Computational Statistics and Data Analysis 55 (2011) 3257–3270

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A new flexible direct ROC regression model: Application to the detection of cardiovascular risk factors by anthropometric measures☆

María Xosé Rodríguez-Álvarez [a,b,c,*], Javier Roca-Pardiñas [d], Carmen Cadarso-Suárez [b,c]

[a] Clinical Epidemiology and Biostatistics Unit, Complexo Hospitalario Universitario de Santiago de Compostela (CHUS), Santiago de Compostela, Spain
[b] Unit of Biostatistics, Department of Statistics and Operations Research, University of Santiago de Compostela, Spain
[c] Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Santiago de Compostela, Spain
[d] Department of Statistics and OR, University of Vigo, Spain

ARTICLE INFO

Article history:
Received 15 October 2010
Received in revised form 9 June 2011
Accepted 9 June 2011
Available online 16 June 2011

Keywords:
ROC curve
Generalised additive models
Bootstrap
Cardiovascular risk factors
Anthropometric measures

ABSTRACT

The receiver operating characteristic (ROC) curve is the most widely used measure for evaluating the accuracy of diagnostic tests in terms of differentiating between two conditions. It is known that, in certain circumstances, the characteristics of the patient or the place where the diagnostic test is performed can modify the test's accuracy. A new estimator for the conditional ROC curve, based on direct modelling, is proposed. In this approach, the effect of covariates and false positive fraction on the ROC curve is modelled non-parametrically using generalised additive models (GAM) combined with local polynomial kernel smoothers. The method allows for incorporation of more than one covariate in the regression model for the ROC curve and the possible interaction between them. The proposed model's performance is examined in an in-depth simulation study. Finally, endocrine data are analysed with the aim of assessing the performance of several anthropometric measures in predicting clusters of cardiovascular risk factors in an adult population in Galicia (NW Spain), with adjustment for age and gender.

1. Introduction

The discriminatory capacity of a continuous marker or diagnostic test Y, is usually measured by means of the receiver operating characteristic (ROC) curve (Metz, 1978; Swets and Pickett, 1982). Under the conventional assumption that high marker values are indicative of disease, classification on the basis of Y of an individual as healthy (D) or diseased ($\bar{D}$) can be made by the choice of a cut-off value c, such that, if $Y \geq c$, the individual is classified as diseased, and if $Y < c$, the individual is classified as healthy. Hence, each cut-off value c chosen will give rise to a true positive fraction, $\text{TPF}(c) = P[Y \geq c|D]$, and a false positive fraction, $\text{FPF}(c) = P[Y \geq c|\bar{D}]$. In such a situation, the ROC curve is defined as the set of all TPF–FPF pairs that can be obtained on the cut-off value c varying, $\{(\text{TPF}(c), \text{FPF}(c)),\ c \in (-\infty, \infty)\}$, or, equivalently, as the function of the form $\text{ROC}(t) = S_D(S_{\bar{D}}^{-1}(t))$ for $t \in (0, 1)$, where S_D and $S_{\bar{D}}$ denote the survival functions of Y in diseased and healthy subjects, respectively.

In many practical situations, however, a marker's discriminatory capacity may be affected by a set of continuous and/or categorical covariates $\boldsymbol{X}$. This is the case of our endocrine study, in which the performance of each anthropometric measure in

☆ Software implementing the flexible direct ROC regression approach can be obtained from the corresponding author. An example of the use of the software can be found in the supplementary material.

* Corresponding address: Clinical Epidemiology and Biostatistics Unit, Complexo Hospitalario Universitario de Santiago de Compostela (CHUS), A Choupana s/n., 15706, Santiago de Compostela, Spain. Tel.: +34 981 950014; fax: +34 981 955053.
E-mail address: mariajose.rodriguez.alvarez@usc.es (M.X. Rodríguez-Álvarez).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.06.008

detecting clusters of cardiovascular risk factors changes with individuals' age and gender (see Section 4 for details). The ROC curve may thus be of little value if important covariates are omitted. Moreover, in such situations, interest must be focused on assessing marker Y's discriminatory capacity with reference to the values assumed by $\boldsymbol{X}$. Hence, if the conditional survival functions of Y_D and $Y_{\bar{D}}$, given $\boldsymbol{X}$, are denoted $S_{D\boldsymbol{X}}$ and $S_{\bar{D}\boldsymbol{X}}$, respectively, the conditional or covariate-specific ROC curve is defined as

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = S_{D\boldsymbol{X}}(S^{-1}_{\bar{D}\boldsymbol{X}}(t)), \quad t \in (0, 1). \tag{1}$$

To study the effect of covariates on the accuracy of a diagnostic test, various ROC regression methodologies have been proposed in the statistical literature, namely, 'induced' and 'direct' methodologies (see e.g. Rodríguez-Álvarez et al. (2011) for a detailed review on this topic). 'Induced' methodology (Tosteson and Begg, 1988; Pepe, 1998; Zheng and Heagerty, 2004; Faraggi, 2003) is based on specifying a model for the diagnostic test result as a function of covariates, in both healthy and diseased populations, so that the induced covariate-specific ROC can then be calculated on the basis of the two models. In contrast to this, in 'direct' methodology (Pepe, 2000; Alonzo and Pepe, 2002; Cai and Pepe, 2002; Cai, 2004) the effect of the covariates is directly evaluated on the ROC curve. Using this methodology, the general form of the covariate-specific ROC curve is given by the following generalised linear model (GLM)

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = g(\boldsymbol{X}'\boldsymbol{\beta} + h_0(t)), \quad t \in (0, 1), \tag{2}$$

where $\boldsymbol{X}$ is a p-dimensional vector of covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters, h_0 is an unknown monotone increasing function of the FPFs and g is a known function (the inverse of the link function), describing the functional relationship between the ROC curve and the covariates. Models such as (2) define the so-called class of ROC-GLMs (Pepe, 2003). Insofar as function h_0 is concerned, while it is assumed to have a parametric form by Alonzo and Pepe (2002), in the case of Cai and Pepe (2002) and Cai (2004) it remains completely unspecified.

Along with the advantage of directly evaluating the effect of the covariate on the ROC curve, direct methodology has some other appealing features, including the following: the ROC's property of being invariant to monotonic transformation of the test result is preserved; and, any possible interaction between covariates and FPFs is easy to incorporate into the regression model.

In some circumstances, however, the linearity assumption of the ROC-GLM regression model might affect the validity of the findings. From an applied point of view, the functional form of the relationship between continuous covariates and the ROC curve could provide important information. Misleading conclusions may be drawn if this relationship is incorrectly specified. This restriction can be avoided by, say, using an extension of (2) as a generalised additive model (GAM) (Hastie and Tibshirani, 1990). In this setting, the researcher, rather than assuming a parametric form for the effects of the continuous covariates, solely assumes that these effects could be represented by arbitrary smooth functions. The ROC-GAM can thus be expressed as:

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = g\left(\alpha + \sum_{k=1}^{p} f_k(X_k) + h_0(t)\right), \quad t \in (0, 1), \tag{3}$$

where f_j and h_0 are assumed to be smooth and unknown functions.

To date, some attempts have been made to include covariates in ROC analysis in the non-parametric framework. Recently, López-de-Ullibarri et al. (2008) studied the effect of covariates on the ROC curve, using a non-parametric approach based on local linear estimation of conditional survival functions in healthy and diseased subjects. In the induced ROC regression context, Yao et al. (2010), González-Manteiga et al. (2011) and Rodríguez-Álvarez et al. (2010) have proposed new non-parametric estimators of the covariate-specific ROC curve, based on kernel-type regression estimators. Although all these approaches have proved useful, they suffer from the limitation of only being able to address a single continuous covariate.

To the best of our knowledge, there have been no approaches proposed in the literature, based on direct methodology, in which the effect of continuous covariates on the ROC curve are modelled non-parametrically. The main goal of this paper is to present a new flexible estimator of the covariate-specific ROC curve based on the ROC-GAM regression model given in (3). So far, several approaches for estimating a GAM have been proposed in the statistical literature, including methods based on penalised regression splines (Eilers and Marx, 1996; Wood, 2003; Lang and Brezger, 2004), or, kernel-type smoothers (Wand and Jones, 1995; Fan and Gijbels, 1996). In this work, use is made of the local scoring algorithm with an inner backfitting loop (Hastie and Tibshirani, 1990; Breiman and Friedman, 1985), and the partial functions in (3), f_j and h_0, are estimated based on local linear kernel smoothers (Wand and Jones, 1995). Among the advantages of using such smoothers is the possible use of binning-type acceleration techniques (Fan and Marron, 2004) to reduce computational time and so ensure that the problem can be adequately addressed in practical situations.

It should be noted that the class of ROC-GAM regression models given in (3) includes, by way of specific examples, other models previously addressed in the literature. For instance, if $g = \Phi$ and $h_0(t) = \alpha_0 + \alpha_1\Phi^{-1}(t)$ – where Φ denotes the cumulative distribution function of a standard normal random variable – the ROC curve follows the classic binormal model. If the effects of the covariates X_j is linear, however, then the corresponding partial functions can be expressed parametrically $f_j(X_j) = \beta_j X_j$, yielding the ROC-GLM regression model (2).

The remainder of the paper is structured as follows: Section 2 introduces the proposed algorithm for the estimation of the ROC-GAM regression model (3). We extend the algorithm proposed by Alonzo and Pepe (2002) to the non-parametric framework. In Section 3, the performance of the estimation procedure is evaluated by means of simulations. In Section 4, we illustrate our method using data from the endocrine field and conclude with a discussion in Section 5. Some technical details have been added by way of an Appendix.

2. ROC-GAM estimation procedure

This section presents an algorithm for the estimation of the ROC-GAM regression model (3). It should be noted that, in order to guarantee the identification of model (3), we introduce a constant α into the model and require a zero mean for the partial functions, $E\{f_j(X_j)\} = 0,\ j = 1, \ldots, p$ and $E\{h_0(t)\} = 0$ (see Hastie and Tibshirani (1990)). For study purposes, deem $\boldsymbol{X}$ to be a p-dimensional covariate vector and let $\{(y_i^{\bar{D}}, \boldsymbol{x}_i^{\bar{D}})\}_{i=1}^{n_{\bar{D}}}$ and $\{(y_j^{D}, \boldsymbol{x}_j^{D})\}_{j=1}^{n_D}$ be two independent random samples drawn from the healthy and diseased populations respectively.

Unlike standard regression analysis, in ROC-GAM regression model (3) the dependent variable is not directly observable. This makes it necessary for another interpretation to be given to the ROC curve (see Pepe (2000) and Pepe and Cai (2004)). The key idea for fitting the model (3) is based on the placement values (Hanley and Hajian-Tilaki, 1997) of Y_D defined as $PV_D \equiv S_{\bar{D}\boldsymbol{X}}(Y_D)$. Given that

$$E[I[PV_D \leq t]|\boldsymbol{X}] = \mathrm{ROC}_{\boldsymbol{X}}(t),$$

the covariate-specific ROC curve can be viewed as the conditional expectation of the binary variable $B_{Dt} = I[PV_D \leq t]$. The ROC-GAM regression model (3) can therefore be viewed as a regression model for B_{Dt}. This suggests (Alonzo and Pepe, 2002) that estimation of the ROC-GAM regression model (3) can be based on the following algorithm:

1. choose a set $T = \{t_l, l = 1, \ldots, n_T\}$ of FPFs;
2. estimate $S_{\bar{D}\boldsymbol{X}}$ on the basis of $\{(y_i^{\bar{D}}, \boldsymbol{x}_i^{\bar{D}})\}_{i=1}^{n_{\bar{D}}}$;
3. for each disease observation, calculate the estimated placement value $PV_j = \hat{S}_{\bar{D}\boldsymbol{x}_j^D}(y_j^D),\ j = 1, \ldots, n_D$;
4. for all $t_l \in T$ and each disease observation, calculate the binary placement value indicator $\hat{B}_{jt_l} = I[PV_j \leq t_l],\ l = 1, \ldots, n_T,\ j = 1, \ldots, n_D$; and
5. fit the following ROC-GAM binary regression model

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = g\left(\alpha + \sum_{k=1}^{p} f_k(X_k) + h_0(t)\right), \tag{4}$$

to the data $\{(\hat{B}_{jt_l}, \{\boldsymbol{x}_j^D, t_l\}),\ l = 1, \ldots, n_T,\ j = 1, \ldots, n_D\}$ and obtain the estimates $\widehat{\mathrm{ROC}}_{\boldsymbol{X}}(t)$.

To estimate the binary GAM (4), the local scoring algorithm with backfitting was used (Breiman and Friedman, 1985; Opsomer, 2000). Briefly, the local scoring algorithm is analogous to the use of iterative reweighted least squares (McCullagh and Nelder, 1989) for solving non-linear regression equations. The backfitting algorithm cycles through each of the covariates $\{X_k\ (k = 1, \ldots, p),\ t\}$ and the estimates $\hat{f}_k$ and $\hat{h}_0$ are obtained by applying local linear kernel smoothers (Wand and Jones, 1995; Fan and Gijbels, 1996) to the corresponding partial residuals. We have chosen to use of local linear kernel smoothers, since these estimators enable us to implement binning (Fan and Marron, 2004) directly into the estimation algorithm and computing time is thus drastically reduced (see Appendix for a detailed description). Finally, to obtain an increasing monotone estimation curve of h_0, we used the monotone smoothing technique proposed by Friedman and Tibshirani (1984).

It should be noted that the observations used to fit the binary GAM (4) are no longer independent, inasmuch as: (a) placement values are estimated rather than observed; and, (b) each disease observation is 'compared' with all $t_l \in T$. It is well known that in the presence of correlated errors, standard bandwidths selectors fail to work and can result in an over (or under) fit (see e.g. Opsomer et al. (2001)). In this study, the cross-validation criterion was used for the automatic choice of bandwidths. As pointed out, this choice of bandwidths may be far from optimal. Nevertheless, the estimation procedure seems to perform reasonably well in the simulation studies presented in Section 3 below.

To implement the estimation procedure presented at the beginning of this section, the conditional survival function in healthy subjects must be estimated, $S_{\bar{D}\boldsymbol{X}}$ (see Step 2). In this study, we propose to model the effect of covariates on $Y_{\bar{D}}$ by a non-parametric location-scale regression model, such that

$$Y_{\bar{D}} = \mu_{\bar{D}}(\boldsymbol{X}) + \sigma_{\bar{D}}(\boldsymbol{X})\varepsilon_{\bar{D}}, \tag{5}$$

where $\mu_{\bar{D}}$ and $\sigma^2_{\bar{D}}$ are the regression and the variance functions respectively, and error $\varepsilon_{\bar{D}}$ is assumed to be independent of the covariates $\boldsymbol{X}$, with zero mean, unit variance and survival function $S_{\bar{D}}$. With this configuration, it can be shown that

$$S_{\bar{D}\boldsymbol{X}}(y) = S_{\bar{D}}\left(\frac{y - \mu_{\bar{D}}(\boldsymbol{X})}{\sigma_{\bar{D}}(\boldsymbol{X})}\right).$$

Our proposal is based on using local polynomial kernel smoothers to estimate the regression, $\mu_{\bar{D}}(\cdot)$, and variance, $\sigma^2_{\bar{D}}(\cdot)$, functions of (5) and then estimating the survival function $S_{\bar{D}}$ on the basis of the empirical survival distribution of the standardised residuals.

If $\boldsymbol{X}$ is univariate with $\boldsymbol{X} = X$ being X a continuous covariate, the estimation of the variance and regression function of (5) can be done as in Rodríguez-Álvarez et al. (2010). For the multivariate case, the following regression model is assumed

$$\begin{aligned} Y_{\bar{D}} &= \mu_{\bar{D}}(\boldsymbol{X}) + \sigma_{\bar{D}}(\boldsymbol{X})\varepsilon_{\bar{D}} \\ &= \alpha_0 + \sum_{k=1}^{p} f_k(X_k) + \exp\left(\alpha_1 + \sum_{k=1}^{p} g_k(X_k)\right)\varepsilon_{\bar{D}}, \end{aligned} \tag{6}$$

where α_0 and α_1 are fixed parameters, and f_k and g_k $(k = 1, \ldots, p)$ are smooth and unknown functions. Note that under (6), we have

$$\begin{aligned} E[\log(Y_{\bar{D}} - \mu_{\bar{D}}(\boldsymbol{X}))^2|\boldsymbol{X}] &= \log(\sigma^2_{\bar{D}}(\boldsymbol{X})) + E[\log(\varepsilon^2_{\bar{D}})] \\ &= 2\alpha_1 + \sum_{k=1}^{p} 2g_k(X_k) + E[\log(\varepsilon^2_{\bar{D}})] \\ &= \alpha' + \sum_{k=1}^{p} g'_k(X_k), \end{aligned}$$

with $\alpha' = 2\alpha_1 + E[\log(\varepsilon^2_{\bar{D}})]$, and $g'_k(\cdot) = 2g_k(\cdot)$. Thus,

$$\sigma^2_{\bar{D}}(\boldsymbol{X}) = E[(Y_{\bar{D}} - \mu_{\bar{D}}(\boldsymbol{X}))^2|\boldsymbol{X}] = \theta \exp\left(\alpha' + \sum_{k=1}^{p} g'_k(X_k)\right),$$

where $\theta = 1/\exp(E[\log(\varepsilon^2_{\bar{D}})])$.

This suggests that estimation of the regression and variance functions of (6) can be based on the following steps:

1. estimate

$$\mu_{\bar{D}}(\boldsymbol{X}) = \alpha_0 + \sum_{k=1}^{p} f_k(X_k)$$

based on the sample $\{(y_i^{\bar{D}}, \boldsymbol{x}_i^{\bar{D}})\}_{i=1}^{n_{\bar{D}}}$ using local linear kernel smoothers and the backfitting algorithm for the unweighted case.

2. estimate in a similar fashion

$$E[\log(Y_{\bar{D}} - \mu_{\bar{D}}(\boldsymbol{X}))^2|\boldsymbol{X}] = \alpha' + \sum_{k=1}^{p} g'_k(X_k)$$

based on the sample $\{(\log(y_i^{\bar{D}} - \hat{\mu}_{\bar{D}}(\boldsymbol{x}_i^{\bar{D}}))^2, \boldsymbol{x}_i^{\bar{D}})\}_{i=1}^{n_{\bar{D}}}$.

3. estimate $\sigma^2_{\bar{D}}(\boldsymbol{X})$ as follows

$$\hat{\sigma}^2_{\bar{D}}(\boldsymbol{X}) = \hat{\theta} \exp\left(\hat{\alpha}' + \sum_{k=1}^{p} \hat{g}'_k(X_k)\right),$$

where $\hat{\theta}$ is the ordinary least squares estimate of θ

$$\hat{\theta} = \frac{\sum_{i=1}^{n_{\bar{D}}} \left((y_i^{\bar{D}} - \hat{\mu}_{\bar{D}}(\boldsymbol{x}_i^{\bar{D}}))^2 \exp\left(\hat{\alpha}' + \sum_{k=1}^{p} \hat{g}'_k(x_{ik}^{\bar{D}})\right)\right)}{\sum_{i=1}^{n_{\bar{D}}} \left(\exp\left(\hat{\alpha}' + \sum_{k=1}^{p} \hat{g}'_k(x_{ik}^{\bar{D}})\right)\right)^2}.$$

Finally, the survival function $S_{\bar{D}}$ is estimated on the basis of the empirical survival distribution of the standardised residuals

$$\hat{S}_{\bar{D}}(y) = \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n_{\bar{D}}} I\left[\frac{y_i^{\bar{D}} - \hat{\mu}_{\bar{D}}(\boldsymbol{x}_i^{\bar{D}})}{\hat{\sigma}_{\bar{D}}(\boldsymbol{x}_i^{\bar{D}})} \geq y\right].$$

3. Simulation study

This section reports the results of a simulation study conducted to study the practical behaviour of the estimation procedure described in Section 2 above. Specifically, three different simulation scenarios were considered, namely: (a) a

linear scenario; (b) a non-linear scenario, which presents a covariate effect far from linear; and (c) a scenario with more than one continuous covariate.

In step 5 of our estimation procedure, the assumption that the observations are independent no longer holds, and so this study first examined the behaviour of the cross-validation criterion used to choose the optimal smoothing parameters. To this end, the performance of the proposed method was compared against that of the induced non-parametric regression approach (Rodríguez-Álvarez et al., 2010), in which this particular issue is of no concern. Moreover, in this approach the optimal smoothing parameters are also chosen by cross-validation. In addition, the efficiency and robustness of our model was studied, by comparing it to the direct parametric approach proposed by Alonzo and Pepe (2002) and the semi-parametric model suggested by Cai (2004).

Data were simulated from three scenarios, namely,

- Scenario I

$$Y_D = 2 + 4X_1 + \varepsilon_D$$

and

$$Y_{\bar{D}} = 1.5 + 3X_1 + 0.5\varepsilon_{\bar{D}}.$$

- Scenario II

$$Y_D = 1 + \sin(\pi(X_1 + 2)) + (.25 + .25(X_1 + 1))\varepsilon_D$$

and

$$Y_{\bar{D}} = 0.5\exp(X_1) + (.25 + .25(X_1 + 1))\varepsilon_{\bar{D}}.$$

- Scenario III

$$Y_D = .5\sin(\pi(X_1 + 1)) + .5\exp(X_1) - X_2^2 + .5\varepsilon_D$$

and

$$Y_{\bar{D}} = .5\exp(X_1) - 2X_2^2 + .5\varepsilon_{\bar{D}}.$$

In all scenarios, X_1 and X_2 are uniformly distributed on $[-1, 1]$, $\varepsilon_{\bar{D}}$ and ε_D have the standard normal distribution, and Φ denotes the CDF of a standard normal variable. With the above configurations, the corresponding covariate-specific $\mathrm{ROC}_{\boldsymbol{X}}(t)$ are respectively

- Scenario I

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = \Phi(0.5 + X_1 + 0.5\Phi^{-1}(t)).$$

- Scenario II

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = \Phi\left(\frac{1 + \sin(\pi(X_1 + 2)) - 0.5\exp(X_1)}{(.25 + .25(X_1 + 1))} + \Phi^{-1}(t)\right).$$

- Scenario III

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = \Phi(\sin(\pi(X_1 + 1)) + 2X_2^2 + \Phi^{-1}(t)).$$

The discrepancy between the estimator of the covariate-specific ROC curve and the true ROC curve was measured in terms of the empirical version of the global mean squared error (MSE):

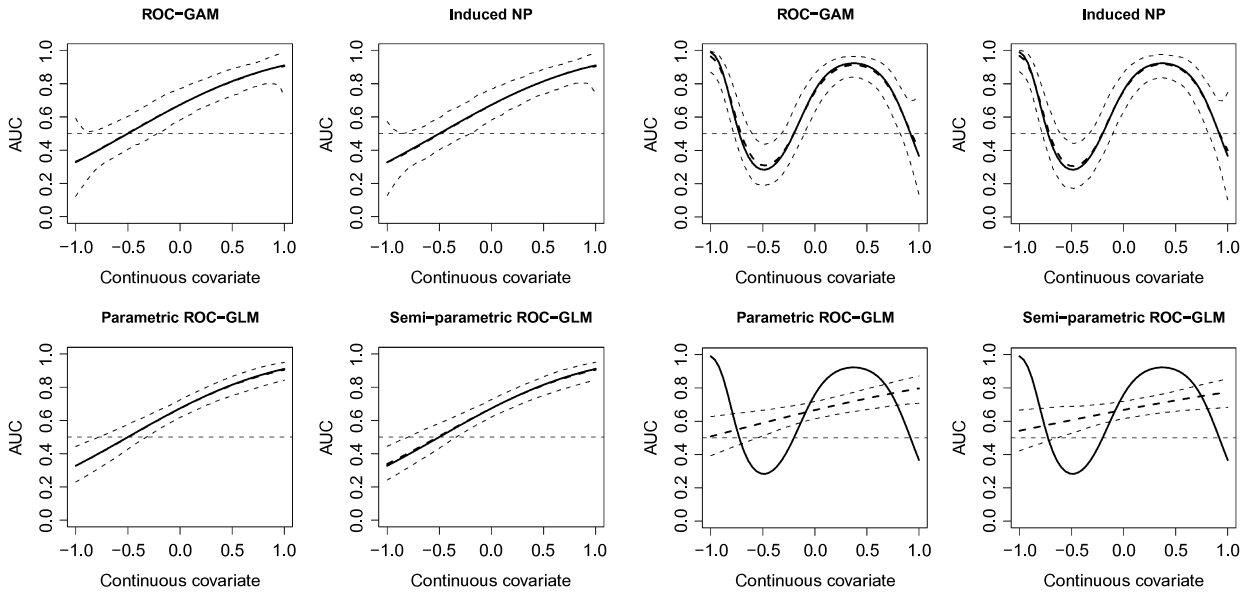
$$\mathrm{MSE} = \frac{1}{n_{X_1}}\sum_{l=1}^{n_{X_1}}\frac{1}{n_{X_2}}\sum_{m=1}^{n_{X_2}}\frac{1}{n_\tau}\sum_{r=1}^{n_\tau}(\widehat{\mathrm{ROC}}_{(X_1,X_2)=(x_l,x_m)}(t_r) - \mathrm{ROC}_{(X_1,X_2)=(x_l,x_m)}(t_r))^2,$$

with $x_l = -1.0 + 2\frac{l-1}{n_{X_1}-1}$, $l = 1, \ldots, n_{X_1}$; $x_m = -1.0 + 2\frac{m-1}{n_{X_2}-1}$, $m = 1, \ldots, n_{X_2}$; $t_r = \frac{r-1}{n_\tau - 1}$, $r = 1, \ldots, n_\tau$ and $n_{X_1} = n_{X_2} = n_\tau = 50$ (for the case of one continuous covariate, the empirical MSE was computed in a similar fashion).

To fit the ROC-GAM model (3), the set T of FPFs must be chosen (step 1 of the estimation algorithm). In the results shown below, $n_T = 50$ and equally spaced values were considered, and the probit function, namely, $g^{-1} = \Phi^{-1}$, was taken as the link function. For Scenario I and II, the covariate-specific ROC curve, $\mathrm{ROC}_{\boldsymbol{X}}(t)$, results in a surface. Accordingly, to facilitate the interpretation of the results of the simulation study, we have summarised them by means of the covariate-specific Area Under the Curve (AUC), $\mathrm{AUC}_{\boldsymbol{X}} = \int_1^0 \mathrm{ROC}_{\boldsymbol{X}}(t)\mathrm{d}t$. Table 1 lists the averages and standard deviations of the MSEs obtained in 1000 data sets simulated from Scenarios I, II and III. In all cases, the same sample size was considered for healthy and diseased subjects, with $n = n_D = n_{\bar{D}} = 50, 100, 200, 500$. As expected, the MSE decreases as the sample sizes increase. As can be seen, the proposed ROC-GAM regression model performs similarly to the induced non-parametric (Induced NP) approach, which suggests that the behaviour of the cross-validation criterion used to choose the optimal

Table 1
Average (standard deviation) of estimated mean squared error (MSE) (x 1000) based on 1000 simulated data sets, yielded by the proposed ROC-GAM, the induced non-parametric approach (Induced NP), Alonzo–Pepe's direct parametric approach (Parametric ROC-GLM) and Cai's direct semi-parametric approach (Semi-parametric ROC-GLM).

Scenario	Model	Sample size			
		50	100	200	500
I	ROC-GAM	17.454 (11.224)	9.393 (5.105)	5.777 (2.306)	3.508 (0.921)
	Induced NP	17.647 (10.054)	9.311 (4.700)	5.163 (2.177)	2.579 (0.894)
	Parametric ROC-GLM	6.125 (5.828)	2.979 (2.945)	1.511 (1.337)	0.595 (0.559)
	Semi-parametric ROC-GLM	9.586 (6.079)	5.596 (3.250)	3.358 (1.486)	1.864 (0.727)
II	ROC-GAM	32.532 (15.031)	15.887 (6.522)	7.706 (3.160)	3.293 (1.212)
	Induced NP	38.923 (17.904)	18.609 (7.832)	8.811 (3.740)	3.454 (1.385)
	Parametric ROC-GLM	64.705 (7.108)	64.089 (4.876)	62.164 (2.759)	61.332(1.628)
	Semi-parametric ROC-GLM	66.414 (7.000)	63.631 (3.971)	61.030 (1.873)	59.618(0.832)
III	ROC-GAM	55.183 (22.863)	24.539 (9.274)	13.185 (4.183)	6.617 (1.951)
	Induced NP	–	–	–	–
	Parametric ROC-GLM	54.330 (12.030)	47.437 (6.982)	44.726 (4.335)	42.975 (2.578)
	Semi-parametric ROC-GLM	54.481 (11.682)	47.690 (6.955)	44.949 (4.402)	43.199 (2.712)



(a) Scenario I. (b) Scenario II.

Fig. 1. True AUC (solid line) versus the average of simulated AUCs (dashed line), along with 2.5 and 97.5 simulation quantiles, for Scenarios I and II and $n_D = n_{\bar{D}} = 200$, based on 1000 estimates. From left to right: proposed ROC-GAM, induced non-parametric (Induced NP) approach, Alonzo–Pepe's direct parametric approach (Parametric ROC-GLM), Cai's direct semi-parametric approach (Semi-parametric ROC-GLM).

smoothing parameters is satisfactory. It should be noted that, for Scenario III, the performance of the ROC-GAM approach was not compared with the Induced NP, inasmuch as this approach is only able to address one continuous covariate. As would be expected under Scenario I, the models that displayed the best performance were the parametric and semi-parametric approaches, though our method also performed satisfactorily (see Fig. 1(a)). For Scenarios II and III, however, the effect of the covariates X_1 and X_2 on the ROC curve were far from linear, and the estimates obtained by the parametric and semi-parametric models were thus not suitable, as can be seen from Table 1 (see also Fig. 1(b)). For Scenario II, the good performance of the proposed model is clearly observable in Fig. 2, where the contour plot of the true ROC surface (solid lines) jointly with the average of estimated ROC surfaces (dashed lines) are shown. The contour plot of the corresponding estimated standard deviations is shown in Fig. 3. Fig. 4 depicts the average of the AUCs estimated by the proposed model, along with the 2.5 and 97.5 simulation quantiles, for the different sample sizes. Finally, for Scenario III, we also investigated the performance of the ROC-GAM when estimating the two centred partial functions $f_1(X_1) = \sin(\pi(X_1 + 1))$ and $f_2(X_2) = 2X_2^2 - \frac{2}{3}$, and the baseline function $h_0(t) = \Phi^{-1}(t)$. Averages of the results are graphically depicted in Fig. 5. The same sample size was considered for both healthy and diseased subjects, with $n = n_D = n_{\bar{D}} = 200$. The good performance of the resulting estimates is evident, with the functional forms of the corresponding true curves being recovered very successfully.

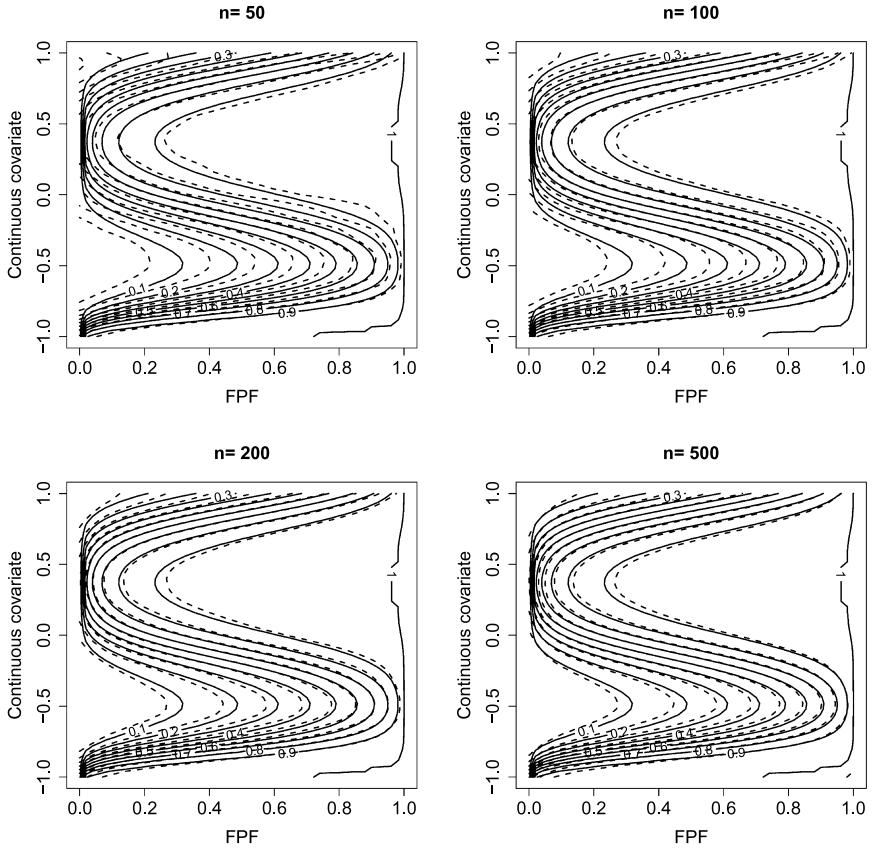


Fig. 2. Contour plot for the true ROC (solid line) versus the average of simulated ROCs (dashed line) yielded by the ROC-GAM, for Scenario II and for different samples sizes ($n = n_D = n_{\bar{D}}$), based on 1000 estimates.

4. Application to real data

Coronary heart and cerebrovascular diseases are one of the leading causes of adult mortality worldwide. There are several cardiovascular disease (CVD) risk factors for developing these vascular diseases, including diabetes mellitus, dyslipidemia, arterial hypertension and obesity. These CVD risks have a metabolic basis and the common characteristic of tending to occur in clusters (e.g., metabolic syndrome), in that when one risk factor is present in any subject, another is also likely to appear in the same individual. Although the metabolic defect ultimately responsible for such clustering is not known, obesity is nonetheless well documented as being a common denominator (Karelis et al., 2004).

Yet, serious concern exists as to which anthropometric measure related to excess body fat best predicts cardiovascular risks. Although there is some evidence to indicate that indices of abdominal fat accumulation, such as waist circumference (WC) or waist-to-hip ratio (WHR), may be better predictors than body mass index (BMI) (Franzosi, 2006; Yusuf et al., 2005), the question of which obesity index is best for predicting CVD risks is still a matter of controversy in biomedical research (Litwin, 2008; Gelber et al., 2008).

In view of the existing gaps in knowledge, we applied the proposed ROC-GAM methodology to an endocrine study (Botana et al., 2007; Tomé et al., 2008), with the aim of assessing the performance of BMI, WC and WHR for predicting clusters of cardiovascular risk factors in an age- and gender-adjusted adult population in Galicia (NW Spain). Since it is well established that anthropometric measures perform differently according to gender, the age-by-gender interaction was included in the ROC-GAM regression models.

4.1. Data source

This study consisted of field work covering a random sample, representative of the Galician adult population (2850 subjects, age range 18–85 years). Direct anthropometric measurements were taken, including weight (in kg), height (in m), WC (in cm) and hip circumference (in cm), and the WHR was subsequently calculated. BMI was computed as weight divided by height squared. A diseased subject was defined as any person presenting with two or more CVD risk factors (raised triglycerides, reduced HDL cholesterol, raised blood pressure and raised fasting plasma glucose) as per the International

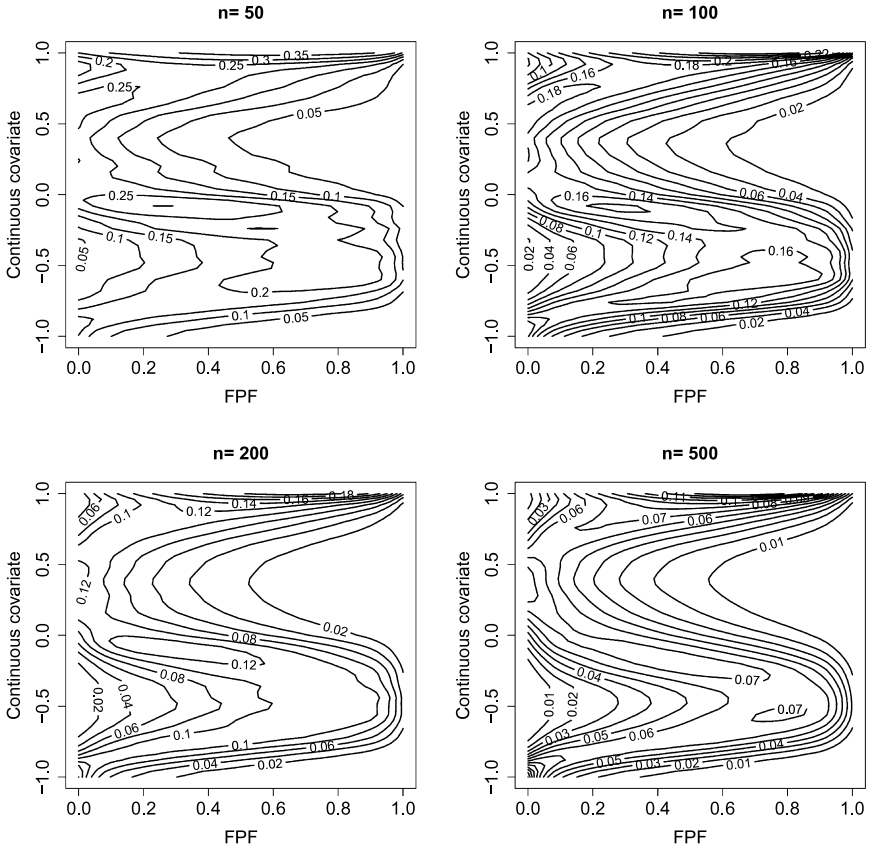


Fig. 3. Contour plot for standard deviation of simulated ROCs yielded by the ROC-GAM, for Scenario II and for different samples sizes ($n = n_D = n_{\bar{D}}$), based on 1000 estimates.

Diabetes Federation criteria (International Diabetes Federation, 2005). Of the total of 2850 subjects, 46.2% were men (899 healthy and 418 diseased) and 53.8% women (1250 healthy and 273 diseased). A detailed description of this dataset can be found in Tomé et al. (2008).

4.2. Statistical analysis

For each anthropometric measure, separate covariate-specific ROC curves were calculated, assuming the following multivariate interaction model:

$$\mathrm{ROC}_{(Age,Gender)}(t) = \Phi\left(\alpha + \sum_{i=0}^{1}\alpha_i I[Gender = i] + f(Age) + \sum_{i=0}^{1} f_i(Age) I[Gender = i] + h_0(t)\right),$$

where α and α_i ($i = 0, 1$) are fixed parameters, f and f_i ($i = 0, 1$) are smooth functions of *Age*, h_0 is a smooth function of the false positive fraction, and *Gender* is a binary variable taking a value of 0 in the case of Men and 1 in the case of Women.

In addition to the estimated covariate-specific ROC curves, another summary measure of accuracy, the AUC, was obtained. Bootstrap-based methods were used for constructing confidence intervals for this measure (Efron and Tibshirani, 1993). For our purposes, the global sample will be denoted as $\{(y_k, \boldsymbol{x}_k, d_k)\}_{k=1}^{n_{\bar{D}}+n_D}$, where d is a binary indicator, taking the value 1 for diseased and 0 for healthy individuals.

Given a point $\boldsymbol{x}$ in the range of $\boldsymbol{X}$ the steps for the construction of the confidence interval were as follows:

1. for $b = 1, \ldots, 500$, draw a random sample of size $n_{\bar{D}} + n_D$, with replacement from the global population; and,
2. from $\{(y^*_{k,b}, \boldsymbol{x}^*_{k,b}, d^*_{k,b})\}_{k=1}^{n_{\bar{D}}+n_D}$ obtain $\widehat{\mathrm{ROC}}^*_{\boldsymbol{X}=\boldsymbol{x},b}(t)$ and compute

$$\widehat{\mathrm{AUC}}^*_{\boldsymbol{X}=\boldsymbol{x},b} = \int_0^1 \widehat{\mathrm{ROC}}^*_{\boldsymbol{X}=\boldsymbol{x},b}(t)\mathrm{d}t.$$

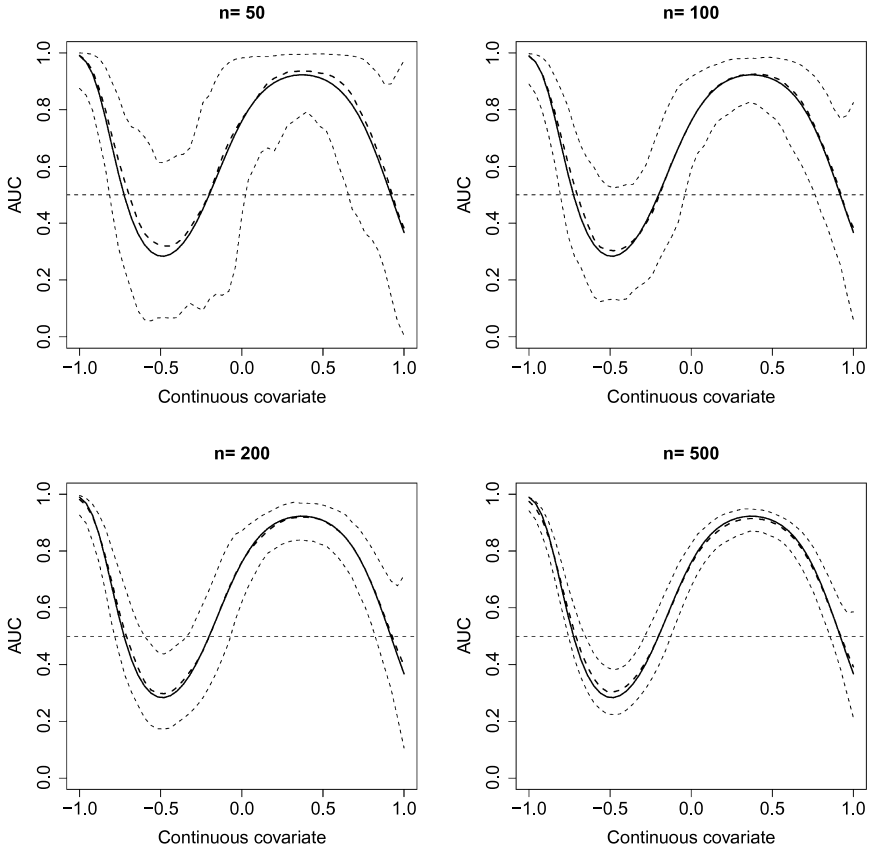


Fig. 4. True AUC (solid line) versus the average of simulated AUCs (dashed line) yielded by the ROC-GAM, along with 2.5 and 97.5 simulation quantiles for the ROC-GAM, in Scenario II and for different samples sizes ($n = n_D = n_{\bar{D}}$), based on 1000 estimates.

Once the above process has been completed, the 100 per cent $x(1-\alpha)$ limits for the confidence interval for the true $\text{AUC}_{X=x}$ are given by

$$\left(\widehat{\text{AUC}}^{\alpha/2}_{\boldsymbol{X}=\boldsymbol{x}}, \widehat{\text{AUC}}^{1-\alpha/2}_{\boldsymbol{X}=\boldsymbol{x}}\right)$$

where $\widehat{\text{AUC}}^{p}_{\boldsymbol{X}=\boldsymbol{x}}$ represents the p-percentile of the estimated $\widehat{\text{AUC}}^{*}_{\boldsymbol{X}=\boldsymbol{x},b}$ $(b = 1, \ldots, 500)$.

4.3. Results

The estimated partial functions together with the corresponding 95% pointwise bootstrap confidence bands (constructed using the methodology described above for the AUC) are plotted in Fig. 6. Fig. 7 depicts the estimated covariate-specific AUCs together with the corresponding 95% pointwise bootstrap confidence bands. For comparison purposes, the parametric approach of Alonzo and Pepe (2002) was also applied to this data. The corresponding AUCs with the 95% pointwise bootstrap confidence bands can be found in the supplementary material.

Particularly in the case of BMI, age displayed a more marked effect among women than among men. These results suggest the need for the incorporation of the age–gender interaction in the ROC-GAM regression model fitted. In the case of men, the accuracy tends to decline progressively and eventually starting to lose significance around age 65. With respect to women, the AUCs indicate very good discriminatory capacity for the youngest women, with values greater than 0.8. From 45 years onwards, the AUC reaches a plateau until subjects are in their sixties. From the age of sixty, the accuracy of BMI declines progressively until age of 65–70 years. From the age of seventy, the AUC increases again, although the bootstrap confidence intervals in this age's period do not allow to extract any biomedical conclusion.

With respect to WHR, both men and women present a similar behaviour. As regards the effect of age, this covariate presents very slight decrease effect between 25 and 75 years. For subjects older than 75 years, the AUC strongly decreases, reaching even values below 0.5. However, the bootstrap confidence intervals in this part include the 0.5, so there is no evidence suggesting that WC can be used to classify those individuals with a higher risk of cardiovascular disease.

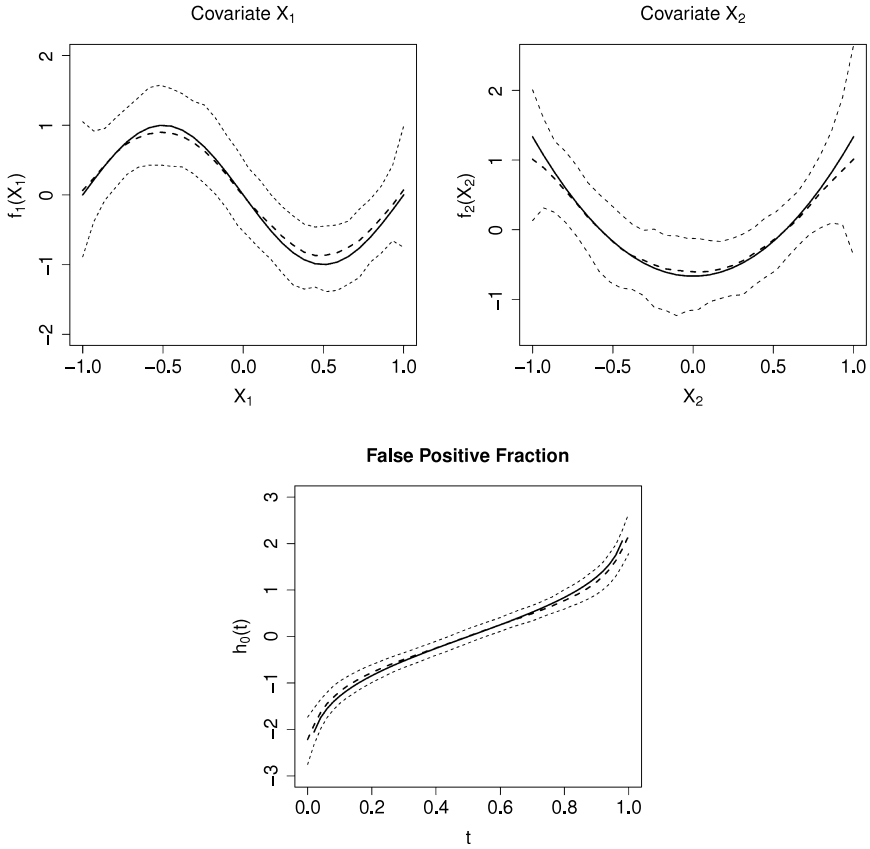


Fig. 5. Simulation results based on 1000 replicated samples obtained under Scenario III for $n_D = n_{\bar{D}} = 200$. From left to right: true curve f_1 (solid line) and average estimate $\hat{f}_1$ (dashed line); true curve f_2 (solid line) and average estimate $\hat{f}_2$(dashed line); and true curve h_0 (solid line) and average estimate $\hat{h}$ (dashed line). In all cases, 2.5 and 97.5 simulation quantiles have also been plotted.

Finally, in the case of WC, accuracy decreases as age increases until subjects are in their seventies. From the age of seventy, the accuracy tend to increase very slightly. Since the confidence intervals here are wide, one should be cautious in extracting any conclusion. As can be observed, WC presents a lower discriminatory capacity in men than in women.

As can be seen from these figures, the discriminatory capacity of the measures studied was affected by sex and age alike. Even so, our study does not suggest that any one measure studied has a discriminatory capacity that is clearly better than that of the others over the span of a subject's lifetime.

5. Discussion

This paper presents a new non-parametric estimator for covariate-specific ROC curve based on direct modelling. The proposed estimator enables a set of continuous and/or categorical covariates and any possible interactions, to be incorporated into the regression model for the ROC curve.

The proposed estimation procedure is based on a combination of local scoring and backfitting algorithms, and uses local linear kernel smoothers. Use is made of the cross-validation criterion to choose optimal smoothing parameters, and binning techniques to speed up computation time. The estimation procedure of the ROC-GAM regression model requires the conditional survival function in the healthy population to be estimated. This paper proposes a location-scale regression model for the test result in the healthy population. Local linear estimators are used to estimate regression and variance functions, and the survival function is empirically estimated on the basis of standardised residuals.

The steps of the proposed estimation algorithm imply the non-independence of the observation upon which the estimation of the ROC-GAM regression model is based. Despite this, the results of the simulation study that compares the new and induced non-parametric approaches would appear to indicate the good performance of our model and, specifically, that of the cross-validation criterion used for automatic selection of the smoothing parameters. We also compared the performance of the ROC-GAM against previous direct approaches, as well as its behaviour in the presence of more than one continuous covariate. The results show that our estimator also performs well in all these situations.

The methodology proposed in this paper was used to analyse an endocrine dataset, in order to evaluate the effect of age and gender on the discriminatory capacity of different anthropometric measures when it came to detecting the presence of

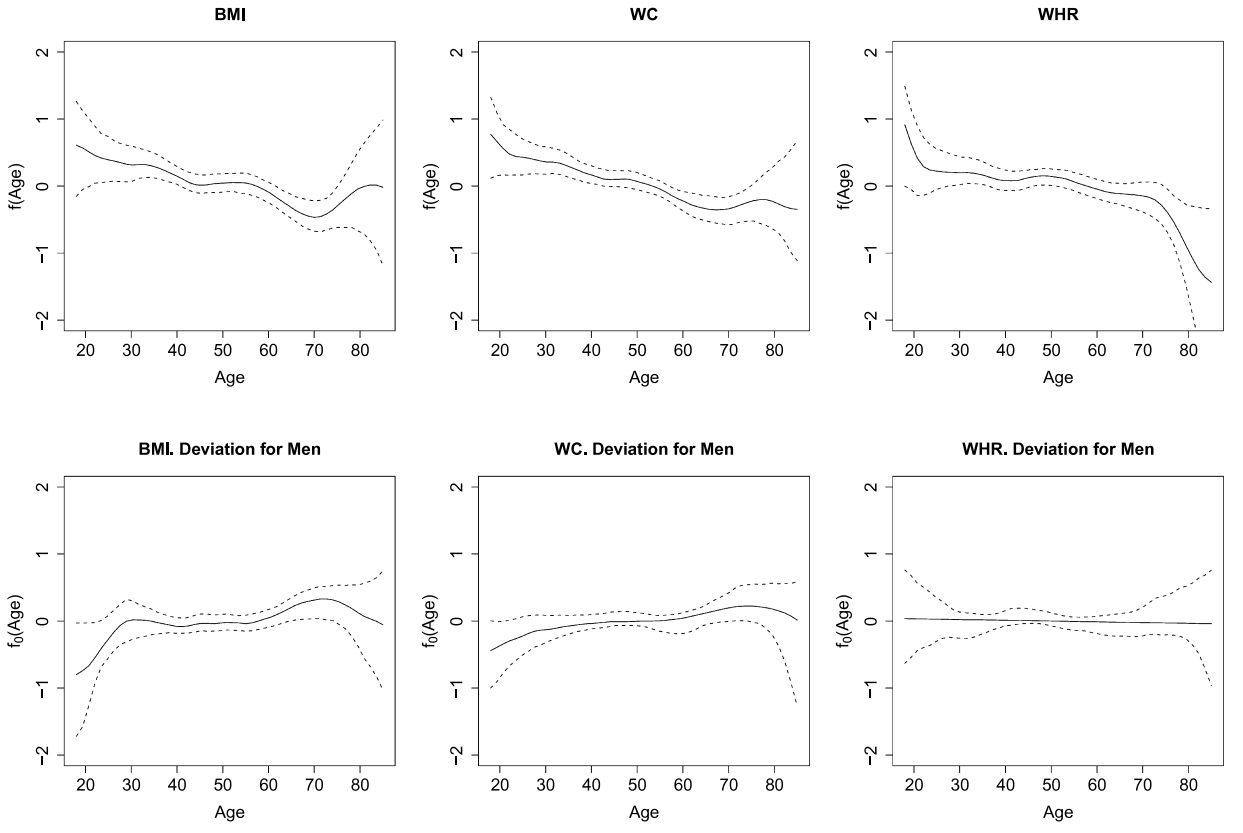


Fig. 6. Estimated partial effect of age, and deviation for men, together with the corresponding 95% pointwise bootstrap confidence bands. From left to right: BMI, WC and WHR.

two or more cardiovascular risk factors. The results show the different behaviour of the measures considered according to individuals' sex and age. It should be noted, however, that, the results obtained in this study are representative of Galicia alone and so cannot be extrapolated to other populations.

Although our methodology was motivated by endocrine data, it could nonetheless be applied in any other biomedical field where the accuracy of a given diagnostic test might depend on a set of continuous and factor-type covariates.

This paper has focused on the presentation of the estimation procedure of the ROC-GAM regression model, yet considerably more work is still needed. For instance, an interesting field for future research would be the development of test statistics to evaluate the possible effects of covariates on the ROC curve, the linearity of the effects of the continuous covariates, or the existence of interactions among covariates, aspects on which – even in the semi-parametric framework – there is still little in the literature. Special attention should be paid to Cai and Zheng (2007)'s paper, in which different procedures to check the goodness of fit for some parametric and semi-parametric ROC regression approaches are proposed.

In our endocrine study, different ROC-GAM regression models were separately fitted for each anthropometric measure. It would, however, be desirable for all the markers to be incorporated into a single regression model. Having more than one observation per subject would mean that, in such a situation, there would be a new source of non-artificial correlation in the data. Consequently, an interesting issue for future research would be to incorporate a structure of correlation among observations pertaining to the same individual into the estimation procedure of the ROC-GAM regression model.

It should be pointed out that, in practical situations, the new methodology may represent a flexible exploratory tool for identifying non-linear covariate effects on the ROC curve. ROC-GAMs can also be used in a diagnostic mode, as an aid for choosing parametric transformation for the covariates (where necessary), in which case a ROC-GLM could thus be used.

Acknowledgments

The authors would like to thank the Galician Endocrinology & Nutrition Foundation (*Fundacion de Endocrinoloxia e Nutricion Galega — FENGA*) for supplying the database used in this study, and would also like to express their gratitude for the support received in the form of the Spanish Ministry of Science and Technology grants MTM2008-01603 and DE2009-0030, Galician Regional Authority (Xunta de Galicia) projects INCITE08PXIB208113PR and PGIDIT07PXIB300191PR, and Instituto de Salud Carlos III (Spanish Ministry of Science and Technology) grant CA09/00539.

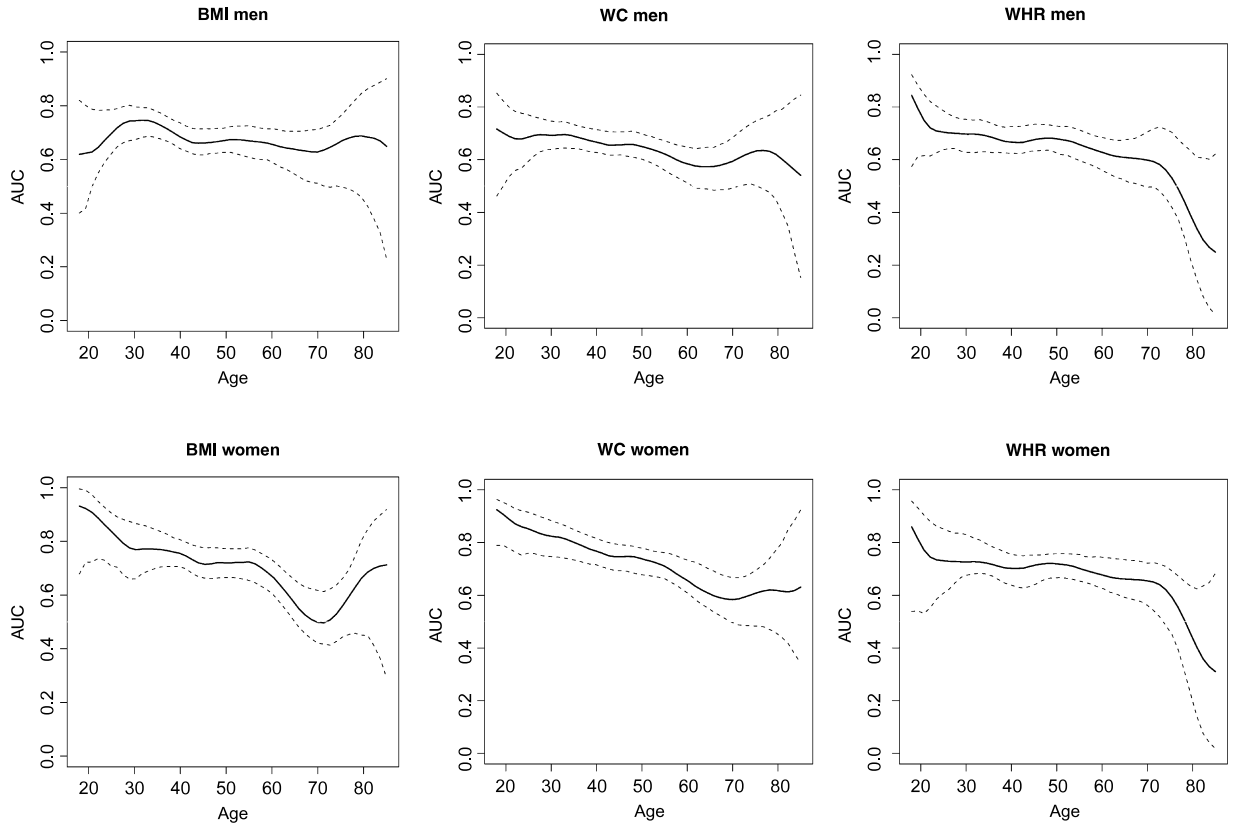


Fig. 7. Estimated AUCs adjusted by age and gender with 95% pointwise bootstrap confidence bands for BMI, WC and WHR.

Appendix. Computational aspects

The local scoring algorithm

Given a sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ of $(Y, \boldsymbol{X})$, the steps of the local scoring algorithm for estimating a GAM for binary response

$$\mu = E[Y|\mathbf{X}] = g(\alpha + f_1(X_1) + \cdots + f_p(X_p)),$$

are as follows:

Initialise. Compute the initial estimates, $\hat{\alpha} = g^{-1}(\bar{y})$, $\hat{f}_j^0 = 0$ $(1 \le j \le p)$ and $\hat{\mu}_i^0 = \bar{y} = n^{-1}\sum_{i=1}^n y_i$ $(i = 1, \ldots, n)$.

Step 1. Form the adjusted dependent variables $\tilde{y}_i = \hat{\eta}_i^0 + (y_i - \hat{\mu}_i^0)/g'(\eta_i^0)$ and weights $\tilde{w}_i = g'(\eta_i^0)^2/\hat{\mu}_i^0(1-\hat{\mu}_i^0)$, where $\hat{\eta}_i^0 = \hat{\alpha} + \hat{f}_1^0(x_{i1}) + \cdots + \hat{f}_p^0(x_{ip})$.

Step 2. Fit an additive model to $\tilde{y}$ using the backfitting algorithm (explained in the following subsection) with weights $\tilde{w}$ and compute the updates $\hat{\alpha}, \hat{f}_j(x_{ij})$, and $\hat{\eta}_i = \hat{\alpha} + \hat{f}_1(x_{i1}) + \cdots + \hat{f}_p(x_{ip})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Step 3. Repeat **Steps** 1 and 2, with $\hat{\mu}_i^0$ being replaced by $\hat{\mu}_i = g(\hat{\eta}_i)$ for $i = 1, \ldots, n$, until

$$\frac{|D(\hat{\boldsymbol{\mu}}^0, \mathbf{Y}) - D(\hat{\boldsymbol{\mu}}, \mathbf{Y})|}{D(\hat{\boldsymbol{\mu}}^0, \mathbf{Y})} \le \varepsilon,$$

where ε is a small threshold and $D(\hat{\boldsymbol{\mu}}, \mathbf{Y}) = -2\sum_{i=1}^n [y_i \log(\hat{\mu}_i) + (1-y_i)\log(1-\hat{\mu}_i)]$.

Fitting weighted additive models

This section describes an algorithm for fitting the additive covariate model

$$E[Y|\mathbf{X}] = \alpha + f_1(X_1) + \cdots + f_p(X_p).$$

The algorithm discussed below is a modification to the weighted case of the backfitting algorithm (Opsomer, 2000). The backfitting algorithm cycles through the covariates X_j $(j = 1, \ldots, p)$ and estimates each f_j by applying local linear kernel smoothers to the partial residuals. These residuals are obtained by removing the estimated effects of the other covariates.

Given a sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ of $(Y, \boldsymbol{X})$ weighted by $\{w_i\}_{i=1}^n$, the steps of the estimation algorithm are as follows:

Initialise. Compute the initial estimates $\hat{\alpha} = \sum_{i=1}^{n} y_i/n$ and $\hat{f}_j^0(x_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Step 1. For $j = 1, \ldots, p$ calculate residuals by removing the estimated effects of all the other covariates:

$$y_i^j = y_i - \hat{\alpha} - \sum_{s=1}^{j-1} \hat{f}_s(x_{is}) - \sum_{s=j+1}^{p} \hat{f}_s^0(x_{is}),$$

and compute the weighted local linear kernel estimators for $i = 1, \ldots, n$ (Wand and Jones, 1995)

$$\hat{f}_j(x_{ij}) = (1\,0) \begin{pmatrix} s_j^0(x_{ij}) & s_j^1(x_{ij}) \\ s_j^1(x_{ij}) & s_j^2(x_{ij}) \end{pmatrix}^{-1} \begin{pmatrix} u_j^0(x_{ij}) \\ u_j^1(x_{ij}) \end{pmatrix} \tag{7}$$

where $s_j^r(x) = \sum_{i=1}^{n}(w_i \cdot L_j^r(x, x_{ij}))$ and $u_j^r(x) = \sum_{i=1}^{n}(w_i \cdot L_j^r(x, x_{ij}) \cdot y_i^j)$, with

$$L_j^r(x, y) = (2\pi)^{-1/2}(x - y)^r \exp(-0.5(h_j^{-1}(x - y))^2),$$

and with h_j being the bandwidth associated with the estimation of $\hat{f}_j$. Finally, in order to meet the identifiability conditions, the resulting estimate $\hat{f}_j(\cdot)$ is replaced by its centred version

$$\hat{f}_j(\cdot) - \frac{1}{n}\sum_{i=1}^{n} \hat{f}_j(x_{ij}).$$

Step 2. Repeat **Step** 1 with $\hat{f}_j^0$ replaced by $\hat{f}_j$, until the convergence criterion

$$\sum_{i=1}^{n} [\hat{f}_j(x_{ij}) - \hat{f}_j^0(x_{ij})]^2 \Big/ \sum_{i=1}^{n} \hat{f}_j^0(x_{ij})^2$$

is below some small threshold ε for all the $j = 1, \ldots, p$.

In the unweighted case, the above backfitting is reduced to that introduced by Opsomer (2000). In such a case, Opsomer and Ruppert (2007) (see also Opsomer and Kauermann (2000)) established the theoretical convergence and consistency of the method. The backfitting algorithm's performance in practice was revised by Nielsen and Sperlich (2005).

In many situations the effect of a continuous covariate on response varies across groups defined by levels of a categorical variable. Although for the sake of simplicity we have not included the so-called factor-by-curve interaction in the above algorithm, this can be adequately modified to incorporate such interaction (see e.g. Roca-Pardiñas et al. (2006)).

Bandwidth selection

It is well known that the estimates obtained for the model depend heavily on the bandwidths $(h_1, \ldots, h_p)$ used in the local linear kernel estimation of the partial functions $(f_1, \ldots, f_p)$. The bandwidths are a trade-off between the bias and variance of the resulting estimates. Various proposals for an optimal selection have been suggested for the additive models, yet the difficulty of asymptotic theory in a backfitting context means that nowadays optimal selection is still a challenging, open problem. Cross-validation was used for the automatic choice of bandwidths.

In each of the cycles of the algorithm, the bandwidth (h_j) used to obtain the estimates $\hat{f}_j$ in Eq. (7) was automatically selected by minimising the following weighted cross-validation error criterion:

$$CV_j = \sum_{i=1}^{n} w_i (y_i^j - \hat{f}_j^{(-i)}(x_{ij}))^2$$

where $\hat{f}_j^{(-i)}$ is the estimate obtained without the ith element of the sample.

Cross-validation implies a high computational cost, inasmuch as it is necessary to repeat the estimation operations several times in order to select the optimal bandwidths. To speed up this process, we used binning-type acceleration techniques (Wand and Jones, 1995; Fan and Gijbels, 1996) to obtain the binning approximations of $\hat{f}_j$ in each of the iterations of the estimation algorithm.

The binning approximations were obtained from the binning sample $\{x_r^{\bullet j}, y_r^{\bullet j}\}$ and the weights $\{w_r^{\bullet}\}$ $(1 \le r \le N)$, with

$$x_1^{\bullet j} < \cdots < x_N^{\bullet j}$$

being a grid of equidistant points along the jth direction. Let us consider δ the distance between consecutive grids. The binning responses $y_r^{\bullet j}$ and the binning weights $W_r^{\bullet}$ are constructed according to $w_r^{\bullet} = \sum_{i=1}^{n} w_r^{\bullet i}$ and $y_r^{\bullet j} = \sum_{i=1}^{n} w_r^{\bullet i} y_i^j$ with

$$w_r^{\bullet i} = w_i(1 - |x_{ij} - x_r^{\bullet j}|/\delta)_+.$$

The binning approximation of the estimator $\hat{f}_j(x)$ is obtained by applying the approximations

$$s_j^r(x) \approx \sum_{l=1}^{N} L_j^r(x, x_l^{\bullet j}; h_j) w_l^{\bullet} \quad \text{and} \quad t_j^r(x) \approx \sum_{l=1}^{N} L_j^r(x, x_l^{\bullet j}; h_j) w_l^{\bullet} y_l^{\bullet j}.$$

As in the estimation algorithm, in the case of the binning technique the cross-validation error CV_j can be approximated by:

$$CV_j \approx \sum_{r=1}^{N} w_r^{\bullet j} \left(\hat{f}_j^{-(r)}(x_r^{\bullet j}) - \frac{y_r^{\bullet j}}{w_r^{\bullet j}} \right)^2,$$

where $\hat{f}_j^{-(r)}$ is obtained without the (r) element of the binning sample.

The finer the grid of points selected, the better the binning approximations. The choice of the number of grid points is a compromise between approximation error and computational speed. In this paper, we used 50 grid points covering the range of each X_j. In practice, depending on the sample size n and the distribution of the covariates, a larger amount of grid points might be more appropriate.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.06.008.

References

Alonzo, T.A., Pepe, M.S., 2002. Distribution-free ROC analysis using binary regression techniques. Biostatistics 3, 421–432.

Botana, M.A., Mato, J.A., Cadarso-Suárez, C., Tomé, M.A., Perez-Fernandez, R., Fernández-Mario, A., Rego-Iraeta, A., Solache, I., 2007. Overweight, obesity and central obesity prevalences in the region of Galicia in Northwest Spain. Obesity and Metabolism 3, 106–115.

Breiman, L., Friedman, J.H., 1985. Estimating optimal transformations for multiple regression and correlations. Journal of the American Statistical Association 80, 580–619 (with discussion).

Cai, T., 2004. Semi-parametric ROC regression analysis with placement values. Biostatistics 5, 45–60.

Cai, T., Pepe, M.S., 2002. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. Journal of the American Statistical Association 97, 1099–1107.

Cai, T., Zheng, Y., 2007. Model checking for ROC regression analysis. Biometrics 63, 152–163.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Eilers, P., Marx, B., 1996. Flexible smoothing with B-splines and penalties. Statisical Science 11, 89–121.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, CRC.

Fan, J., Marron, J.S., 2004. Fast implementation of non-parametric curve estimators. Journal of Computational and Graphical Statistics 3, 35–56.

Faraggi, D., 2003. Adjusting receiver operating characteristic curves and related indices for covariates. The Statistician 52, 179–192.

Franzosi, M.G., 2006. Should we continue to use BMI as a cardiovascular risk factor. Lancet 268, 624–625.

Friedman, J., Tibshirani, R., 1984. The monotone smoothing of scatterplots. Technometrics 26, 243–250.

Gelber, R.P., Gaziano, J.M., Orav, E.J., Manson, J.E., Buring, J.E., Kurth, T., 2008. Measures of obesity and cardiovascular risk among men and women. Journal of the American College of Cardiology 52, 605–615.

González-Manteiga, W., Pardo Fernández, J.C., Van Keilegom, I., 2011. ROC curves in nonparametric location-scale regression models. Scandinavian Journal of Statistics 38, 169–184.

Hanley, J.A., Hajian-Tilaki, K.O., 1997. Sampling variability of non-parametric estimates of the area under receiver operating characteristic curves: an update. Academic Radiology 4, 49–58.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.

International Diabetes Federation. The IDF consensus worldwide definition of the metabolic syndrome http://www.idf.org/webdata/docs/IDF_Meta_def_final.pdf, 2005 (accessed 14.10.10).

Karelis, A.D., St-Pierre, D.H., Conus, F., Rabasa-Lhoret, R., Poehlman, E.T., 2004. Metabolic and body composition factors in subgroups of obesity: what do we know? Journal of Clinical Endocrinology and Metabolism 89, 2569–2575.

Lang, S., Brezger, A., 2004. Bayesian P-splines. Journal of Computational and Graphical Statistics 13, 183–212.

Litwin, S.E., 2008. Which measures of obesity best predict cardiovascular risk? Journal of the American College of Cardiology 52, 616–619.

López-de-Ullibarri, I., Cao, R., Cadarso-Suárez, C., Lado, M.J., 2008. Nonparametric estimation of conditional ROC curves: application to discrimination tasks in computerized detection of early breast cancer. Computational Statistics & Data Analysis 52, 2623–2631.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second edition. Chapman & Hall, London.

Metz, C.E., 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine 8, 183–298.

Nielsen, J.P., Sperlich, S., 2005. Smooth backfitting in practice. Journal of the Royal Statistical Society Series B 10, 43–61.

Opsomer, J.D., 2000. Asymptotic properties of backfitting estimators. Journal of Multivariate Analysis 73, 166–179.

Opsomer, J.D., Kauermann, G., 2000. Weighted local polynomial regression weighted additive models and local scoring. Preprint Series #00-7, Department of Statistics, Iowa State University.

Opsomer, J.D., Ruppert, D., 2007. Fitting a bivariate additive model by local polynomial regression. Annals of Statistics 25, 186–211.

Opsomer, J.D., Wang, Y., Yang, Y., 2001. Nonparametric regression with correlated errors. Statistical Science 16, 134–153.

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

Pepe, M.S., 2000. An interpretation for the ROC curve and inference using GLM procedures. Biometrics 56, 52–359.

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

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

Roca-Pardiñas, J., Cadarso-Suárez, C., Nácher, V., Acuña, C., 2006. Bootstrap-based methods for testing factor-by-curve interactions in generalized additive models: assessing prefrontal cortex neural activity related to decision-making. Statistics in Medicine 25, 2483–2501.

Rodríguez-Álvarez, M.X., Roca-Pardiñas, J., Cadarso-Suárez, C., 2010. ROC curve and covariates: extending induced methodology to the non-parametric framework. Statistics and Computing doi:10.1007/s11222-010-9184-1.

Rodríguez-Álvarez, M.X., Tahoces, P.G., Cadarso-Suárez, C., Lado, M.J., 2011. Comparative study of ROC regression techniques. Applications for the computer-aided diagnostic system in breast cancer detection. Computational Statistics and Data Analysis 55, 888–902.

Swets, J.A., Pickett, R.M., 1982. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press, New York.

Tomé, M.A., Botana, M.A., Cadarso-Suárez, C., Rego-Iraeta, A., Fernández-Mario, A., Mato, J.A, Solache, I., Pérez-Fernández, R., 2008. Prevalence of metabolic syndrome in Galicia (NW Spain) on four alternative definitions and association with insulin resistance. Journal of Endocrinological Investigation 32, 505–511.

Tosteson, A.N., Begg, C.B., 1988. A general regression methodology for ROC curve estimation. Medical Decision Making 8, 204–215.

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, London.

Wood, S., 2003. Thin plate regression splines. Journal of the Royal Statistical Society Series B 65, 95–114.

Yao, F., Craiu, R.V., Reiser, B., 2010. Nonparametric covariate adjustment for receiver operating characteristic curves. The Canadian Journal of Statistics 38, 27–46.

Yusuf, S., Hawken, S., Ounpuu, S., Bautista, L., Franzosi, M.G., Commerford, P., Lang, C.C., Rumboldt, Z., Onen, C.L., Lisheng, L., Tanomsup, S., Wangai Jr., P., Razak, F., Sharma, A.M., Anand, S.S., 2005. Obesity and the risk of myocardial infarction in 27 000 participants from 52 countries: a case-control study. Lancet 366, 1640–1649. On behalf of the INTERHEART Study Investigators.

Zheng, Y., Heagerty, P.J., 2004. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. Biostatistics 4, 615–632.