INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0833

ROC CURVES IN NONPARAMETRIC LOCATION-SCALE REGRESSION MODELS

GONZALEZ-MANTEIGA, W., PARDO-FERNANDEZ, J. C.
and I. VAN KEILEGOM

ROC curves in nonparametric location-scale regression models

Wenceslao González-Manteiga*

Departamento de Estatística e IO

Universidade de Santiago de Compostela

Juan Carlos Pardo-Fernández*

Departamento de Estatística e IO

Universidade de Vigo

Ingrid Van Keilegom*

Institute of Statistics

Université catholique de Louvain

December 22, 2008

Abstract

The receiver operating characteristic curve (ROC curve) is a tool of extensive use to analyse the discrimination capability of a diagnostic variable in medical studies. In certain situations, the presence of a covariate related to the diagnostic variable can increase the discriminating power of the ROC curve. In this article we model the effect of the covariate over the diagnostic variable by means of nonparametric location-scale regression models. We propose a new nonparametric estimator of the conditional ROC curve and study its asymptotic properties. We also present some simulations and an illustration to a data set concerning diagnosis of diabetes.

Key Words: Area under the curve; conditional ROC curve; location-scale regression models; nonparametric regression; relative distribution.

*The research of the three authors is supported by the Spanish Ministerio de Educación y Ciencia (project MTM2005-00820, including European FEDER support). The research of J.C. Pardo-Fernández is also supported by Xunta de Galicia and Universidade de Vigo. The research of I. Van Keilegom is also supported by IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650.

1 Introduction

1.1 ROC curves

In medical studies, or in general in health studies, the diagnosis of an individual or a patient is very often based on a characteristic of interest, which may lead to some classification errors. These classification errors are calibrated on the basis of two indicators: *sensitivity* (probability of diagnosing a diseased person as diseased) and *specificity* (probability of diagnosing a healthy person as healthy).

When the diagnostic characteristic, or diagnostic variable, is of a continuous type, here denoted by Y, the classification will necessarily be based on a *cutoff* value, c: if $Y \geq c$ then the individual is classified as diseased, and if $Y < c$ then the individual is classified as healthy. Let F_1 denote the distribution of Y in the diseased population, and let F_0 denote the distribution of Y in the healthy population. In that case, the geometrical locus is of special interest :

$$\big\{ \left(1 - F_0(c), 1 - F_1(c)\right), c \in \mathbb{R} \big\}, \tag{1.1}$$

which is obtained by varying the cutoff values in the complement of the specificity versus the sensitivity. The geometrical locus (1.1) is called the *receiver operating characteristic curve* (ROC curve), and it is a very extensively used tool to analyse the discrimination power of the diagnostic variable. In practice, the ROC curve is usually reparametrized in the interval $(0, 1)$, as follows:

$$\big\{ \left(p, 1 - F_1(F_0^{-1}(1-p))\right), p \in (0,1) \big\}.$$

The estimation of the ROC curve has been intensively treated in the literature, specially during the last ten years, both from parametric and non-parametric points of view. The book of Pepe (2004) is a general and good reference on this topic.

Several estimators have been proposed when the ROC curve is identified as

$$\mathrm{ROC}(p) = 1 - F_1(F_0^{-1}(1-p)), \;\; 0 < p < 1.$$

For that, assume that two samples, $\{Y_{01}, \ldots, Y_{0n_0}\}$ and $\{Y_{11}, \ldots, Y_{1n_1}\}$, are available from the populations F_0 and F_1, respectively. Those estimates are of the form

$$\widehat{\mathrm{ROC}}(p) = 1 - \hat{F}_1(\hat{F}_0^{-1}(1-p)),$$

where $\hat{F}_0$ and $\hat{F}_1$ are either empirical estimates $\hat{F}_j(t) = F_{jn_j}(t) = n_j^{-1} \sum_{i=1}^{n_j} I(Y_{ji} \leq t)$, or smooth estimates $\hat{F}_j(t) = (F_{jn_j} * K_h)(t)$ (here $K_h(u) = \int_{-\infty}^{u} h^{-1} k(h^{-1}u)du$ is the cumulative distribution function of the rescaled version of the kernel k, h is a bandwidth or smoothing parameter, and $*$ denotes convolution). See, among others, the aforementioned book of Pepe (2004) and the papers by Lloyd (1998), Lloyd and Yong (1999), Zou, Hall and Shapiro (1997), Zhou and Harezlak (2002) and Hall and Hyndman (2003). Other smoothing procedures are treated in the papers by Qiu and Le (2001) and by Peng and Zhou (2004), while Wan and Zhang (2007) present a semiparametric approach. Besides, the ROC curve can also be interpreted in terms of the relative distribution or relative density, see e.g. Handcock and Morris (1999) and Molanes-López (2007).

Related to the ROC curve, several markers, such as the area under the curve (AUC) or the index of Youden, are considered as summaries of the discrimination capability of the ROC curve. The AUC is the most commonly used one and it is given by

$$\text{AUC} = \int_0^1 \text{ROC}(p)dp.$$

Clearly, under the assumption of independence between populations, $\text{AUC} = P(Y_1 > Y_0)$, where Y_0 and Y_1 are random variables with distributions F_0 and F_1, respectively. The AUC takes values between 0.5 (low discrimination power) and 1 (high discrimination power).

A widely used family of ROC curves is obtained when the distributions F_0 and F_1 only differ from their location parameters, μ_0 and μ_1, and scale parameters σ_0 and σ_1. More specifically, when the distributions F_0 and F_1 are Gaussian, the obtained ROC curve is called a *binormal ROC curve*:

$$\text{ROC}(p) = \Phi(a + b\ \Phi^{-1}(p)),$$

where Φ is the cumulative distribution function of a standard normal, Φ^{-1} is the corresponding quantile function, $a = (\mu_1 - \mu_0)/\sigma_1$ and $b = \sigma_0/\sigma_1$. In that case, the area under the curve is simply $\text{AUC} = \Phi(a/\sqrt{1+b^2})$ (see e.g. Pepe (2004), page 83).

1.2 ROC curves with covariates

In many studies, a covariate (or vector of covariates), X, is available along with the diagnostic variable, Y. The information contained in X may increase the discrimination

capability of the ROC curve. A general framework to incorporate the information in the covariate is given by location-scale regression models:

$$Y_0 = \mu_0(X_0) + \sigma_0(X_0)\varepsilon_0, \tag{1.2}$$

$$Y_1 = \mu_1(X_1) + \sigma_1(X_1)\varepsilon_1, \tag{1.3}$$

where, for $j = 0, 1$, $\mu_j(\cdot) = E(Y_j|X_j = \cdot)$ and $\sigma_j^2(\cdot) = Var(Y_j|X_j = \cdot)$ are the conditional mean and conditional variance of the response Y_j given the covariate X_j in each population, respectively, and the error ε_j is independent of X_j.

The parametric case, with $\mu_j(x) = \alpha_j + \beta_j x$ $(j = 0, 1)$ and constant variances, has been studied and applied in the recent literature. See, for instance, Pepe (1997, 1998, 2004) or Faraggi (2003). In the latter paper by Faraggi, a data set concerning fingerstick glucose measurements as a marker for diabetes is analysed and the age of the patients is considered as the covariate. This data set was previously discussed in Smith and Thompson (1996), and we will reconsider it in our illustration in Section 5.

More recently, Zheng and Heagerty (2004) in a context where the diagnostic marker changes over time, estimated the ROC curve induced from model (1.2)-(1.3) on the basis of pilot spline estimators for the mean functions and variance functions.

In other contributions in nonparametric setups, the ROC curve is directly modelled through a generalized linear model of a semiparametric type where the ROC curve is considered as the response variable (see, for instance, Cai and Pepe, 2002).

In this paper, we present a new nonparametric estimator of the conditional ROC curve under the general model (1.2)-(1.3). The estimating process, which makes use of the estimation of the distribution of the regression errors, is described in Section 2. In Section 3 we state several theoretical results concerning the asymptotic behaviour of the proposed estimator. Some simulations are presented in Section 4, and Section 5 contains an illustration to the abovementioned data set. Finally, the appendix contains the proofs of the theoretical results.

2 Methodology

Consider that along with the diagnostic variables in the healthy population, Y_0, and in the diseased population, Y_1, we have two univariate continuous covariates, X_0 and X_1. The

relation between the diagnostic variables and the covariates is established in terms of the nonparametric location-scale regression model (1.2)-(1.3), where we assume for $j = 0, 1$ that $\mu_j(\cdot) = E(Y_j|X_j = \cdot)$ and $\sigma_j^2(\cdot) = \mathrm{Var}(Y_j|X_j = \cdot)$ are unknown smooth functions, and ε_j is independent of X_j. For $j = 0, 1$, let $G_j(y) = P(\varepsilon_j \le y)$, $F_j(y|x) = P(Y_j \le y|X_j = x)$ and $F_{X_j}(x) = P(X_j \le x)$, and denote the support of X_j by R_{X_j}. The intersection of R_{X_0} and R_{X_1} is denoted by R_X and is supposed to be non-empty. The probability density functions of the above distributions will be denoted by lower case letters (i.e., $g_j(y)$, $f_j(y|x)$ and f_{X_j}, for $j = 0, 1$).

For a fixed value x in R_X, the conditional ROC curve is defined by, for $0 < p < 1$,

$$\begin{aligned}\mathrm{ROC}_x(p) &= 1 - F_1(F_0^{-1}(1-p|x)|x) \\ &= 1 - G_1\Big(\sigma_1^{-1}(x)\{G_0^{-1}(1-p)\sigma_0(x) + \mu_0(x) - \mu_1(x)\}\Big) \\ &= 1 - G_1\Big(G_0^{-1}(1-p)b(x) - a(x)\Big),\end{aligned}$$

where

$$a(x) = \frac{\mu_1(x) - \mu_0(x)}{\sigma_1(x)} \qquad \text{and} \qquad b(x) = \frac{\sigma_0(x)}{\sigma_1(x)},$$

and where for any distribution function F and any $0 \le s \le 1$, $F^{-1}(s) = \inf\{y : F(y) \ge s\}$. Suppose we have a sample $(X_{01}, Y_{01}), \ldots, (X_{0n_0}, Y_{0n_0})$ of i.i.d. data generated from model (1.2) and another sample $(X_{11}, Y_{11}), \ldots, (X_{1n_1}, Y_{1n_1})$ of i.i.d. data generated from model (1.3), that is independent of the first sample. Let $N = n_0 + n_1$. Based on these data, we propose the following estimator of the conditional ROC curve:

$$\widehat{\mathrm{ROC}}_x(p) = 1 - \int \hat{G}_1\Big(\hat{G}_0^{-1}(1-p+hu)\hat{b}(x) - \hat{a}(x)\Big)k(u)du, \tag{2.1}$$

where k is a probability density function (kernel), $h = h_N$ is a bandwidth sequence, and for $j = 0, 1$,

$$\begin{aligned}&\hat{G}_j(y) = n_j^{-1}\sum_{i=1}^{n_j} I(\hat{\varepsilon}_{ji} \le y), \\ &\hat{\varepsilon}_{ji} = \frac{Y_{ji} - \hat{\mu}_j(X_{ji})}{\hat{\sigma}_j(X_{ji})} \quad (i = 1, \ldots, n_j), \\ &\hat{\mu}_j(x) = \sum_{i=1}^{n_j} W_{ji}(x, g)Y_{ji}, \qquad \hat{\sigma}_j^2(x) = \sum_{i=1}^{n_j} W_{ji}(x, g)[Y_{ji} - \hat{\mu}_j(X_{ji})]^2,\end{aligned}$$

and

$$W_{ji}(x,g) = \frac{k_g(x - X_{ji})}{\sum_{l=1}^{n_j} k_g(x - X_{jl})},$$

with $g = g_N$ a second bandwidth sequence, and $k_g(\cdot) = k(\cdot/g)/g$. Finally, $\hat{a}(x) = [\hat{\mu}_1(x) - \hat{\mu}_0(x)]/\hat{\sigma}_1(x)$ and $\hat{b}(x) = \hat{\sigma}_0(x)/\hat{\sigma}_1(x)$. Note that $\widehat{\mathrm{ROC}}_x(p)$ can also be written as :

$$\widehat{\mathrm{ROC}}_x(p) = 1 - \frac{1}{n_1}\sum_{i=1}^{n_1} K\left(\frac{\hat{G}_0(\{\hat{\varepsilon}_{1i} + \hat{a}(x)\}/\hat{b}(x)) - 1 + p}{h}\right),$$

where K is the distribution function corresponding to the kernel k.

The ROC curve is defined in terms of distribution functions of continuous random variables, and hence it is a continuous curve. This motivates the construction of the smooth estimator proposed in (2.1), which ensures that the estimated ROC curve is also continuous. The bandwidth h determines the smoothness of the estimated ROC curve.

Also note that the estimator of the conditional ROC curve given in (2.1) can be considered simply in terms of empirical distributions of the regression residuals, without adding any smoothing to the ROC curve, by taking $h = 0$:

$$\widetilde{\mathrm{ROC}}_x(p) = 1 - \hat{G}_1\Big(\hat{G}_0^{-1}(1-p)\hat{b}(x) - \hat{a}(x)\Big). \tag{2.2}$$

This estimator, which we can call the "empirical" conditional ROC curve estimator, is also a valid estimator of the conditional ROC curve, but it has the drawback of not being continuous.

On the other hand, the bandwidth g is used to locally estimate the regression and variance functions. In principle, one could use different bandwidths for each of the curves $\mu_0(x), \mu_1(x), \sigma_0(x)$ and $\sigma_1(x)$, but for simplicity of presentation we will restrict here to one bandwidth.

Other estimators of $\mathrm{ROC}_x(p)$ can be considered, based on smoothing of each of the empirical distributions $\hat{G}_0(\cdot)$ and $\hat{G}_1(\cdot)$. See e.g. Hall and Hyndman (2003) and Qiu and Le (2001) for the case without covariates. We follow here the approach used, among others, by Peng and Zhou (2004) and López-de Ullibarri et al. (2008) and apply smoothing on the ROC curve itself.

3 Main result

The following result is an i.i.d. representation for the ROC-process $\widehat{\mathrm{ROC}}_x(p) - \mathrm{ROC}_x(p)$. Note that the main term of this representation does not depend on the bandwidth h, as its contribution is asymptotically negligible. The assumptions under which the results below are valid, are given in the appendix.

Theorem 3.1 *Assume (A1)-(A3). Then, for $0 < p < 1$ and for a fixed x in R_X,*

$$\begin{aligned}
&\widehat{ROC}_x(p) - ROC_x(p) \\
&= g_1(G_0^{-1}(1-p)b(x) - a(x))\Big\{\hat{A}_x + G_0^{-1}(1-p)\hat{B}_x\Big\} + g^2\beta_x(p) + \hat{R}_x(p),
\end{aligned}$$

where

$$\hat{A}_x = \sigma_1^{-1}(x)\sum_{j=0}^{1}(-1)^{j+1}f_{X_j}^{-1}(x)n_j^{-1}\sum_{i=1}^{n_j}k_g(x - X_{ji})(Y_{ji} - \mu_j(X_{ji})),$$

$$\hat{B}_x = \frac{1}{2}\sigma_1^{-2}(x)\sum_{j=0}^{1}(-1)^{j+1}\Big(\frac{\sigma_0(x)}{\sigma_1(x)}\Big)^{2j-1}f_{X_j}^{-1}(x)n_j^{-1}\sum_{i=1}^{n_j}k_g(x - X_{ji})\sigma_j^2(X_{ji})(\varepsilon_{ji}^2 - 1),$$

$$\begin{aligned}
\beta_x(p) = &-\frac{1}{2}\mu_2^k\int\frac{\partial^2}{\partial t^2}E[\varphi(t, Y_1, c_x(1-p))|X_1 = v]|_{t=v}\,dF_{X_1}(v) \\
&+\frac{1}{2}\mu_2^k g_1(G_0^{-1}(1-p)b(x) - a(x))\Big\{\sigma_1^{-1}(x)\sum_{j=0}^{1}(-1)^{j+1}\Big[\mu_j''(x) + 2\mu_j'(x)\frac{f_{X_j}'(x)}{f_{X_j}(x)}\Big] \\
&+G_0^{-1}(1-p)\sigma_1^{-2}(x)\frac{1}{2}\sum_{j=0}^{1}(-1)^{j+1}\Big(\frac{\sigma_0(x)}{\sigma_1(x)}\Big)^{2j-1}\Big[(\sigma_j^2(x))'' + 2(\sigma_j^2(x))'\frac{f_{X_j}'(x)}{f_{X_j}(x)}\Big]\Big\},
\end{aligned}$$

$$\varphi(x, y, z) = g_1(z)\sigma_1^{-1}(x)\Big[y - \mu_1(x) + \frac{z}{2\sigma_1(x)}\big\{(y - \mu_1(x))^2 - \sigma_1^2(x)\big\}\Big],$$

and where $\mu_2^k = \int u^2k(u)\,du$ and $\sup_{\delta<p<1-\delta}|\hat{R}_x(p)| = o_P((Ng)^{-1/2})$, for any small $\delta > 0$.

As a consequence, we get the weak convergence of the ROC-process. The proof can be obtained by applying the central limit theorem for triangular arrays to the random variables $(Ng)^{1/2}\hat{A}_x$ and $(Ng)^{1/2}\hat{B}_x$. Both the case of undersmoothing ($C = 0$) and the optimal bandwidth $g = C^{1/5}N^{-1/5}$ with $0 < C < \infty$ are considered.

Corollary 3.2 *Assume (A1)-(A3). Then, for a fixed x in R_X and for a small $\delta > 0$, the process $(Ng)^{1/2}(\widehat{ROC}_x(p) - ROC_x(p))$ $(\delta < p < 1-\delta)$ converges weakly to a Gaussian process*

$$W_x(p) = g_1(G_0^{-1}(1-p)b(x) - a(x))\Big\{W_{1x} + G_0^{-1}(1-p)W_{2x}\Big\} + C^{1/2}\beta_x(p),$$

where C is defined in assumption (A1), and where W_{1x} and W_{2x} are normal random variables with zero mean, and

$$Var(W_{1x}) = \sigma_1^{-2}(x)\|k\|_2^2 \sum_{j=0}^{1} f_{X_j}^{-1}(x)\lambda_j^{-1}\sigma_j^2(x)$$

$$Var(W_{2x}) = \frac{1}{4}\frac{\sigma_0^2(x)}{\sigma_1^2(x)}\|k\|_2^2 \sum_{j=0}^{1} f_{X_j}^{-1}(x)\lambda_j^{-1}E(\varepsilon_j^4 - 1)$$

$$Cov(W_{1x}, W_{2x}) = \frac{1}{2}\frac{\sigma_0(x)}{\sigma_1^2(x)}\|k\|_2^2 \sum_{j=0}^{1} f_{X_j}^{-1}(x)\lambda_j^{-1}\sigma_j(x)E(\varepsilon_j^3),$$

with $\lambda_j = \lim_{N\to\infty} n_j/N$ $(j = 0, 1)$, and where $\|k\|_2^2 = \int k^2(u)\,du$.

This result can now be used to obtain the limiting distribution of any continuous functional of the ROC-process. A well known particular case is the conditional version of the so-called *area under the curve* (AUC), which, for a fixed x in R_X, we define by

$$\mathrm{AUC}_x = \int_\delta^{1-\delta} \mathrm{ROC}_x(p)\,dp. \tag{3.1}$$

For technical reasons, we restrict the integration to the interval to $[\delta, 1-\delta]$, which can however be made arbitrarily close to $[0, 1]$. The estimator is

$$\widehat{\mathrm{AUC}}_x = \int_\delta^{1-\delta} \widehat{\mathrm{ROC}}_x(p)\,dp.$$

The proof of the following result is an immediate consequence of the continuous mapping theorem.

Corollary 3.3 *Assume (A1)-(A3). Then, for a fixed x in R_X,*

$$(Ng)^{1/2}(\widehat{AUC}_x - AUC_x) \xrightarrow{d} N(0, s_x^2),$$

where

$$
\begin{aligned}
s_x^2 &= Var\Big(\int_\delta^{1-\delta} W_x(p)\, dp \Big) \\
&= \gamma_{1x}^2\, Var(W_{1x}) + \gamma_{2x}^2\, Var(W_{2x}) + 2\gamma_{1x}\gamma_{2x}\, Cov(W_{1x}, W_{2x}), \\
\gamma_{1x} &= \int_\delta^{1-\delta} g_1\big(G_0^{-1}(1-p)b(x) - a(x)\big)\, dp, \\
\gamma_{2x} &= \int_\delta^{1-\delta} g_1\big(G_0^{-1}(1-p)b(x) - a(x)\big) G_0^{-1}(1-p)\, dp.
\end{aligned}
$$

4 Simulations

In this section we present a small simulation study. We are mainly interested in the global performance of the proposed estimator of the conditional ROC curve and in the effect of the smoothing parameter h. We have simulated data from two scenarios:

- **Scenario 1:**

  Regression functions: $\mu_0(x) = 0$; $\mu_1(x) = x$.

  Conditional variance functions: $\sigma_0^2(x) = \sigma_1^2(x) = 0.5^2$.

- **Scenario 2:**

  Regression functions: $\mu_0(x) = 0.5\sin(2\pi x)$; $\mu_1(x) = \sin(\pi x)$.

  Conditional variance functions: $\sigma_0^2(x) = \sigma_1^2(x) = (0.25 + 0.5x)^2$.

In both scenarios, the covariates X_0 and X_1 are uniformly distributed on [0,1], and the regression errors ε_0 and ε_1 have standard normal distribution. The true ROC curves, presented here as a surface, and the true conditional AUC, presented as a function of the values of the covariate, are depicted in Figure 1 (scenario 1) and Figure 4 (scenario 2).

The estimator of the conditional ROC curves was calculated on a grid of points of the form $\{(x_l, p_r) \in (0,1) \times (0,1), l = 1, \ldots, n_x, r = 1, \ldots, n_p\}$. More precisely, in all cases we take

$$
\begin{aligned}
x_l &= 0.05 + (l-1)\frac{0.90}{n_x - 1}, \text{ for } l = 1, \ldots, n_x, \\
p_r &= 0.05 + (r-1)\frac{0.90}{n_p - 1}, \text{ for } r = 1, \ldots, n_p,
\end{aligned}
$$

with $n_x = 25$ and $n_p = 25$. The estimators of the regression curves, $\mu_0(\cdot)$ and $\mu_1(\cdot)$, and variance curves, $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$, which are needed in the construction of the estimator of the conditional ROC curve, are based on the kernel of Epanechnikov $k(u) = 0.75(1 - u^2)I(|u| < 1)$ and on cross-validation bandwidths: for $j = 0, 1$, a regular cross-validation procedure is used to estimate μ_j, and then the same bandwidth is used to estimate σ_j^2.

The discrepancy between the estimator and the true ROC surface is measured in terms of the empirical version of the global mean squared error (MSE):

$$\text{MSE} = \frac{1}{n_x}\sum_{l=1}^{n_x}\frac{1}{n_p}\sum_{r=1}^{n_p}\left(\widehat{\text{ROC}}_{x_l}(p_r) - \text{ROC}_{x_l}(p_r)\right)^2.$$

Table 1 displays the averages and standard deviations of the MSEs obtained in 1000 data sets simulated from Scenario 1. The estimators of the ROC curves were calculated with different values of the smoothing parameter h, ranging from 0 to 0.25. The case $h = 0$ corresponds to the empirical estimator given in (2.2). As expected, the MSE decreases as the sample sizes increase. The effect of the parameter h is not very important, although introducing a small amount of smoothing in the estimator produces a better behaviour in terms of MSE with respect to the empirical estimator. The required amount of smoothing to improve the MSE decreases as the sample sizes get larger. Figure 2 shows the boxplots of the 1000 estimated MSEs for several sample sizes and several values of the smoothing parameter. Finally, we have also considered the estimation of the conditional AUC, as defined in (3.1), where we take $\delta = 0.05$. Figure 3 shows the average of the estimated AUC for several sample sizes, with $h = 0.10$. As a reference, we have also included in the graph two bands which correspond to ± 2 times the standard deviation of the estimator of the AUC in the 1000 data sets. The general performance of the estimator of the conditional AUC is good.

Table 2, Figure 5 and Figure 6 show the corresponding results when the data sets are simulated from Scenario 2. Similar conclusions can be stated in this case. The lowest values of the MSE are achieved with values of the smoothing parameter h smaller than the corresponding ones in Scenario 1.

5 Data analysis

As an illustration of the proposed methodology, we present an application to a data set concerning diagnosis of diabetes. This data set has also been analysed in Faraggi (2003) and Smith and Thompson (1996).

The data come from a population-based pilot survey of diabetes mellitus in Cairo (Egypt), and consist of post-prandial blood glucose measurements of 286 subjects obtained from a fingerstick. According to the gold standard criteria of the World Health Organization for diagnosing diabetes, 88 subjects were classified as diseased and 198 subjects were classified as healthy. The age of the subject was considered as a relevant covariate in this example, because due to medical reasons (see Smith and Thompson (1996) for the details) glucose levels are expected to be higher for older persons who do not suffer from diabetes.

Figure 7 shows the scatter plot of the data for both the healthy and diseased population. The glucose concentration is considered as the diagnostic variable, and the age of the subject as a covariate. We have estimated the conditional ROC curves with the methodology proposed in Section 2 in the values of the covariate $x = 20, 21, \ldots, 90$. The analysis has been performed with several values for the smoothing parameter h, and very similar results were obtained. Figure 8-(a) shows the complete ROC surface estimated with $h = 0.10$. We will keep this value of the smoothing parameter in the rest of the figures. To check visually the effect of the age on the ROC curves, the conditional ROC curves for ages 30, 50 and 70 are depicted in Figure 8-(b). Clearly, the aging process reduces the capability of the ROC curve to discriminate between diseased and healthy subjects.

The effect of the age on the discrimination power of the ROC curve can be summarized by means of the AUC. Figure 9 shows the AUC as a function of the values of the covariate. As in the simulation study, we use definition (3.1) with $\delta = 0.05$. We have also included in the graph confidence intervals for the AUC obtained by bootstrap. Asymptotic confidence intervals for AUC_x could be obtained from Corollary 3.3, but the asymptotic variance of the estimator depends on certain unknown quantities that are difficult to estimate. Alternatively, we use the following bootstrap procedure: for fixed x, and for $b = 1, ..., B$,

1. For $j = 0, 1$, let $\{\varepsilon^*_{ji,b}, i = 1, \ldots, n_j\}$ be an i.i.d. sample from $\hat{G}_j$.

2. Reconstruct the bootstrap samples $\{(X_{ji}, Y^*_{ji}), i = 1, \ldots, n_j\}$, for $j = 0, 1$, where $Y^*_{ji,b} = \hat{\mu}_j(X_{ji}) + \hat{\sigma}_j(X_{ji})\varepsilon^*_{ji,b}$.

3. Repeat the estimation process with the bootstrap samples to obtain $\mathrm{AUC}^*_{x,b}$.

Let $\mathrm{AUC}^*_{x,(b)}$ be the order statistics of the values $\mathrm{AUC}^*_{x,1}, \ldots, \mathrm{AUC}^*_{x,B}$ obtained in step 3. According to the percentile method, $(\mathrm{AUC}^*_{x,(\lfloor B\alpha/2 \rfloor)}, \mathrm{AUC}^*_{x,(\lfloor B(1-\alpha/2) \rfloor)})$ is a bootstrap confidence interval for AUC_x of confidence level $1 - \alpha$ ($\lfloor \cdot \rfloor$ denotes the integer part). In the graph, we represent the bootstrap confidence intervals of levels 90% and 95% obtained with $B = 1000$ replications for the AUC with respect to the values of the covariate. As seen before, the age of the subject clearly has an important impact on the discrimination power of the glucose measurements as an indicator of diabetes.

Similar conclusions can be found in Faraggi (2003), although this author works under a much more restrictive model (linear regression models with homoscedastic normal errors). The advantage of our method is the flexibility incorporated by the nonparametric and heteroscedastic regression models.

Appendix : Proofs

Assumptions

(A1) (*i*) $n_j/N \to \lambda_j$ for some $0 < \lambda_j < 1$ $(j = 0, 1)$. Moreover, $Ng^5 \to C$ for some $0 \leq C < \infty$, $Ng^{3+\alpha}(\log g^{-1})^{-1} \to \infty$ for some $\alpha > 0$ and $Nh^4 g \to 0$.
(*ii*) R_{X_j} is a bounded interval in $\mathbb{R}$ $(j = 0, 1)$.
(*iii*) k has compact support, $\int uk(u)du = 0$ and k is twice continuously differentiable.

(A2) (*i*) F_{X_j} is three times continuously differentiable and $\inf_{x \in R_{X_j}} f_{X_j}(x) > 0$ $(j = 0, 1)$.
(*ii*) μ_j and σ_j are twice continuously differentiable and $\inf_{x \in R_{X_j}} \sigma_j(x) > 0$ $(j = 0, 1)$.

(A3) G_j is three times continuously differentiable and $\sup_y |y^2 G_j^{(k)}(y)| < \infty$ for $k = 1, 2, 3$ and $j = 0, 1$. Moreover, for any $\delta > 0$, $\inf_{\delta < p < 1-\delta} g_0(G_0^{-1}(p)) > 0$.

Proof of Theorem 3.1. For any $0 < s < 1$, let $c_x(s) = G_0^{-1}(s)b(x) - a(x)$ and $\hat{c}_x(s) = \hat{G}_0^{-1}(s)\hat{b}(x) - \hat{a}(x)$. Write

$$\begin{aligned}
&\widehat{\mathrm{ROC}}_x(p) - \mathrm{ROC}_x(p) \\
&= -\int \big\{\hat{G}_1(\hat{c}_x(1-p+hu)) - E[\hat{G}_1(s)]|_{s=\hat{c}_x(1-p+hu)}\big\}k(u)\,du \\
&\quad -\int \big\{E[\hat{G}_1(s)]|_{s=\hat{c}_x(1-p+hu)} - G_1(\hat{c}_x(1-p+hu))\big\}k(u)\,du \\
&\quad -\int \big\{G_1(\hat{c}_x(1-p+hu)) - G_1(c_x(1-p+hu))\big\}k(u)\,du \\
&\quad -\int \big\{G_1(c_x(1-p+hu)) - G_1(c_x(1-p))\big\}k(u)\,du \\
&= T_{1x}(p) + T_{2x}(p) + T_{3x}(p) + T_{4x}(p).
\end{aligned}$$

We start with $T_{1x}(p)$. Using Corollary 2 in Akritas and Van Keilegom (2001) it follows that $\sup_y |\hat{G}_1(y) - E[\hat{G}_1(y)]| = O_P(N^{-1/2})$, and hence, $\sup_{\delta<p<1-\delta} |T_{1x}(p)| = o_P((Ng)^{-1/2})$. On the other hand,

$$\begin{aligned}
T_{2x}(p) &= -\frac{1}{2}g^2\mu_2^k \int\int \frac{\partial^2}{\partial t^2} E[\varphi(t, Y_1, s)|X_1 = v]|_{t=v, s=\hat{c}_x(1-p+hu)}\, dF_{X_1}(v)k(u)\,du + o_P(g^2) \\
&= -\frac{1}{2}g^2\mu_2^k \int \frac{\partial^2}{\partial t^2} E[\varphi(t, Y_1, c_x(1-p))|X_1 = v]|_{t=v}\, dF_{X_1}(v) + o_P(g^2).
\end{aligned}$$

Next, by condition (A3) we have that $\sup_{\delta<p<1-\delta} |T_{4x}(p)| = O(h^2) = o((Ng)^{-1/2})$ if $Nh^4g \to 0$. It remains to consider $T_{3x}(p)$:

$$\begin{aligned}
T_{3x}(p) &= -\int g_1(G_0^{-1}(1-p+hu)b(x) - a(x))\Big\{G_0^{-1}(1-p+hu)[\hat{b}(x) - b(x)] \\
&\qquad -[\hat{a}(x) - a(x)]\Big\}k(u)\,du + O_P((Ng)^{-1}\log N) + O_P(n_0^{-1/2}(\log n_0)^{1/2}) \\
&= -g_1(G_0^{-1}(1-p)b(x) - a(x))\Big\{G_0^{-1}(1-p)[\hat{b}(x) - b(x)] - [\hat{a}(x) - a(x)]\Big\} \\
&\qquad +O_P(N^{-1/2}(\log N)^{1/2}) + O(h^2), \qquad (A.1)
\end{aligned}$$

which follows from Lemma A.1 below and since $\hat{\mu}_j(x) - \mu_j(x) = O_P((Ng)^{-1/2})$ and $\hat{\sigma}_j(x) - \sigma_j(x) = O_P((Ng)^{-1/2})$ $(j = 0, 1)$. Next, note that

$$\begin{aligned}
&G_0^{-1}(1-p)[\hat{b}(x) - b(x)] - [\hat{a}(x) - a(x)] \\
&= G_0^{-1}(1-p)\sigma_1^{-2}(x)\Big[(\hat{\sigma}_0(x) - \sigma_0(x))\sigma_1(x) - (\hat{\sigma}_1(x) - \sigma_1(x))\sigma_0(x)\Big] \\
&\quad -\sigma_1^{-1}(x)\Big[\hat{\mu}_1(x) - \mu_1(x) - \hat{\mu}_0(x) + \mu_0(x)\Big] + O_P((Ng)^{-1}\log N), \qquad (A.2)
\end{aligned}$$

and that for $j = 0, 1$,

$$\hat{\mu}_j(x) - \mu_j(x) = f_{X_j}^{-1}(x) n_j^{-1} \sum_{i=1}^{n_j} k_g(x - X_{ji})(Y_{ji} - \mu_j(X_{ji}))$$
$$+ \frac{g^2}{2}\Big[\mu_j''(x) + 2\mu_j'(x)\frac{f_{X_j}'(x)}{f_{X_j}(x)}\Big]\mu_2^k + o_P((Ng)^{-1/2}), \tag{A.3}$$

$$\hat{\sigma}_j(x) - \sigma_j(x) = \frac{1}{2}\sigma_j^{-1}(x) f_{X_j}^{-1}(x) n_j^{-1} \sum_{i=1}^{n_j} k_g(x - X_{ji})[(Y_{ji} - \mu_j(X_{ji}))^2 - \sigma_j^2(X_{ji})]$$
$$+ \frac{g^2}{4\sigma_j(x)}\Big[(\sigma_j^2(x))'' + 2(\sigma_j^2(x))'\frac{f_{X_j}'(x)}{f_{X_j}(x)}\Big]\mu_2^k + o_P((Ng)^{-1/2}). \tag{A.4}$$

The result now follows, by combining (A.1), (A.2), (A.3) and (A.4). □

Lemma A.1 *Assume (A1)-(A3). Then, for any small $\delta > 0$,*

$$\sup_{\delta < s < 1-\delta} |\hat{G}_0^{-1}(s) - G_0^{-1}(s)| = O_P(n_0^{-1/2}).$$

Proof. Let $I_\delta = [\delta, 1-\delta]$, let $\alpha_n = K_\varepsilon n_0^{-1/2}$ for some $K_\varepsilon > 0$ and some $\varepsilon > 0$. Then,

$$P\Big(\sup_{s \in I_\delta} |\hat{G}_0^{-1}(s) - G_0^{-1}(s)| > \alpha_n\Big)$$
$$\le P\Big(\hat{G}_0^{-1}(s) > G_0^{-1}(s) + \alpha_n \text{ for some } s \in I_\delta\Big)$$
$$+ P\Big(\hat{G}_0^{-1}(s) < G_0^{-1}(s) - \alpha_n \text{ for some } s \in I_\delta\Big)$$
$$= T_1 + T_2.$$

In what follows, we consider the term T_1. The term T_2 can be treated in a very similar way.

$$T_1 \le P\Big(\hat{G}_0(G_0^{-1}(s) + \alpha_n) < s \text{ for some } s \in I_\delta\Big)$$
$$\le P\Big(\sup_y |\hat{G}_0(y) - G_0(y)| > G_0(G_0^{-1}(s) + \alpha_n) - s \text{ for some } s \in I_\delta\Big)$$
$$= P\Big(\sup_y |\hat{G}_0(y) - G_0(y)| > \inf_{s \in I_\delta}\{G_0(G_0^{-1}(s) + \alpha_n) - s\}\Big)$$
$$\le P\Big(\sup_y |\hat{G}_0(y) - G_0(y)| > K_1\alpha_n\Big),$$

since $\inf_{s\in I_\delta}\{G_0(G_0^{-1}(s)+\alpha_n)-s\} > \inf_{\delta/2<s<1-\delta/2} g_0(G_0^{-1}(s))\alpha_n > K_1\alpha_n$ for some $K_1 > 0$. The latter probability is bounded by ε for K_ε and n_0 large enough, since $\sup_y |\hat{G}_0(y) - G_0(y)| = O_P(n_0^{-1/2})$ (see Corollary 2 in Akritas and Van Keilegom (2001)). □

References

Akritas, M. and Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scandinavian Journal of Statistics*, 28, 549-567.

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

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

Hall, P.G. and Hyndman, R.J. (2003). Improved methods for bandwidth selection when estimating ROC curves. *Statistics and Probability Letters*, 64, 181-189.

Handcock, M.S. and Morris, M. (1999). *Relative distribution methods in the social sciences*. Springer, New York.

Lloyd, C.J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, 93, 1356-1364.

Lloyd, C.J. and Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical. *Statistics and Probability Letters*, 44, 221-228.

López-de Ullibarri, I., Cao, R., Cadarso-Suárez, C. and 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.

Molanes-López, E.M. (2007). *Nonparametric statistical inference for relative curves in two-sample problems*. PhD thesis. University of A Coruña.

Peng, L. and Zhou, X.-H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference*, 118, 129-143.

Pepe, M.S. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84, 595-608.

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. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, Oxford.

Qiu, P. and Le, C. (2001). ROC curve estimation based on local smoothing. *Journal of Statistical Computation and Simulation*, 70, 55-69.

Smith, P.J. and Thompson, T.J. (1996). Correcting for confounding in analyzing receiver operating characteristic curves. *Biometrical Journal*, 38, 857-863.

Wan, S. and Zhang, B. (2007). Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. *Statistics in Medicine*, 26, 2565-2586.

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

Zhou, X.-H. and Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine*, 21, 2045-2055.

Zou, K.H, Hall, W.J and Shapiro, D.E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16, 2143-2156.

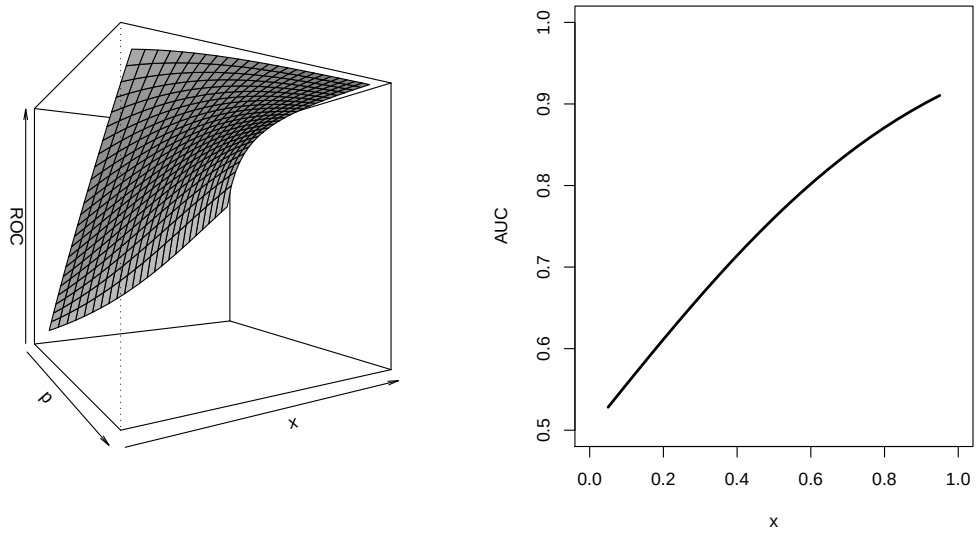
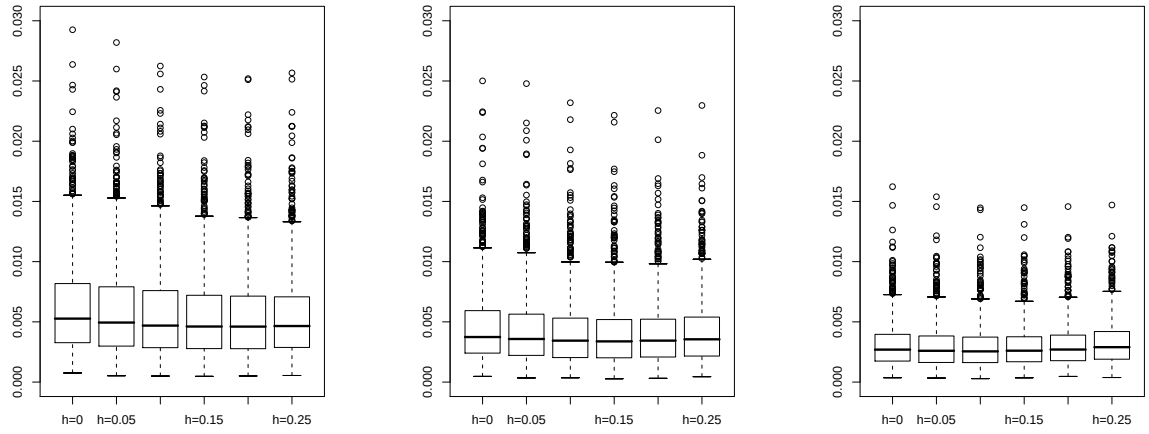


Figure 1: Conditional ROC curves (left) and conditional AUC (right) for Scenario 1.

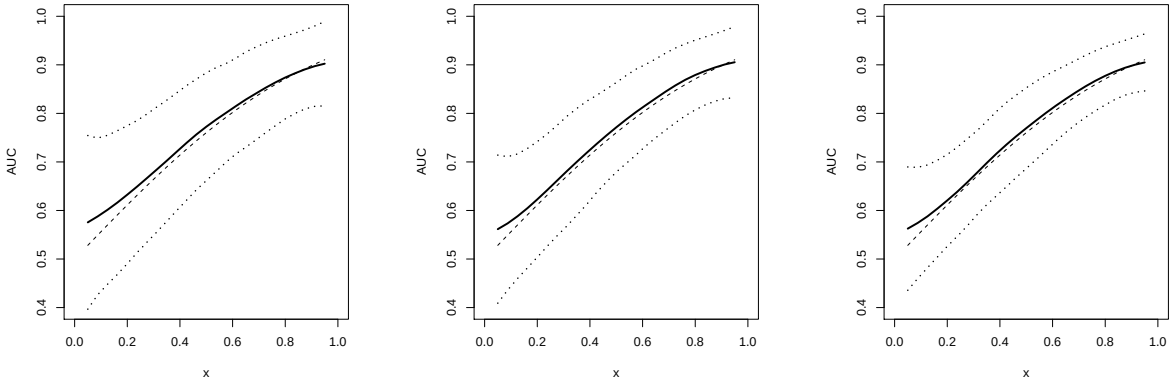
			h					
n_0	n_1		0.00	0.05	0.10	0.15	0.20	0.25
100	100	average	6.411	6.097	5.820	5.644	5.570	5.597
		sd	4.370	4.304	4.178	4.058	3.965	3.903
100	200	average	4.622	4.372	4.170	4.074	4.077	4.178
		sd	3.160	3.111	2.998	2.909	2.856	2.836
200	200	average	3.122	2.994	2.927	2.954	3.067	3.269
		sd	1.905	1.880	1.836	1.818	1.829	1.866

Table 1: Average and standard deviation (sd) of the estimated MSE ($\times 1000$) obtained from 1000 data sets simulated according to Scenario 1, for different sample sizes and different values of the smoothing parameter h.



$n_0 = 100$, $n_1 = 100$ $n_0 = 100$, $n_1 = 200$ $n_0 = 200$, $n_1 = 200$

Figure 2: Boxplots of the estimated MSE obtained from 1000 data sets simulated from Scenario 1, for different sample sizes and different values of the smoothing parameter h.



$n_0 = 100$, $n_1 = 100$ $n_0 = 100$, $n_1 = 200$ $n_0 = 200$, $n_1 = 200$

Figure 3: Average of the estimated conditional AUC (solid line) ± 2 times its standard deviation (dotted lines) obtained from 1000 data sets simulated from Scenario 1, for different sample sizes. In all cases $h = 0.10$. The dashed line represents the true AUC.

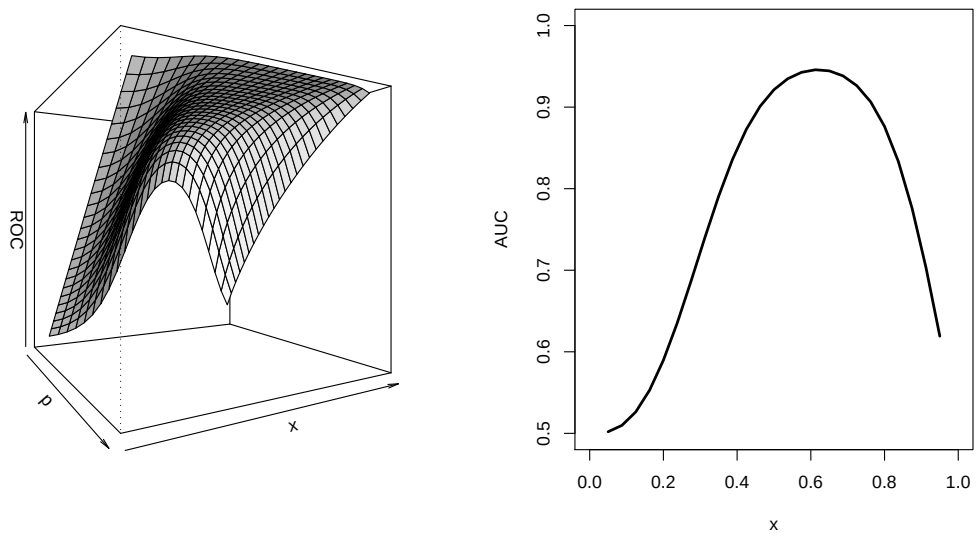
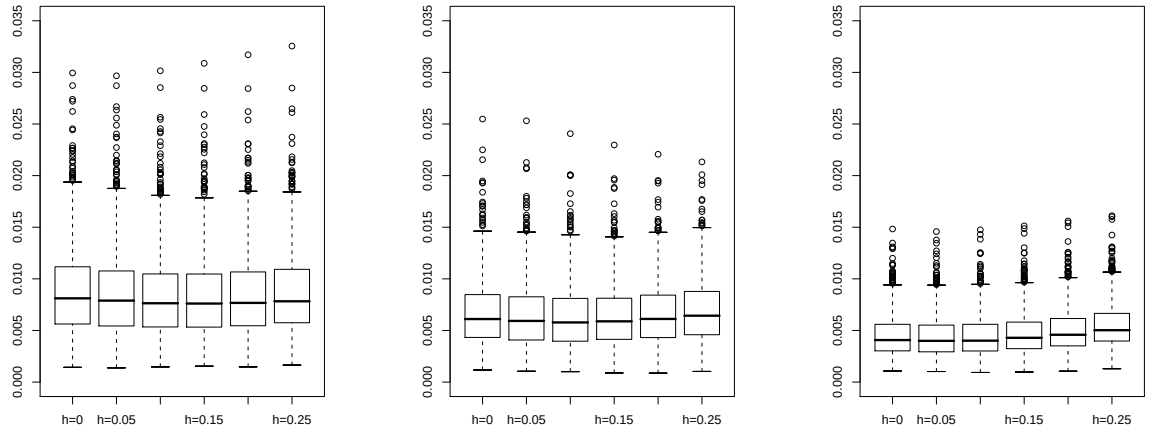


Figure 4: Conditional ROC curves (left) and conditional AUC (right) for Scenario 2.

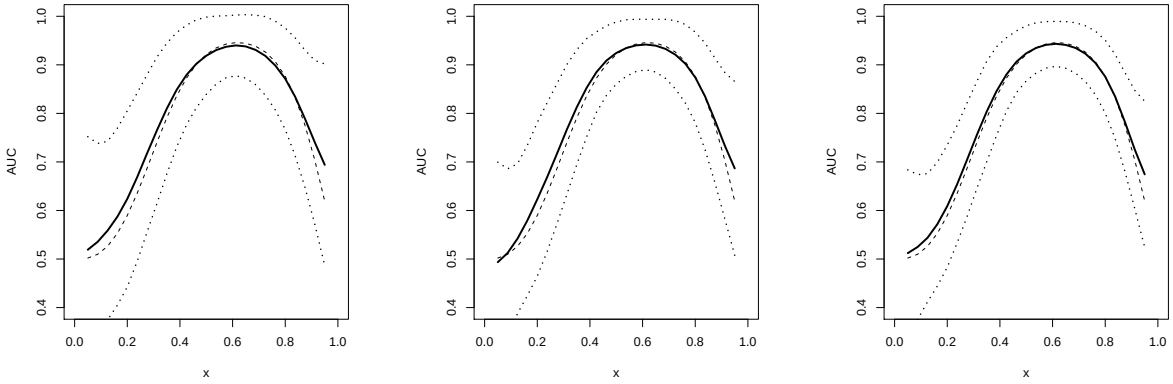
			h					
n_0	n_1		0.00	0.05	0.10	0.15	0.20	0.25
100	100	average	8.849	8.587	8.383	8.360	8.477	8.714
		sd	4.376	4.310	4.224	4.149	4.095	4.062
100	200	average	6.704	6.492	6.349	6.398	6.587	6.896
		sd	3.215	3.155	3.086	3.038	3.014	3.006
200	200	average	4.529	4.424	4.460	4.678	5.021	5.473
		sd	2.092	2.071	2.068	2.089	2.128	2.179

Table 2: Average and standard deviation (sd) of the estimated MSE ($\times 1000$) obtained from 1000 data sets simulated according to Scenario 2, for different sample sizes and different values of the smoothing parameter h.



$n_0 = 100$, $n_1 = 100$ $n_0 = 100$, $n_1 = 200$ $n_0 = 200$, $n_1 = 200$

Figure 5: Boxplots of the estimated MSE obtained from 1000 data sets simulated from Scenario 2, for different sample sizes and different values of the smoothing parameter h.



$n_0 = 100$, $n_1 = 100$ $n_0 = 100$, $n_1 = 200$ $n_0 = 200$, $n_1 = 200$

Figure 6: Average of the estimated conditional AUC (solid line) ± 2 times the standard deviation (dotted lines) obtained from 1000 data sets simulated from Scenario 2, for different sample sizes. In all cases $h = 0.10$. The dashed line represents the true AUC.

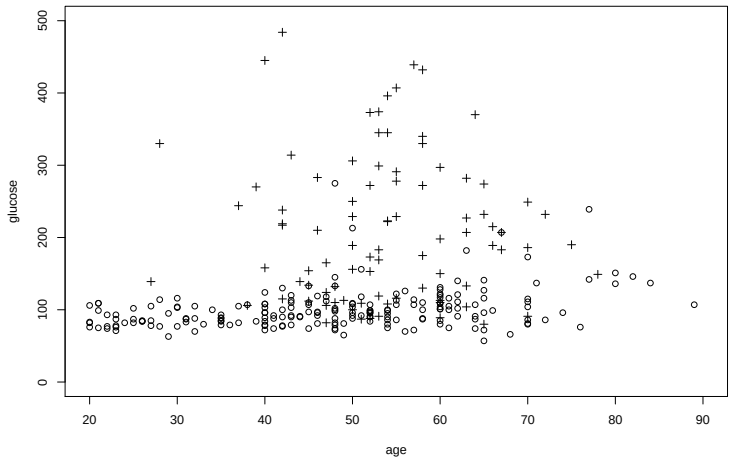
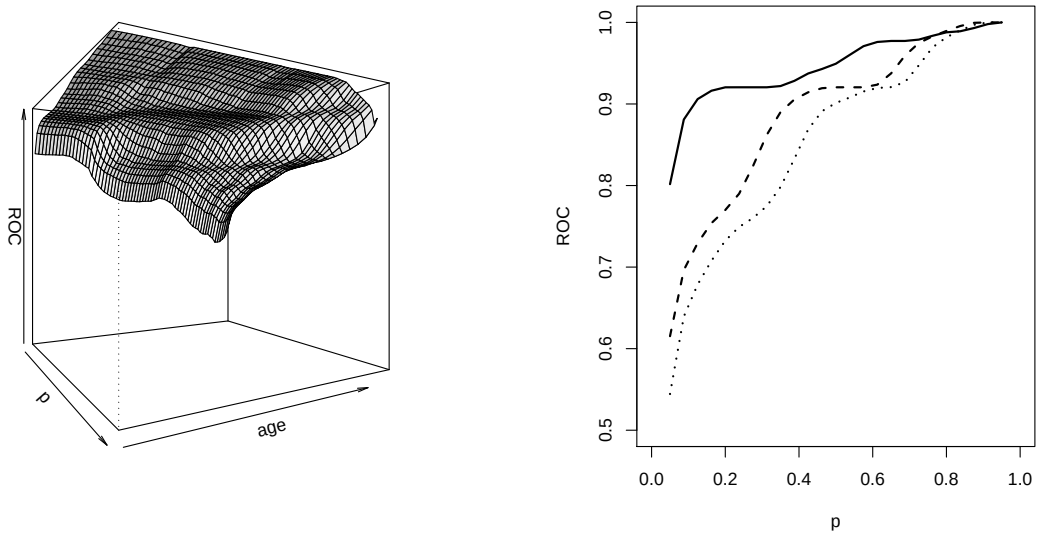


Figure 7: Scatter plot of the diagnostic variable 'glucose concentration' with respect to the covariate 'age of the subject'. The diseased population is represented by crosses and the healthy population is represented by circles.



(a) (b)

Figure 8: (a) Estimated conditional ROC curves. (b) Conditional ROC curves for ages 30 (solid line), 50 (dashed line) and 70 (dotted line).

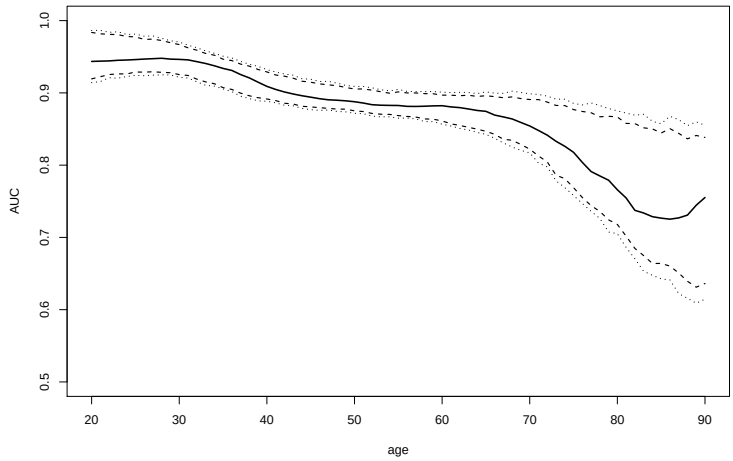


Figure 9: AUC as a function of age (solid line). The dotted and dashed lines represent 90% and 95% pointwise bootstrap confidence intervals, respectively.