

Bootstrap-based procedures for inference in nonparametric ROC regression analysis

María Xosé Rodríguez-Álvarez^{1,2}, Javier Roca-Pardiñas¹,
Carmen Cadarso-Suárez³ and Pablo G. Tahoces⁴

¹ Department of Statistics and Operations Research
and Biomedical Research Centre (CINBIO), University of Vigo
Campus Lagoas-Marcosende s/n, 36310 Vigo, Spain.

mxrodriguez@uvigo.es

² BCAM - Basque Center for Applied Mathematics, Bilbao, Spain

³ Center for Research in Molecular Medicine and Chronic Diseases (CiMUS),
University of Santiago de Compostela, Santiago de Compostela, Spain

⁴ Centro Singular de Investigación en Tecnologías de la Información (CITIUS)
University of Santiago de Compostela, Santiago de Compostela, Spain

August 20, 2016

Abstract

Before the use of a diagnostic test in a routine clinical setting, the rigorous evaluation of its diagnostic accuracy is an essential step. 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 needs to be also assessed. In this paper, attention is focused on an estimator for the covariate-specific ROC curve based on direct regression modelling and nonparametric smoothing techniques. This approach defines the class of generalized 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 over 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 computed-aided diagnostic (CAD) system for the automatic detection of tumour masses in breast cancer is analysed.

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 indicating 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. $TPF = P(Y = 1 | D = 1)$, whereas the FPF or '1-specificity' is the probability of a healthy individual being falsely classified as diseased, i.e. $FPF = P(Y = 1 | D = 0)$.

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, Zhou, Obuchowski and McClish 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, i.e. it relies on all possible transformations 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 $TPF(c) = P(Y \geq c | D = 1)$ and $FPF(c) = P(Y \geq c | D = 0)$. When the diagnostic test Y is continuous, the ROC curve is usually represented as $ROC(p) = S_D(S_D^{-1}(p))$ for $0 \leq p \leq 1$, where $S_D(c) = P(Y \geq c | D = 1)$ and $S_{\bar{D}}(c) = P(Y \geq c | D = 0)$.

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 \mathbf{X} the d -variate vector of covariates we are interested in, the conditional or covariate-specific ROC curve, given a covariate value \mathbf{x} , is

defined as

$$ROC_{\mathbf{x}}(p) = S_D \left(S_D^{-1}(p | \mathbf{x}) | \mathbf{x} \right), \quad 0 \leq p \leq 1, \quad (1)$$

where, by a slight abuse of notation, $S_D(c | \mathbf{x}) = P(Y \geq c | D = 1, \mathbf{X} = \mathbf{x})$ and $S_{\bar{D}}(c | \mathbf{x}) = P(Y \geq c | D = 0, \mathbf{X} = \mathbf{x})$. Note that a continuum of different ROC curves (and, therefore, a continuum of different diagnostic accuracies) is obtained by varying the value \mathbf{x} in the range of \mathbf{X} . As a consequence, the conditional ROC curve can be viewed as a tool which helps identifying 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. \(2011c\)](#), whereas the paper by [Pardo-Fernández et al. \(2014\)](#) mainly focuses on the nonparametric counterparts. In brief, there are two main strategies to approach estimation: one based on estimating the conditional cumulative survival functions involved in the definition given in (1) ([Inácio de Carvalho et al. 2013](#), [López-de-Ullibarri et al. 2008](#)); 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 and Martínez 2014](#), [Rodríguez-Álvarez et al. 2011a](#), [Yao et al. 2010](#), [Zheng and Heagerty 2004](#)). On the other hand, direct methodology directly regresses the ROC curve. This methodology has yielded to (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 generalized 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. 2011b](#)), along the line of the generalized additive model (GAM, [Hastie and Tibshirani 1990](#)).

The aims of this paper are 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 covariates' effects. In spite of its importance, to the best of our knowledge this topic has received little attention in the statistical literature, especially in the non-parametric 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. \(2011b\)](#), 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 framed within the ROC-GAM context.

Secondly, nowadays there is an undeniable need for software developments 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 2016) 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. (2011a) and Rodríguez-Álvarez et al. (2011b) respectively, 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 present in Section 5. Section 6 describes the `npROCRegression` R-package. We illustrate our approach and the usage of the package in Section 7 using data from a computed-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. (2011c) and Pardo-Fernández et al. (2014) for a more in-deep survey. However, with this section we aim at setting in a context different modelling strategies for the incorporation of covariates on the ROC curve. More precisely, we focus here on those that can be viewed as coming within the general framework of regression, namely ‘induced’ and ‘direct’ methodologies. A qualitative comparison of both approaches is also presented. The section ends up presenting 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

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

where, for $j \in \{\bar{D}, D\}$, $\mu_j(\mathbf{x}) = E(Y_j \mid \mathbf{X} = \mathbf{x})$ and $\sigma_j^2(\mathbf{x}) = Var(Y_j \mid \mathbf{X} = \mathbf{x})$ are the conditional mean and the conditional variance of Y_j given $\mathbf{X} = \mathbf{x}$, respectively. The error ε_j is assumed independent of the covariate \mathbf{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 between the errors and the covariates, it follows that

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

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 on 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. \(2011a\)](#) and [Rodríguez and Martínez \(2014\)](#). 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 \(2014\)](#), 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 .

2.1.2 Direct ROC regression methodology.

In contrast to induced, 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

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

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 \mathbf{X} on the ROC curve is incorporated parametrically, i.e., $\mu(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}$. Thus, models such as (2) define the so-called class of ROC-GLMs ([Pepe 2003](#)), due to their resemblance to generalized linear models (GLMs). To the extent of our knowledge, up to date, only the paper by [Rodríguez-Álvarez et al. \(2011b\)](#) 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 seen, 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 relies on modelling covariate effects on the result of the diagnostic test. Even though, on the one hand, it implies that the modelling of covariate effects on the ROC curve is indirect, on the other hand (a) it allows 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 that on the ROC curve, which can help understanding and explaining the covariate impact on the accuracy of the test. Secondly, for ease of reading, throughout all of our presentation, covariates affecting the test results in healthy and diseased population were assumed to be the same while this is not necessarily the case in practice (e.g., disease stage). Induced methodology allows incorporating specific covariates of healthy or diseased populations, or even both. Lastly, note that 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, if the covariate vector \mathbf{X} is multidimensional (and no restrictions are imposed), heteroskedasticity would also imply that the interaction among

all covariates (and possibly the FPF) is to be implicitly incorporated into the “model” for the conditional ROC curve. This can make the interpretation and visualization 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 above, 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 incorporating disease-specific covariates. It does not, however, permit health-related information.

2.2 Conditional summary statistics

It is common to summarize 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 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

$$AUC_{\mathbf{x}} = \int_0^1 ROC_{\mathbf{x}}(p) dp. \quad (3)$$

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

The most obvious way to estimate the conditional AUC is to just 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 $AUC_{\mathbf{x}}$ have been proposed in the literature. For instance, Faraggi (2003) discusses a fully parametric estimation approach based on induced modelling. In Dodd and Pepe (2003a;b) 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 $AUC_{\mathbf{x}}$ estimation. This method has been recently extended to functional covariates by Inácio de Carvalho et al. (2016).

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

$$\begin{aligned} YI_{\mathbf{x}} &= \max_{c_{\mathbf{x}}} \{TPF(c_{\mathbf{x}} | \mathbf{x}) - FPF(c_{\mathbf{x}} | \mathbf{x})\} \\ &= \max_{c_{\mathbf{x}}} \{S_D(c_{\mathbf{x}} | \mathbf{x}) - S_{\bar{D}}(c_{\mathbf{x}} | \mathbf{x})\} \end{aligned} \quad (4)$$

$$= \max_{p_{\mathbf{x}}} \{ROC_{\mathbf{x}}(p_{\mathbf{x}}) - p_{\mathbf{x}}\}, \quad (5)$$

where we use the notation $c_{\mathbf{x}}$ and $p_{\mathbf{x}}$ to emphasize that these values depend on covariate \mathbf{x} . The $YI_{\mathbf{x}}$ takes values between 0.5, in the case of an uninformative test, and 1.0 in a perfect test. The value $c_{\mathbf{x}}^*$ which maximises (4) is frequently used in practice as a threshold value to separate diseased from healthy status (in those individuals with covariate value \mathbf{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 but 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

$$AROC(p) = \int ROC_{\mathbf{x}}(p) dH_D(\mathbf{x}), \quad (6)$$

where $H_D(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x} | D = 1)$. 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 imply that the conditional ROC curve (which in this case is common to all covariate values) coincides with ROC curve obtained on pooling the data regardless of the values of the covariates. It coincides, however, with the AROC curve (see Janes and Pepe 2009, Pardo-Fernández et al. 2014; for more details). As a consequence, even in those situations where the accuracy of a test does not vary along 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. (2011a), in the context of nonparametric induced modelling approaches.

3 The ROC-GAM regression model

As said 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 generalized additive model. Specifically, the ROC-GAM regression model is expressed as

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

where \mathbf{x}_u and \mathbf{x}_v denote subsets of the covariate vector \mathbf{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 $E(f_k(X_{kv})) = 0$ ($k = 1, \dots, V$) and $\int_0^1 h_0(p) dp = 0$.

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

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

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

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

jointly with

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

Note that, given the previous constraints, model (8) has been parametrized so it is *hierarchical*. As a consequence, f is the smooth main effect of covariate X_v , and thus f^l ($l = 1, \dots, M$) 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. \(2011b\)](#) 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, $S_{\bar{D}}(\cdot | \mathbf{x})$ (Step 2). For that purpose, Rodríguez-Álvarez et al. (2011b) suggest to model the effect of covariates on $Y_{\bar{D}}$ by a nonparametric location-scale regression model

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

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

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

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 our interest 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, that 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

$$H_0 : f^1(x_v) = \dots = f^M(x_v) = 0,$$

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 interactions curves f^l ($l = 1, \dots, M$).

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

4.1 Testing for continuous covariate effect

The test for the null hypothesis

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

versus the general hypothesis

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

is based on the estimate \hat{f}_r . For this purpose, L_1 and L_2 norms are considered yielding the following test statistics

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

Note that the proposed statistics are measures of deviations from the estimated nonparametric function \hat{f}_r to its mean $n_D^{-1} \sum_{j=1}^{n_D} \hat{f}_r(x_{jvr}^D)$ 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 $P(T > T^o | H_0) < \alpha$, 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 $\left\{ \left(\mathbf{x}_i^{\bar{D}}, y_i^{\bar{D}} \right) \right\}_{i=1}^{n_{\bar{D}}}$ as explained in Appendix A. Let $\hat{\mu}_{\bar{D}}(\cdot)$, $\hat{\sigma}_{\bar{D}}(\cdot)$, and $\hat{G}_{\bar{D}}(\cdot)$ be these estimates.

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

For $b = 1, \dots, B$

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

$$y_{i,b}^{\bar{D}*} = \hat{\mu}_{\bar{D}}(\mathbf{x}_i^{\bar{D}}) + \hat{\sigma}_{\bar{D}}(\mathbf{x}_i^{\bar{D}}) \varepsilon_{i,b}^{\bar{D}*}, \quad (11)$$

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

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

Step 4. From $\{(\mathbf{x}_i^{\bar{D}}, y_{i,b}^{\bar{D}^*})\}_{i=1}^{n_{\bar{D}}}$ and $\{(\mathbf{x}_j^D, y_{j,b}^{D*})\}_{j=1}^{n_D}$ obtain T^b (T_{\parallel}^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_{\parallel} or T_2) under H_0 . Thus, the test rule based on T (T_{\parallel} or T_2) consists of rejecting the null hypothesis if $T^o > T_{\alpha}^B$ where T_{α}^B is the empirical $(1-\alpha)$ -percentile of the values of T^1, \dots, T^B obtained in Step 4.

4.2 Testing for factor-by-curve interaction

The test for the null hypothesis

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

versus the general hypothesis

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

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

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

As before, the proposed statistics are measures of deviations from the estimated nonparametric interaction curves \hat{f}^l ($l = 1, \dots, M$) to their mean $n_D^{-1} \sum_{j=1}^{n_D} \hat{f}^l(x_{jv}^D) I(x_{ju}^D = l)$ which 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 the Step 2 of algorithm, which now must be

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

4.3 Resampling under the null hypothesis

As said, 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 meet this requirement.

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

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

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

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

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

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

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

Finally, it is worth emphasizing 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:

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)).
2. 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, in (14) the conditional ROC curve under the null hypothesis (see models (10) and (13)) is put in place of $ROC_{\mathbf{x}}(\cdot)$.

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. Data were simulated from three different scenarios, namely,

Scenario I

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

Scenario II

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

Scenario III

$$\begin{aligned} Y_{\bar{D}} &= -0.25X_{v1}^3 + 0.5X_{v1}^2 + 0.5X_{v1}^2 X_{u1} - 0.5X_{v1}^2(1 - X_{u1}) + 0.5 \varepsilon_{\bar{D}}, \\ Y_D &= 0.25X_{v1}^3 + (a + 1) (.5X_{v1}^2 + .5X_{v1}^2 X_{u1} - 0.5X_{v1}^2(1 - X_{u1})) + 0.5 \varepsilon_D. \end{aligned}$$

In all cases, a is a real constant, X_{v1} was 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 was simulated from a uniform distribution on $[-1, 1]$, and Scenario III represents the factor-by-curve case. Here $X_{u1} \sim \text{Bernoulli}(0.5)$. We note that Scenario I was also considered in [Rodríguez-Álvarez et al. \(2011a\)](#).

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

Scenario I

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

Scenario II

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

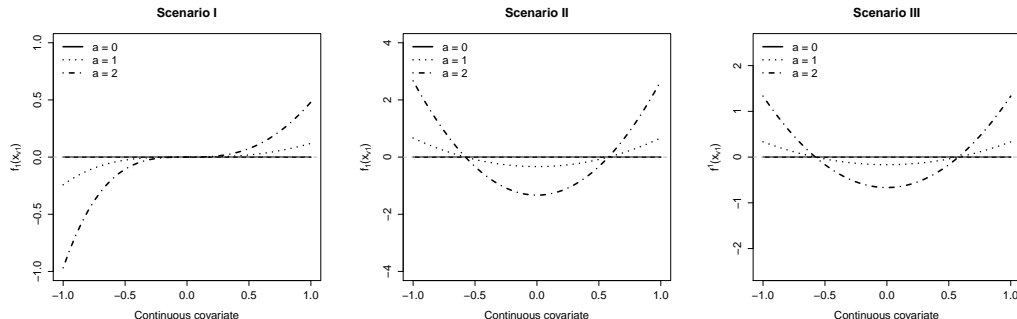


Figure 1: For Scenario I and II: centred nonparametric function of X_{v1} . For Scenario III: centred interaction curve of X_{v1} for $X_{u1} = 1$. In all cases, the partial functions are shown for different values of a . The dotted grey line represents the null hypothesis (no effect/interaction), which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

Scenario III

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

where Φ denotes the cumulative distribution function of a standard normal random variable.

In order to fit a ROC-GAM regression model, several choices need to be made. In all results shown below, the probit function, namely $g^{-1} = \Phi^{-1}$, was 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 were considered. We should note that our implementation makes use of binning type acceleration techniques (Fan and Marron 1994) to reduce computational time. In this study, we used 30 equally spaced binning points along the range of each continuous covariates. The bandwidths involved in the local-linear kernel smoothers were 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 were considered for a , ranging from 0 to 2. 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 Scenario III, 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 three 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 Scenario I.

The bootstrap procedure described in Section 4.1 was applied to Scenarios I and II, and that presented in Section 4.2 to Scenario III. In all cases, critical values and p -values were

determined using $B = 400$ bootstrap samples. Both type I errors and powers were calculated as the proportion of rejections of H_0 in 1000 runs, and the same sample size was considered for both healthy and diseased subjects, with $n = n_D = n_{\bar{D}} = 50, 100, 200, 500, 1000$. For the sake of brevity, only the results for $n = 50, 200, 1000$ are shown below. Results for the remaining sample sizes are in concordance 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 was 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 performed 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). Note that there were only two situations where the KS test rejected the null hypothesis of a uniform distribution: for Scenario I and a sample size of $n = 50$, and for Scenario III and a sample size of $n = 1000$. We were especially concerned about the result for Scenario III with such a large sample size. Accordingly, we evaluated the behaviour of the tests (under the same conditions) for a sample size of $n = 2000$. In this case (results not shown), the KS test gave a p -value of 0.223 and 0.471 for $S_{||}$ and S_2 tests, respectively.

Power as a function of constant a and different significance levels is shown in Tables 2 and 3, and Figure 3 shows the power curves at 0.05 significance level. All tests registered satisfactory power curves, with the probability of rejection rising in response to any increase in the value of the constant a and the sample size. Note that the power curves also depict the expected behaviour according to the plots shown in in Figure 1.

6 Software implementation: the npROCRegression package

This section contains a brief description of the developed R-package that accompany this paper. The package can be freely downloaded from <https://cran.r-project.org/package=npROCRegression>, where a more detailed depiction 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. 2011b) and induced (Rodríguez-Álvarez et al. 2011a) regression approaches. Numerical and graphical summaries of the fitted models can be obtained by calling the functions print(), summary() and plot().

	Sample size	Test	Level					KS p -value
			0.01	0.05	0.10	0.15	0.20	
Scenario I	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
	200	$T_{ }$	0.008	0.056	0.119	0.163	0.207	0.683
		T_2	0.011	0.057	0.109	0.161	0.204	0.939
	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	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
	200	$T_{ }$	0.020	0.051	0.114	0.161	0.205	0.482
		T_2	0.015	0.060	0.113	0.167	0.209	0.736
	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	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
	200	$S_{ }$	0.024	0.061	0.102	0.149	0.191	0.217
		S_2	0.021	0.061	0.102	0.148	0.192	0.170
	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

Table 1: For Scenarios I, II and III: 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.

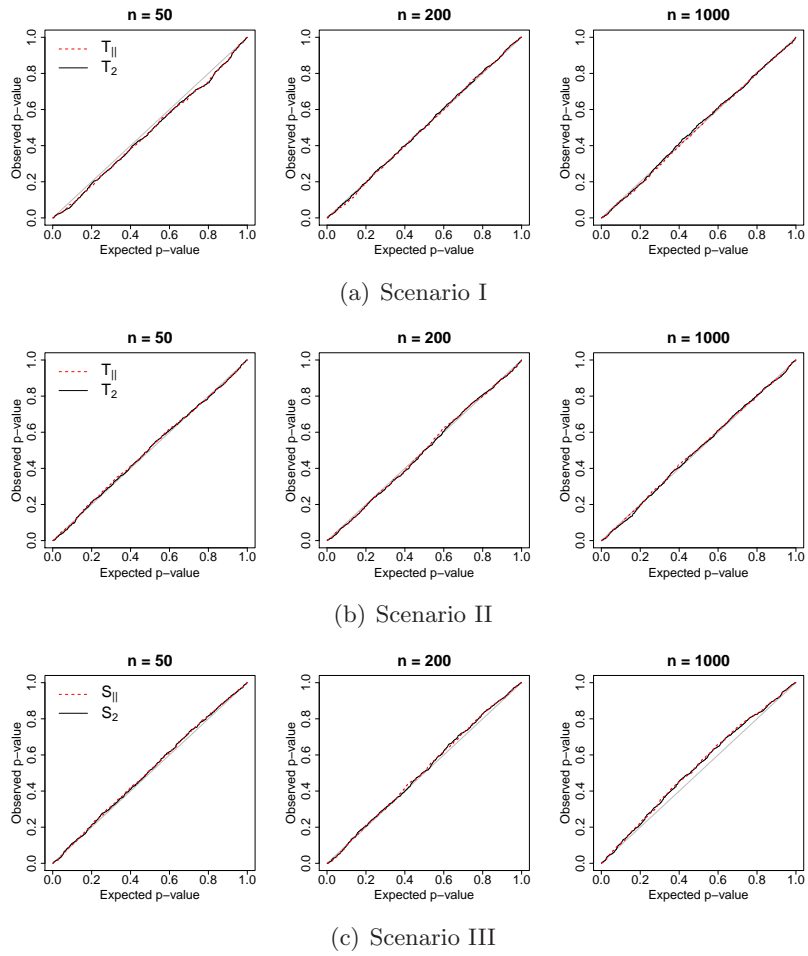


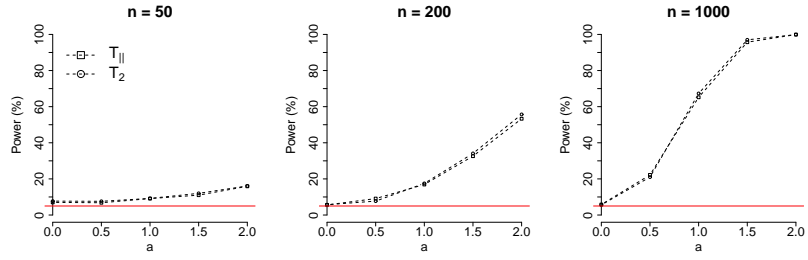
Figure 2: For Scenarios I, II and III: Quantile-quantile plot for the expected p-values vs. the observed p-values when the null hypothesis is correct.

			Level					
	Sample size	Test	0.01	0.05	0.10	0.15	0.20	
Scenario I	$a = 0.5$	50	$T_{ }$	0.023	0.068	0.133	0.184	0.227
			T_2	0.021	0.076	0.131	0.180	0.223
		200	$T_{ }$	0.024	0.091	0.159	0.212	0.260
			T_2	0.024	0.077	0.145	0.211	0.265
		1000	$T_{ }$	0.097	0.222	0.318	0.392	0.475
			T_2	0.091	0.210	0.304	0.396	0.467
	$a = 1.0$	50	$T_{ }$	0.027	0.091	0.151	0.208	0.269
			T_2	0.034	0.092	0.158	0.209	0.265
		200	$T_{ }$	0.070	0.169	0.251	0.339	0.403
			T_2	0.065	0.175	0.261	0.333	0.396
		1000	$T_{ }$	0.417	0.652	0.764	0.832	0.874
			T_2	0.468	0.673	0.779	0.834	0.875
$a = 2.0$	50	$T_{ }$	0.060	0.159	0.241	0.311	0.382	
		T_2	0.074	0.161	0.247	0.318	0.381	
	200	$T_{ }$	0.305	0.533	0.652	0.711	0.770	
		T_2	0.346	0.557	0.668	0.726	0.777	
	1000	$T_{ }$	0.990	0.999	1.000	1.000	1.000	
		T_2	0.997	0.999	1.000	1.000	1.000	
Scenario II	$a = 0.5$	50	$T_{ }$	0.037	0.097	0.176	0.226	0.268
			T_2	0.032	0.097	0.164	0.225	0.280
		200	$T_{ }$	0.258	0.447	0.566	0.635	0.697
			T_2	0.225	0.435	0.548	0.619	0.672
		1000	$T_{ }$	0.993	0.996	0.998	0.998	0.999
			T_2	0.990	0.996	0.998	0.998	0.999
	$a = 1.0$	50	$T_{ }$	0.142	0.305	0.398	0.473	0.540
			T_2	0.126	0.293	0.390	0.456	0.512
		200	$T_{ }$	0.951	0.982	0.992	0.995	0.997
			T_2	0.928	0.979	0.985	0.993	0.995
		1000	$T_{ }$	1.000	1.000	1.000	1.000	1.000
			T_2	1.000	1.000	1.000	1.000	1.000
$a = 2.0$	50	$T_{ }$	0.574	0.783	0.867	0.920	0.942	
		T_2	0.534	0.751	0.841	0.904	0.933	
	200	$T_{ }$	1.000	1.000	1.000	1.000	1.000	
		T_2	1.000	1.000	1.000	1.000	1.000	
	1000	$T_{ }$	1.000	1.000	1.000	1.000	1.000	
		T_2	1.000	1.000	1.000	1.000	1.000	

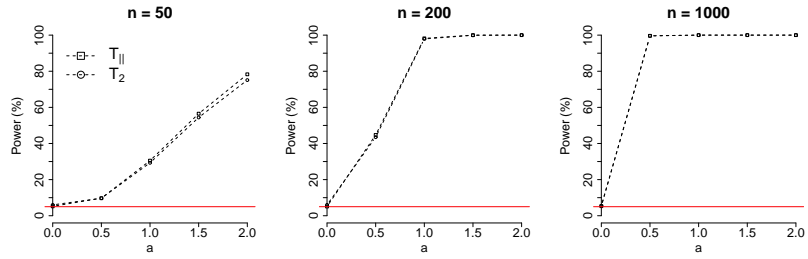
Table 2: For Scenarios I and II: estimated rejection probabilities registered by the proposed tests under the alternative hypothesis for different values of a , significance levels and sample sizes.

		Level					
	Sample size	Test	0.01	0.05	0.10	0.15	0.20
$a = 0.5$	50	$S_{ }$	0.020	0.081	0.146	0.207	0.257
		S_2	0.020	0.083	0.143	0.203	0.257
	200	$S_{ }$	0.112	0.211	0.307	0.391	0.458
		S_2	0.092	0.214	0.300	0.386	0.459
	1000	$S_{ }$	0.567	0.769	0.853	0.902	0.929
		S_2	0.527	0.753	0.848	0.890	0.917
$a = 1.0$	50	$S_{ }$	0.052	0.130	0.210	0.284	0.358
		S_2	0.050	0.127	0.208	0.276	0.347
	200	$S_{ }$	0.392	0.606	0.712	0.766	0.807
		S_2	0.373	0.585	0.701	0.753	0.799
	1000	$S_{ }$	0.998	1.000	1.000	1.000	1.000
		S_2	0.998	0.999	1.000	1.000	1.000
$a = 2.0$	50	$S_{ }$	0.111	0.270	0.372	0.464	0.527
		S_2	0.095	0.235	0.361	0.438	0.518
	200	$S_{ }$	0.924	0.976	0.988	0.993	0.994
		S_2	0.894	0.969	0.986	0.991	0.992
	1000	$S_{ }$	1.000	1.000	1.000	1.000	1.000
		S_2	1.000	1.000	1.000	1.000	1.000

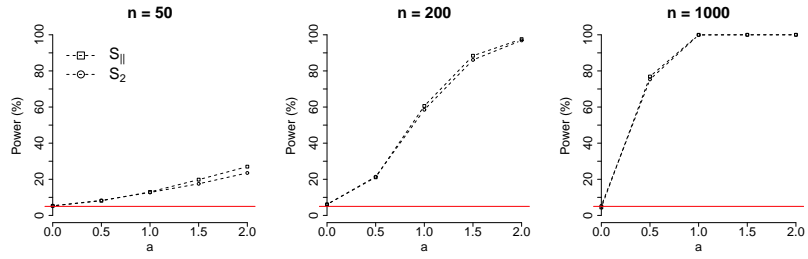
Table 3: For Scenario III: estimated rejection probabilities registered by the proposed tests under the alternative hypothesis for different values of a , significance levels and sample sizes.



(a) Scenario I



(b) Scenario II



(c) Scenario III

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

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 above. 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 through argument `marker`. The nonparametric location-scale regression model for healthy population (see (9)) is specified through `formula.h`. This argument should be as a vector (of length 2) of right-hand formulas (atomic values are also valid, being recycled). The first right-hand formula is the model for the regression function, $\mu_{\bar{D}}(\mathbf{x})$, and the second one is the model for the (logarithm) of the variance function, $\sigma_{\bar{D}}^2(\mathbf{x})$. 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 `~ 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 discussed in Section 3. By way of example, to include the interaction between `age` and `gender` we need to specify `~ 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 into the formula. All 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 specified in argument `group`, and in `tag.healthy` the value codifying the healthy individuals in this variable. The `data` argument is a data frame representing the data and containing all needed variables.

Pointwise bootstrap confidence intervals for each component of the additive predictor of the ROC-GAM, as well as the conditional AUCs (with the integral being 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 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 well as the link function or the number of bootstrap resamples and 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 based on the induced nonparametric ROC regression approach as presented in [Rodríguez-Álvarez et al. \(2011a\)](#). The call to the function 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 have to 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 representing the data and containing 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 by means of the test presented in [Rodríguez-Álvarez et al. \(2011a\)](#).

By default, the `INPROCreg()` function returns the estimated regression and variance functions both in healthy and diseased populations. As far as accuracy measures is concerned, the function provides the estimated conditional ROC curve, the associated conditional AUCs (with the integral being 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 these 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. We should note that, even in this case, covariate-specific thresholds will be obtained ([Rodríguez-Álvarez et al. 2011a](#)).

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`. For instance, this argument can be used to select the number of binning points or the order of the polynomial associated to the kernel smoothers.

7 Application to a CAD system

Computer aided diagnosis has been defined as the diagnosis made by a radiologist who takes into account the results of quantitative computer analysis of medical diagnosis (Doi 2007). Those kind of systems have demonstrated its usefulness in situations where the radiologists have to discriminate positive cases among hundreds or thousands of normal cases. Those screening programs suppose a challenge for the physicians and the presence of a second reader in form of computer algorithm has demonstrated to be useful (Nishikawa 2007). The base of a CAD system is a computer vision algorithm where several parameters extracted from some features particularly selected from the image, are combined by means of a classifier specially trained for discriminating between normal and abnormal cases. CAD schemes have been developed for screening programs related to breast, chest, colon, etc.

Quality in terms of image can be described in terms of three basic concepts: contrast, spatial resolution and noise. For x-ray images the subjective contrast (the contrast measured before the capture of the image by the recording system) is modelled by three main factors: the energy of the photons that constitute the x-ray beam, the thickness of the object to visualize and the linear attenuation coefficient of the material that compose the object, which is related with its density.

The breast is mainly made up of fat and glandular tissue. The attenuation coefficient of fat and glandular tissue is quite different, having glandular tissue a higher attenuation coefficient. To achieve mammographies of good quality in terms of contrast, the whole breast is physically compressed, and x-ray photons of low energy are employed. But still, the volume of the breast and its composition differ from breast to breast. Thus, the contrast and even the average grey level (AGL) 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 cancers, become more difficult.

For breast cancer detection, the identification of masses and microcalcifications constitutes the main issue. Microcalcifications are small radiopacities of high contrast, whereas masses are radiopacities of median size and low contrast. The main problem for mass detection is the presence of several structures related with glandular tissue that has similar attenuation coefficient as masses. In some cases the presence of such structures is massive

(dense breast) hiding the presence of masses. By the contrary, when the presence of glandular tissue is negligible, masses are clearly visualize from the background compose mainly of fat and its detection become relatively easy.

One important issue in the development of a CAD scheme is to analyse the impact that a particular variable has over the performance of the whole system. For breast cancer detection, its behaviour with respect to the presence of glandular tissue is also an essential aspect. To this aim, in this paper we have analysed this effect over a CAD scheme for breast masses detection previously developed (see [Varela et al. 2007](#); for a detailed description). The system is based on the extraction of several image features that can be grouped into three main categories: contour related, texture and tone. Among the texture features, the variables related with the iris filter have a special importance in terms of the performance of the system. The iris filter is an algorithm specially designed to enhance round brilliant structures within an image. Since a mass has this appearance in a mammogram, the variables related with that feature are of special interest.

7.1 Data set

The database contain 580 mammograms where a total of 190 images were classified as abnormal (lesion present), and the remaining 390 as normal images (no lesion present). From the 580 original mammograms, a total of 2796 regions suspicious of being a malignant mass were detected by the computer in a first step. Of these, 384 corresponded to true masses, and the remainder, a total of 2412, corresponded to false detections.

7.2 Data analysis

One of the main purposes of this study on CAD systems was to assess the effects of some covariates as 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. \(2011b\)](#) 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 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 suggested in [Section 4](#) are used to check formally the possible effects on the ROC curve of the covariate AGL and the tissue-by-AGL interaction.

As a first step of the analysis, the discriminatory capacity of the iris filter was evaluated without taking into account the effect of the covariates. The AUC value corresponding to the pooled ROC curve was 0.69. ROC analysis was also performed for dense and fatty tissues separately, yielding pooled AUCs of 0.64 and 0.75, respectively. Additionally, we also estimated the AGL-adjusted ROC curves (AROC), both for dense and fatty tissues.

The areas under these AROC curves, were in this case 0.60 and 0.73. Note that they are slightly lower than the pooled AUCs, possibly indicating that the pooled analyses ‘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 than for dense tissue.

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

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

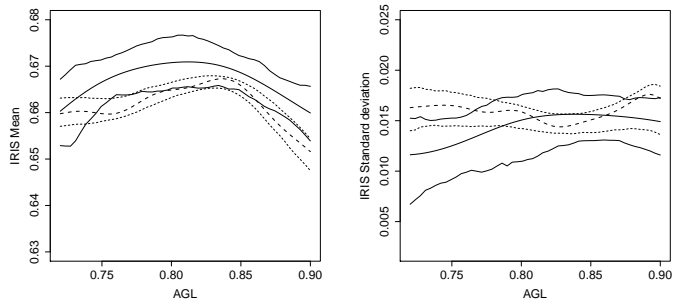
Figure 4 depicts the estimated effect of AGL both on the mean and standard deviation of filter output, according to breast tissue type, along with 95% pointwise confidence intervals. For masses, in both fatty and dense tissues, mean values rise to a peak approximately midway through the interval and 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 find on its diagnostic accuracy. Second, discrimination based on iris output is much more complex in dense 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 also seem to indicate the existence of a possible interaction between average grey levels and tissue type. Accordingly, we fitted a ROC-GAM regression model including such interaction

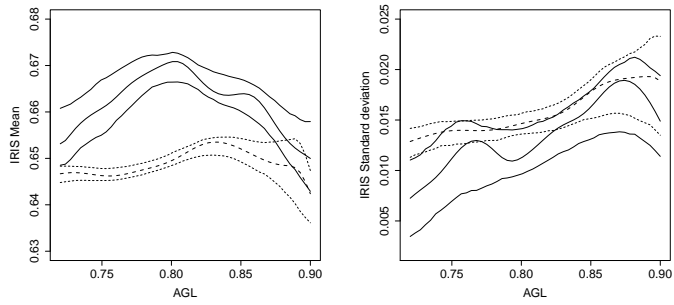
$$\text{ROC}_{\text{AGL},\text{TIS}}(p) = g \left(\beta_0 + \sum_{l=1}^2 \beta_l I(\text{TIS} = l) + f(\text{AGL}) + \sum_{l=1}^2 f^l(\text{AGL}) I(\text{TIS} = l) + h_0(p) \right). \tag{16}$$

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

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. We would like to point out that, the estimated AUCs obtained using the induced approach were quite similar to those depicted in Figure 6. For clarity in the presentation, these AUC estimates are not shown. As can be seen in Figure 6,



(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.

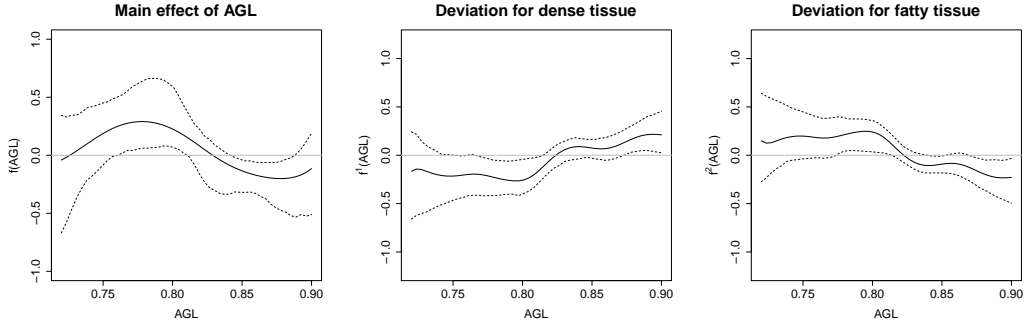


Figure 5: Estimated main effect of AGL in IRIS's accuracy, and deviation for dense and fatty tissue, together with 95% pointwise bootstrap confidence intervals.

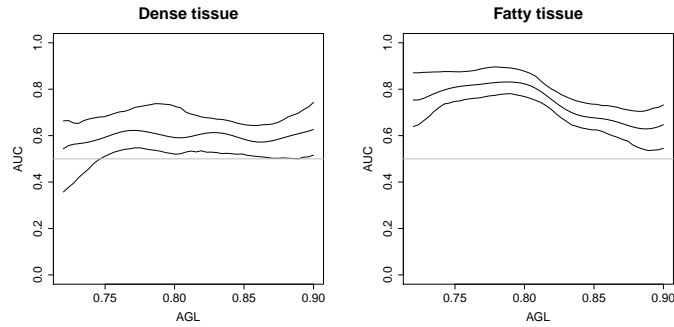


Figure 6: Estimated conditional AUC for the CAD system according to AGL and type of tissue. The dashed lines represent the 95% pointwise bootstrap confidence intervals.

the behaviour of 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 in average, pixel values for fatty breast are relatively low. When the average pixel value rises it is probably because the overall contrast of the whole breast decreases, due to size of the breast, the energy of the x-ray beam or both. In any case, the quality of the image get worse, and the results achieved by the filter are not as good. By the contrary, for dense breast, results are almost similar along the range of the pixel values. In this case, the presence of structures related with the glandular tissue makes the iris filter do not works 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 the previous results suggest that the presence of interaction between the type of tissue and the average grey levels might be plausible. For the purpose of statistically

verifying this hypothesis, we performed the tests presented in Section 4.2. The resulting p -values were 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 $S_{||}$ was lower than 0.05, thus rendering the interaction term significant.

Although the results of both tests could lead to different conclusions, based on all results presented we are 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 was to check for the effect of AGL on the accuracy of IRIS. To that aim, the following ROC-GAM model was fitted separately in fatty and dense tissues

$$ROC_{AGL}(p) = g(\beta_0 + f(AGL) + h_0(p)),$$

and we performed the tests for continuous covariate effects outlined in Section 4.1 above. Whereas the p -value obtained in the case of dense tissue was 0.54, in the case of fatty tissue was < 0.001 . So we can conclude that the accuracy of IRIS is constant along AGL in the dense tissue while in 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 above is now given. For the nonparametric induced approach presented in (15), the following code was 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 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, 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 test statistics to evaluate effects of continuous covariate 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 sample size increases and as one moves further away from the null hypothesis. The behaviour of the tests can be considered satisfactory.

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 in

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, it could be useful to extend the package, and more precisely 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 important 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 also ask for further research efforts. On the one hand, the parametrization used for the factor-by-curve interaction model allows evaluating the presence of such 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 parametrizations that would permit to find out which groups present a significant continuous covariate effect, but without the need to fit separate models. On the other hand, 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.

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 of image feature extraction. Furthermore, the possibility of introducing a more deeper analysis of side effects in the behaviour of the algorithm, such as the interactions between covariates or the introduction of a factor, open the possibility to propose design alternatives that would allow to create more complex and useful algorithms.

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. The first author thanks the Agrupamento INBIOMED from

A ROC-GAM estimation procedure

This appendix describes the estimation process associated to 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. \(2011b\)](#) for more details. More precisely, in [Rodríguez-Álvarez et al. \(2011b\)](#) 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 proposed procedure can be summarized 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 $S_{\bar{D}}(\cdot | \mathbf{x})$, say $\widehat{S}_{\bar{D}}(\cdot | \mathbf{x})$, on the basis of the sample $\left\{ \left(\mathbf{x}_i^{\bar{D}}, y_i^{\bar{D}} \right) \right\}_{i=1}^{n_{\bar{D}}}$;

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

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

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

Note that in Step 5 above the binary indicators, \widehat{B}_{jp_l} , are the response variable. This suggests the use of GAM estimation techniques for binary response data. In [Rodríguez-Álvarez et al. \(2011b\)](#) it is 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 used local-linear smoothers jointly with binning-type acceleration techniques to speed up computation ([Fan and Marron 1994](#)). The optimal bandwidths were 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 $\widehat{\varepsilon}_i^{\bar{D}} = \left(y_i^{\bar{D}} - \widehat{\mu}_{\bar{D}} \left(\mathbf{x}_i^{\bar{D}} \right) \right) / \widehat{\sigma}_{\bar{D}} \left(\mathbf{x}_i^{\bar{D}} \right)$, $i = 1, \dots, n_{\bar{D}}$ (see [Rodríguez-Álvarez et al. 2011b](#); for more details).

References

- Alonzo TA and Pepe MS. Distribution-free ROC analysis using binary regression techniques. *Biostatistics* 2002; **3**: 421–432.
- Cai T. Semi-parametric ROC regression analysis with placement values. *Biostatistics* 2004; **5**: 45–60.
- Cai T and Dodd LE. Regression analysis for the partial area under the ROC curve. *Statistical Sinica* 2008; **97**: 817–836.
- Cai T and Pepe MS. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* 2002; **97**: 1099–1107.
- Cai T and Zheng Y. Model checking for ROC regression analysis. *Biometrics* 2007; **63**: 152–163.
- Dodd LE and Pepe MS. Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association* 2003a; **98**: 409 – 417.
- Dodd LE and Pepe MS. Partial AUC estimation and regression. *Biometrics* 2003b; **59**: 614–623.
- Doi K. Computer-aided diagnosis in medical imaging: Historical review current status and future potential. *Computerized Medical Imaging and Graphics* 2007, **31**: 198–211.
- Fan J and Gijbels I. *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1996.
- Fan J and Marron JS. Fast implementation of non-parametric curve estimators. *Journal of Computational and Graphical Statistics* 1994; **3**: 35–56.
- Faraggi D. Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician* 2003; **52**: 179–192.
- González-Manteiga W, Pardo Fernández JC and van Keilegom I. ROC curves in nonparametric location-scale regression models. *Scandinavian Journal of Statistics* 2011; **38**: 169–184.
- Hanley JA and Hajian-Tilaki KO. Sampling variability of non-parametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology* 1997; **4**: 49–58.

- Hastie TJ and Tibshirani RJ. *Generalized additive models*. Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, 1990.
- Inácio de Carvalho V, de Carvalho M, Alonzo TA and González-Manteiga W. Functional covariate-adjusted partial area under the specificity-ROC curve with an application to metabolic syndrome diagnosis, *The Annals of Applied Statistics* 2016 (to appear).
- Inácio de Carvalho V, Jara A, Hanson TE and de Carvalho M. Bayesian nonparametric ROC regression modeling, *Bayesian Analysis* 2013; **3**: 623–646.
- Janes H and Pepe MS. Adjusting for covariates in studies of diagnostic, screening, or prognosis markers: an old concept in a new setting. *American Journal of Epidemiology* 2008; **168**: 89–97.
- Janes H and Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted ROC curve. *Biometrika* 2009; **96**: 371–382.
- Krzanowski WJ and Hand DJ. *ROC curves for continuous data*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 2009.
- López-de-Ullibarri I, Cao R, Cadarso-Suárez C and Lado MJ. Nonparametric estimation of conditional ROC curves: application to discrimination tasks in computerized detection of early breast cancer. *Computational Statistics & Data Analysis* 2008; **52**: 2623–2631.
- McCullagh P and Nelder JA. *Generalized linear models*, Monographs on Statistics and Applied Probability, Second Edition, Chapman & Hall/CRC, 1989.
- Nishikawa RM. Current status and future directions of computer-aided diagnosis in mammography. *Computerized Medical Imaging and Graphics* 2007; **31**: 224–235.
- Pardo-Fernández J, Rodríguez-Álvarez MX and van Keilegom I. A review on ROC curves in the presence of covariates. *REVSTAT – Statistical Journal* 2014; **12**: 21–41.
- Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* 1998; **54**: 124–135.
- Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, New York, 2003.
- Pepe MS and Cai. The analysis of placement values for evaluating discriminatory measures. *Biometrics* 2004; **60**: 528–535.
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>, 2016.

- López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C and Gude F. OptimalCut-points: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *Journal of Statistical Software* 2014; **61**(8): 1–36. URL <http://www.jstatsoft.org/v61/i08/>.
- Rodríguez A and Martínez JC. Bayesian semiparametric estimation of covariate-dependent ROC curves, *Biostatistics* 2014; **15**: 353–369.
- Rodríguez-Álvarez MX, Roca-Pardiñas J and Cadarso-Suárez C. ROC curve and covariates: extending induced methodology to the non-parametric framework. *Statistics and Computing* 2011a, **21**: 483–499.
- Rodríguez-Álvarez MX, Roca-Pardiñas J and Cadarso-Suárez C. A new flexible direct ROC regression model - Application to the detection of cardiovascular risk factors by anthropometric measures. *Computational Statistics & Data Analysis* 2011b, **55**: 3257–3270.
- Rodríguez-Álvarez MX, Tahoces PG, Cadarso-Suárez C and Lado MJ. Comparative study of ROC regression techniques – Applications for the computer-aided diagnostic system in breast cancer detection, *Computational Statistics & Data Analysis* 2011c; **55**: 888–902.
- Tosteson AN and Begg CB. A general regression methodology for ROC curve estimation. *Medical Decision Making* 1988; **8**: 204–215.
- Varela C Tahoces PG, Méndez AJ, Souto M, Vidal JJ. Detection of breast tumours in digitized mammograms. *Computers in Biology and Medicine* 2007; **37**: 214–226.
- Xu T, Wang J and Fang Y. A model-free estimation for the covariate-adjusted Youden index and its associated cut-point. *Statistics in Medicine* 2014; **33**: 4963–4974.
- Yao F, Craiu RV and Reiser B. Nonparametric covariate adjustment for receiver operating characteristic curves. *The Canadian Journal of Statistics* 2010; **38**: 27–46.
- Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**: 32–35.
- Zheng Y and Heagerty PJ. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* 2004; **4**: 615–632.
- Zhou XH, Obuchowski NA and McClish DK. *Statistical methods in diagnostic medicine*. Wiley, New York, 2002.