively. Assuming a significance level of 0.05, the result based on S_2 does not suggest the presence of interaction. However, the p-value obtained with $S_{||}$ is lower than 0.05, thus rendering the interaction term significant.

Although both tests could lead to different conclusions, the results of the simulation study seem to indicate that the tests based on the L_1 norm are more powerful or at least as good as the L_2-based tests. In addition, based on all results presented we are also prone to accept the presence of interaction. However, note that a significant interaction does not say anything about the effect of AGV on the accuracy of IRIS: it solely indicates that this

effect is different in dense and fatty tissues. Thus, the last step in our analyses is to check for the effect of AGL on the accuracy of IRIS. To that aim, the following ROC-GAM model is fitted separately in fatty and dense tissues

$$ROC_{\text{AGL}}(p) = \Phi\left(\beta_0 + f(\text{AGL}) + h_0(p)\right),$$

and the tests for continuous covariate effect outlined in Section 4.1 are performed. Whereas the p-values obtained in the case of dense tissue are 0.60 ($T_{||}$) and 0.41 (T_2), in the case of fatty tissue both are < 0.001. So we can conclude that the accuracy of IRIS is constant along AGL in the dense tissue, while in the fatty tissue population the accuracy of IRIS depends significantly on AGL.

7.3 Source code

The R-code used to fit the models presented in Section 7.2 is now given. For the nonparametric induced approach presented in (15), the following code is used

```
library(npROCRegression)

# Set several parameters controlling the fitting process
# p:    order of the local polynomial kernel smoother to be used
#       for estimating the conditional mean functions.
# kbin: number of binning points to be used for the binning
#       approximation.

control.ind = controlINPROCreg(p = 1, kbin = 50)

# Dense tissue
mod.dense <- INPROCreg(marker = "IRIS", covariate = "AGL",
group = "MASS", tag.healthy = 0,
data = subset(masses, TIS == "Dense"),
ci.fit = TRUE, test = TRUE, control = control.ind)

# Fatty tissue
mod.fatty <- marker = "IRIS", covariate = "AGL",
group = "MASS", tag.healthy = 0,
data = subset(masses, TIS == "Fatty"),
ci.fit = TRUE, test = TRUE, control = control.ind)
```

Regarding the ROC-GAM model including the AGL-by-tissue interaction, model (16), the R-code is listed below

```
# card.P: cardinality of the set of FPF to be used for estimating
```

```
#         the ROC-GAM model.
# kbin:   number of binning points to be used for the binning
#         approximation.

control.d = controlDNPROCreg(card.P = 50, kbin = 50)

# Fit the model
mod.int <- DNPROCreg(marker = "IRIS",
formula.h = ~ TIS + s(AGL) + s(AGL, by = TIS),
formula.ROC = ~ TIS + s(AGL) + s(AGL, by = TIS),
group = "MASS", tag.healthy = 0,
data = masses, control = control.d,
ci.fit = TRUE, test.partial = 3)
```

Note that we include the interaction between AGL and TIS not only in the ROC-GAM (`formula.ROC`), but also in the nonparametric location-scale regression model assumed for the healthy population (see (9)), in both the conditional mean and the logarithm of the conditional variance (`formula.h`). Also, by specifying `test.partial = 3` we test for the interaction between the AGL and TIS, which is modelled by means of the third component of the ROC-GAM formula, i.e., `s(AGL, by = TIS)`.

8 Discussion

This paper proposes and investigates L_1- and L_2-norm based test statistics to evaluate the effect of continuous covariates and factor-by-curve interactions in a ROC-GAM regression model. The practical implementation of the proposed tests relies on approximating their distribution under the null hypotheses by means of bootstrap techniques. To that aim, a resampling mechanism that obeys the null hypothesis is proposed. Simulation results show that the proposed procedures yield type I errors relatively close to nominal errors, regardless of sample size. As expected, the power grows as the sample size increases and as one moves further away from the null hypothesis. In general, all tests present a similar power. However, both the simulation study and the real data analysis also suggest that, in some circumstances, the tests based on the L_1 norm (i.e., $T_{||}$ and $S_{||}$) could be more powerful. In practice, we recommend the use of both tests. If the conclusions derived from them are not concordant, as for our CAD system, the L_1 norm test can be considered more reliable. However, we also suggest basing the conclusions not only on p-values (which serve as guidance), but on a comprehensive analysis and understanding of the data.

Our calculations were done with the R-package `npROCRegression` that can be freely downloaded from https://cran.r-project.org/package=npROCRegression. The R-code used for the simulations can also be found at https://bitbucket.org/mxrodriguez/rocgam_inference. The package covers a variety of nonparametric regression approaches

for the inclusion of covariate information on the ROC curve. However, it would be worthwhile to include some extensions of interest. For instance, we could extend the ROC-GAM approach implementation to allow for the presence of two or more diagnostic tests, and to provide inferential procedures for comparing the accuracy of these tests. Also, the incorporation of additional optimal threshold criteria may constitute another issue to cover in the future (López-Ratón et al., 2014). Currently, estimates of the conditional AUC and Youden Index (and associated threshold values) are obtained by simply plugging-in an estimate for the conditional ROC curve in (3) and (5), respectively. Further work is warranted to implement direct estimators, such as those presented in Yao et al. (2010) and Xu et al. (2014).

The methods presented in this paper pave the way for further research efforts, discussed here. Firstly, the parametrisation used for the factor-by-curve interaction model allows us to evaluate the presence of an interaction component. If the result of the tests bring to reject the absence of interaction, it would be of great interest to study in which groups defined by the categorical covariate, the continuous covariate has an impact on the accuracy of the diagnostic test. In the data analyses presented in this paper, this question is answered by fitting a separate ROC-GAM model in fatty and dense tissue. We are currently working on alternative parametrisations that would permit us to find out which groups present a significant continuous covariate effect, without the need to fit separate models. Secondly, this paper focuses on testing for effects modelled by means of univariate nonparametric functions. The extension of both the estimation algorithm and the testing procedures to bivariate nonparametric functions (curve-by-curve interactions) represents an interesting line of research. Note that these extensions would allow incorporating (and testing) the interaction between continuous covariates and the FPF. However, as pointed out before, estimation in this case should ensure monotonicity in the FPF direction. Thirdly, in order to apply the ROC-GAM approach in practice, it would be interesting to know which covariates have a linear effect and which have a nonlinear effect. The graphical display of the estimates of each partial effect, jointly with the corresponding pointwise confidence intervals, can be used for that purpose. A nonlinear effect will be detected if it is not possible to plot a line inside the limits given by the confidence intervals. Otherwise the effect can be considered linear. However, since the confidence limits are only pointwise, this approach should be taken with caution. We believe that the results presented in this paper can be extended to propose bootstrap-based procedures for testing for nonlinearity effects. Finally, the generalisation of the methods presented in this paper to correlated data constitutes an area to be further explored.

Concerning the application of the results of this paper to the development of CAD systems, the possibility of analysing and evaluating covariate effects can help in the development of new algorithms for image feature extraction. The methods discussed allow for a deeper analysis of side effects in the behaviour of the algorithm, and this opens the possibility to propose alternative designs that would permit us to create more complex and useful algorithms.

9 Acknowledgements

This research was supported by the Spanish Ministry of Economy and Competitiveness MINECO grants MTM2014-55966-P, MTM2014-52975-C2-1-R, and BCAM Severo Ochoa excellence accreditation SEV-2013-0323, and by the Basque Government through the BERC 360 2014-2017.

A ROC-GAM estimation procedure

This appendix describes the estimation process associated with the ROC-GAM regression models (7) and (8). We present here the main steps of the algorithm, and refer the reader to Rodríguez-Álvarez et al. (2011a) for more details. More precisely, in Rodríguez-Álvarez et al. Rodríguez-Álvarez et al. (2011a) the algorithm proposed by Alonzo and Pepe (2002) for the estimation of ROC-GLMs was extended to allow for nonparametric covariate effects. The steps of the proposed procedure can be summarised as follows

Step 1. Choose a set of FPFs $P = \{p_l\}_{l=1}^{n_P} \subset (0,1)$ where the conditional ROC curve will be evaluated;

Step 2. Estimate $S_{\bar{D}}(\cdot \mid \boldsymbol{x})$, say $\widehat{S}_{\bar{D}}(\cdot \mid \boldsymbol{x})$, on the basis of the sample $\left\{\left(\boldsymbol{x}_i^{\bar{D}}, y_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$;

Step 3. For each observation in the diseased population, calculate the estimated 'placement value' (Hanley and Hajian-Tilaki, 1997) $\widehat{S}_{\bar{D}}\left(y_j^D \mid \boldsymbol{x}_j^D\right)$, $1 \leq j \leq n_D$;

Step 4. For each $p_l \in P$ and each disease observation, calculate the binary placement value indicator $\widehat{B}_{jp_l} = I\left(\widehat{S}_{\bar{D}}\left(y_j^D \mid \boldsymbol{x}_j^D\right) \leq p_l\right)$, $1 \leq l \leq n_P$, $1 \leq j \leq n_D$; and

Step 5. Fit the ROC-GAMs (7) or (8) to the data $\left\{\left(\left\{\boldsymbol{x}_j^D, p_l\right\}, \widehat{B}_{jp_l}\right), l = 1, \ldots, n_P, j = 1, \ldots, n_D\right\}$ and obtain the estimates $\widehat{ROC}_{\boldsymbol{x}}(p)$.

Note that in Step 5, the binary indicators, $\widehat{B}_{jp_l}$, are the response variable. This suggests the use of GAM estimation techniques for binary response data. Rodríguez-Álvarez et al. (2011a) proposed the use of the local scoring estimation algorithm with backfitting (Hastie and Tibshirani, 1990), and estimates of f_k, f^l and h_0 are obtained by applying local polynomial kernel smoothers (Fan and Gijbels, 1996). In the present paper, both for the simulations and the real data analyses, we use local-linear smoothers jointly with binning-type acceleration techniques to speed up computation (Fan and Marron, 1994). The optimal bandwidths are selected by means of cross validation.

As far as model (9) is concerned (involved in Step 2), nonparametric estimates of $\mu_{\bar{D}}(\cdot)$ and $\sigma_{\bar{D}}(\cdot)$, say $\widehat{\mu}_{\bar{D}}(\cdot)$ and $\widehat{\sigma}_{\bar{D}}(\cdot)$, are obtained by means of local-linear kernel smoothers and

the backfitting algorithm, and the cumulative survival function of the regression error $G_{\bar{D}}$ is estimated by the corresponding empirical cumulative survival function of the estimated residuals, i.e., $\widehat{G}_{\bar{D}}(c) = n_{\bar{D}}^{-1} \sum_{i=1}^{n_{\bar{D}}} I\left(\widehat{\varepsilon}_i^{\bar{D}} \geq c\right)$, where $\widehat{\varepsilon}_i^{\bar{D}} = \left(y_i^{\bar{D}} - \widehat{\mu}_{\bar{D}}\left(\boldsymbol{x}_i^{\bar{D}}\right)\right) / \widehat{\sigma}_{\bar{D}}\left(\boldsymbol{x}_i^{\bar{D}}\right)$, $i = 1, \ldots, n_{\bar{D}}$ (see Rodríguez-Álvarez et al., 2011a, for more details).

References

Alonzo, T. A. and M. S. Pepe (2002). Distibution-free ROC analysis using binary regression techniques. *Biostatistics 3*(3), 421–432.

Cai, T. (2004). Semiparametric ROC regression analysis with placement values. *Biostatistics 5*(1), 45–60.

Cai, T. and L. E. Dodd (2008). Regression analysis for the partial area under the ROC curve. *Stat Sinica 18*(3), 817–836.

Cai, T. and M. S. Pepe (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *J Am Stat Assoc 97*(460), 1099–1107.

Cai, T. and Y. Zheng (2007). Model checking for ROC regression analysis. *Biometrics 63*(1), 152–163.

Dodd, L. E. and M. S. Pepe (2003a). Partial AUC estimation and regression. *Biometrics 59*(3), 614–623.

Dodd, L. E. and M. S. Pepe (2003b). Semiparametric regression for the area under the receiver operating characteristic curve. *J Am Stat Assoc 98*(462), 409–417.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review current status and future potential. *Comput Med Imag Grap 31*, 198–211.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Fan, J. and J. S. Marron (1994). Fast implementations of nonparametric curve estimators. *J Comput Graph Stat 3*(1), 35–56.

Faraggi, D. (2003). Adjusting receiver operating curves and related indices for covariates. *Statistician 52*, 179–192.

González-Manteiga, W., J. C. Pardo-Fernández, and I. van Keilegom (2011). ROC curves in non-parametric location-scale regression models. *Scand J Stat 38*(1), 169–184.

Hanley, J. A. and K. O. Hajian-Tilaki (1997). Sampling variability of non-parametric estimates of the areas under receiver operating characteristic curves: An update. *Acad Radiol 4*, 49–58.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Inácio de Carvalho, V., M. de Carvalho, T. A. Alonzo, and W. González-Manteiga (2016, 09). Functional covariate-adjusted partial area under the specificity–ROC curve with an application to metabolic syndrome diagnosis. *Ann Appl Stat 10*(3), 1472–1495.

Inácio de Carvalho, V., A. Jara, T. E. Hanson, and M. de Carvalho (2013, 09). Bayesian nonparametric ROC regression modeling. *Bayesian Anal 8*(3), 623–646.

Janes, H. and M. S. Pepe (2008). Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. *Am J Epidemmiol 168*(1), 89–97.

Janes, H. and M. S. Pepe (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika 96*(2), 371–382.

Krzanowski, W. J. and D. J. Hand (2009). *ROC Curves for Continuous Data* (1st ed.). Chapman & Hall/CRC.

López-de-Ullibarri, I., R. Cao, C. Cadarso-Suárez, and M. J. Lado (2008). Nonparametric estimation of conditional ROC curves: Application to discrimination tasks in computerized detection of early breast cancer. *Comput Stat Data An 52*(5), 2623–2631.

López-Ratón, M., M. X. Rodríguez-Álvarez, C. Cadarso-Suárez, and F. Gude-Sampedro (2014). OptimalCutpoints: An R package for selecting optimal cutpoints in diagnostic tests. *J Stat Softw 61*(1), 1–36.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall / CRC.

Nishikawa, R. M. (2007). Current status and future directions of computer-aided diagnosis in mammography. *Comput Med Imag Grap 31*(4–5), 224–235.

Pardo-Fernández, J. C., M. X. Rodríguez-Álvarez, and I. van Keilegom (2014). A review on ROC curves in the presence of covariates. *REVSTAT-Stat J 12*(1), 21–41.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics 54*(1), 124–135.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Sciences Series. Oxford University Press, New York.

Pepe, M. S. and T. Cai (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics 60*(2), 528–535.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rodríguez, A. and J. C. Martínez (2013). Bayesian semiparametric estimation of covariate-dependent ROC curves. *Biostatistics 15*(2), 353–369.

Rodríguez-Álvarez, M. X., J. Roca-Pardiñas, and C. Cadarso-Suárez (2011a). A new flexible direct ROC regression model: Application to the detection of cardiovascular risk factors by anthropometric measures. *Comput Stat Data An 55*(12), 3257–3270.

Rodríguez-Álvarez, M. X., J. Roca-Pardiñas, and C. Cadarso-Suárez (2011b). ROC curve and covariates: extending induced methodology to the non-parametric framework. *Stat Comput 21*(4), 483–499.

Rodríguez-Álvarez, M. X., P. G. Tahoces, C. Cadarso-Suárez, and M. J. Lado (2011). Comparative study of ROC regression techniques – applications for the computer-aided diagnostic system in breast cancer detection. *Comput Stat Data An 55*(1), 888–902.

Sickles, E. A. (1984). Mammographic features of "early" breast cancer. *Am J Roentgenol 143*, 461–464.

Varela, C., P. G. Tahoces, A. J. Méndez, M. Souto, and J. J. Vidal (2007). Computerized detection of breast masses in digitized mammograms. *Comp in Bio and Med 37*(2), 214–226.

Xu, T., J. Wang, and Y. Fang (2014). A model-free estimation for the covariate-adjusted Youden index and its associated cut-point. *Stat Med 33*(28), 4963–4974.

Yao, F., R. V. Craiu, and B. Reiser (2010). Nonparametric covariate adjustment for receiver operating characteristic curves. *Can J Stat 38*(1), 27–46.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer 3*(1), 32–35.

Zheng, Y. and P. J. Heagerty (2004). Semiparametric estimation of timedependent ROC curves for longitudinal marker data. *Biostatistics 5*(4), 615–632.

Zou, X.-H., N. A. Obuchowski, and D. K. McClish (2002). *Statistical Methods in Diagnostic Medicine*. Wiley: New York.