REVSTAT – Statistical Journal
Volume 12, Number 1, March 2014, 1–20

ROC CURVE ESTIMATION: AN OVERVIEW

Authors: Luzia Gonçalves
– Unidade de Saúde Pública Internacional e Bioestatística,
Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa,
CEAUL
luziag@ihmt.unl.pt

Ana Subtil
– Departamento de Estatística e Investigação Operacional,
Faculdade de Ciências da Universidade de Lisboa,
CEAUL
asubtil@ihmt.unl.pt

M. Rosário Oliveira
– Departamento de Matemática, Instituto Superior Técnico,
Universidade de Lisboa, Portugal,
CEMAT
rsilva@math.tecnico.ulisboa.pt

Patricia de Zea Bermudez
– Departamento de Estatística e Investigação Operacional,
Faculdade de Ciências da Universidade de Lisboa,
CEAUL
pcbermudez@fc.ul.pt

Abstract:

- This work overviews some developments on the estimation of the Receiver Operating Characteristic (ROC) curve. Estimation methods in this area are constantly being developed, adjusted and extended, and it is thus impossible to cover all topics and areas of application in a single paper. Here, we focus on some frequentist and Bayesian methods which have been mostly employed in the medical setting. Although we emphasize the medical domain, we also describe links with other fields where related developments have been made, and where some modeling concepts are often known under other designations.

Key-Words:

- *Bayesian analysis; bi-normal; kernel; receiver operating characteristic curve; robustness.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

The Receiver Operating Characteristic (ROC) curve was developed by engineers during World War II for detecting enemy objects in battlefields (Collison, 1998). Its expansion to other fields was prompt and, for instance, in psychology it was used to study the perceptual detection of stimuli (Swets, 1996). Over the years, it has been widely applied in many fields including atmospheric sciences, biosciences, experimental psychology, finance, geosciences, and sociology (Marzaban, 2004; Krzanowski and Hand, 2009, and the references therein). ROC analysis has also been increasingly used in machine learning and data mining, and other relevant applications have also emerged in economics (Lasko *et al.*, 2005). Yet in another setting, Morrison *et al.* (2003) described the ROC curve as a simple and effective method to compare the accuracies of reference variables of bacterial beach water quality. Since several fields have contributed independently to the development of ROC analysis, many concepts and techniques are often known under different names in different communities.

This paper provides an overview on some inference methods used in ROC analysis—which have been mostly employed in the medical setting—, and points out the usefulness of transferring knowledge from one field to another. The estimation target of interest is the so-called ROC curve which is a graphical representation of the relationship between false positive and true positive rates or, using an epidemiological language, it is a graphical representation of Se as a function of $1 - \text{Sp}$, where Se is the sensitivity and Sp is the specificity of a diagnostic test. Se is the probability that a truly diseased individual has a positive test result, and Sp is the probability that a truly non-diseased individual has a negative test result. Using the true/false positive/negative rates or the specificity and sensitivity, we deal with conditional probabilities of belonging to a particular predicted class given the true classification (Krzanowski and Hand, 2009), in a two-class classification (*e.g.*, diseased and nondiseased subjects, email messages are spam or not, credit card transactions are fraudulent or not).

In medicine, one of the earliest applications of ROC analysis was published in the 1960s (Lusted, 1960), although the ROC curve only gained its popularity in the 1970s (Martinez *et al.*, 2003; Zhou *et al.*, 2011). Nowadays, medical technologies offer a vast array of ways to diagnose a disease, or to predict the disease progression, and new diagnostic tests and biomarkers are continuously being studied. ROC analysis is widely used for evaluating the discriminatory performance of a continuous variable representing a diagnostic test, a marker, or a classifier.

According to different aims, the ROC analysis is useful to: (i) evaluate the discriminatory ability of a continuous marker to correctly assign into a two-group

classification; (ii) find an optimal cut-off point to least misclassify the two-group subjects; (iii) compare the efficacy of two (or more) diagnostic tests or markers; and (iv) study the inter-observer variability when two or more observers measure the same continuous variable.

Many parametric, semiparametric, and nonparametric estimation methods have been proposed for estimating the ROC curve and its associated summary measures. Here, we focus on some frequentist and Bayesian methods which have been mostly employed in the medical setting. In Section 2 we introduce notation and the basic modeling concepts. Frequentist and Bayesian approaches are reviewed in Section 3 and Section 4, respectively. The paper ends with a short discussion in Section 5.

2. DEFINITIONS AND MODELING FRAMEWORK

Let X and Y be two independent random variables, respectively denoting the diagnostic test measure for a healthy population ($D = 0$) and for a diseased population ($D = 1$), defined using a gold standard. Without loss of generality, and for an appropriate cut-off point c, the test result is positive if it is greater than c and negative otherwise.

Let F and G be the distribution functions of the random variables X and Y, respectively. The sensitivity of the test is given by $\mathrm{Se}(c) = 1 - G(c)$, and the specificity is defined as $\mathrm{Sp}(c) = F(c)$. An example is presented in Figure 1.

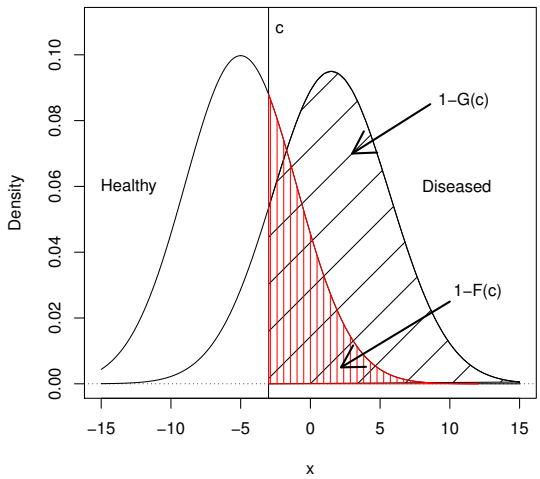

Figure 1: Distribution of the diagnostic test measures for the healthy and the diseased populations.

The ROC curve is defined as a plot of $\mathrm{Se}(c)$ versus $1-\mathrm{Sp}(c)$ for $-\infty \leq c \leq \infty$, or equivalently as a plot of

$$\mathrm{ROC}(t) = 1 - G\big(F^{-1}(1-t)\big)\,, \tag{2.1}$$

over $t \in [0,1]$, where $F^{-1}(1-t) = \inf\big\{x \in \mathbb{R}\colon F(x) \geq 1-t\big\}$.

The ROC curve is increasing and invariant under any monotone increasing transformation of the variables X and Y. Several ROC curve summary measures have been proposed in the literature, such as the area under the curve (AUC) or the Youden index ($\max_c\{\mathrm{Se}(c) + \mathrm{Sp}(c) - 1\}$). They are considered as summaries of the discriminatory accuracy of a test. The AUC is given by

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(u)\,\mathrm{d}u\ . \tag{2.2}$$

Different approaches to estimate the ROC curve lead to different estimates of the AUC. The AUC can be interpreted as the probability that, in a randomly selected pair of nondiseased and diseased individuals, the diagnostic test value is higher for the diseased subject, *i.e.*, $\mathrm{AUC} = P(Y > X)$. Values of AUC close to 1 suggest a high diagnostic accuracy of the test or marker. Bamber (1975) established an important link with the popular nonparametric test of Mann–Whitney. The area of the empirical ROC curve is equal to the Mann–Whitney U statistic that provides an unbiased nonparametric estimator for the AUC (Faraggi and Reiser, 2002). Since the seminal work of Bamber (1975), several authors have proposed refining the nonparametric approach to obtain smoothed ROC curves, for example, by using the kernel method to be described below. Parametric estimation of the ROC curve is also an active area of research and several proposals for F and G are considered. The most widely used parametric ROC model is the bi-normal, which is described in the next section.

3. FREQUENTIST METHODS

3.1. Parametric approaches

3.1.1. The bi-normal estimator

Parametric methods are used when F and G in nondiseased and diseased populations are known. The bi-normal model is commonly considered, and it is applicable when both diseased and nondiseased test outcomes follow normal distributions (Faraggi and Reiser, 2002). If data are actually bi-normal, or a Box–Cox transformation, such as the logarithm or the square root, makes the data

bi-normal, then the relevant parameters can be easily estimated by the means and variances of test values in diseased and nondiseased populations.

Let X and Y be independent normal variables with mean values μ_0, μ_1 and variances σ_0^2, σ_1^2. Then, the ROC curve can be summarized in the following way:

$$\text{ROC}(t) = \Phi\big\{a + b\,\Phi^{-1}(t)\big\}\ , \qquad 0 \le t \le 1\ , \tag{3.1}$$

where, Φ is the standard normal distribution function and a and b are the separation and the symmetry coefficients, respectively, given by $a = (\mu_1 - \mu_0)/\sigma_1$ and $b = \sigma_0/\sigma_1$. In this case, the AUC has a closed form given by

$$\text{AUC} = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right). \tag{3.2}$$

Returning to the example presented in Figure 1, the graphical representation of the ROC curve is illustrated in Figure 2.

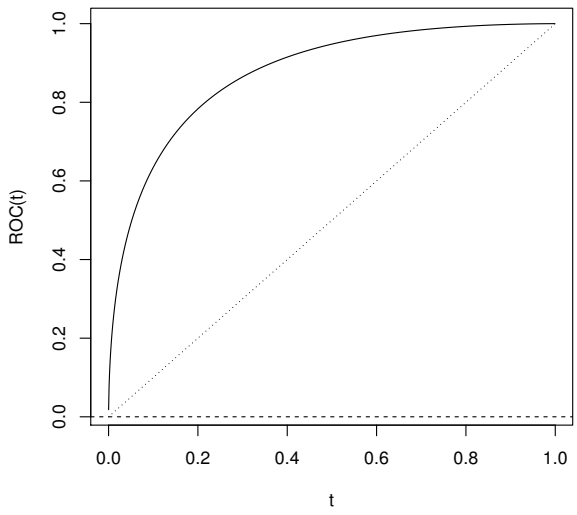

Figure 2: Example of an ROC curve for a bi-normal model, constructed using Equation (3.1).

The bi-normal model leads to convenient maximum likelihood estimates (and corresponding asymptotic variances) of the ROC curve parameters.

In this example, the normal distributions for healthy and diseased populations have the same variance and, hence, the curve is concave. Concavity is a characteristic of *proper* ROC curves (Dorfman *et al.*, 1996). This is a desirable property because it guarantees that the ROC curve will never cross the main

diagonal. Moreover, it is a property of the optimal ROC curve to establish decision rules (Huang and Pepe, 2009). However, a problem with using the bi-normal ROC model is that it is not concave in (0, 1) unless $b = 1$, as noted by Huang and Pepe (2009). Hughes and Bhattacharya (2013) characterize the symmetry properties of bi-normal and bi-gamma ROC curves in terms of the Kullback–Leibler divergences. Considering the negative diagonal of the plot, a ROC curve may be symmetric or skewed towards the left-hand axis or the upper axis of the plot. ROC curves with different symmetry properties may have the same AUC value. Not all continuous parametric ROC curves are proper. It is well known that the bi-normal ROC curve is not proper in general, while the bi-gamma ROC curve is proper (Dorfman *et al.*, 1996; Hughes and Bhattacharya, 2013). Several alternative models have been explored and compared in simulation studies, considering bi-gamma, bi-beta, bi-logistic, bi-exponential (a particular case of bi-gamma), bi-lognormal, bi-Rayleigh and even other proposals, such as the triangular distribution with constrained or unconstrained support (Dorfman *et al.*, 1996; Zou *et al.*, 1997; Marzaban, 2004; Tang *et al.*, 2010; Pundir and Amala, 2012; Tang and Balakrishnan, 2011; Hussain, 2012; Hughes and Bhattacharya, 2013).

3.1.2. Robustness of the bi-normal estimator

The choice of the bi-normal estimator to fit a ROC curve is usually justified by theoretical considerations, mathematical tractability, familiarity with the normal model or just by convenience. Hanley (1988) presents a table summarizing the most common arguments in favor of the use of this estimator. But some authors also argue that the bi-normal estimator is robust. The word robust can have many different meanings. Here it is used in the sense of robust statistics, *i.e.* meaning that in the presence of a certain amount of observations coming from a non-normal distribution the bi-normal estimator will yield reliable results. Lately, the impact of model misspecification in the parametric or semiparametric models used in health sciences is gaining importance, since practitioners are aware that theoretical models are only approximations of reality, and statistical procedures that give reliable results under model departures are essential for solving real problems. This concern is addressed by Heritier *et al.* (2009) and Farcomeni and Ventura (2010).

In the case of the bi-normal estimator of the ROC curve, authors like Swets (1986) argue that "*Empirical ROC's drawn from experimental psychology and several practical fields, (...) are fitted well on a binormal graph...*". This statement is reinforced by Hanley (1988), who claims that "*...the binormal-based fits are certainly good enough for all practical purposes.*". Hajian-Tilaki *et al.* (1997) state that, "*The results suggested that the AUC is robust to departures from binomality if one uses the binormal model as implemented in* `LARROC` *program.*". Neverthe-

less, these authors were more cautious adding that a possible explanation relies in the use of ranks instead of the original data, in both estimation procedures.

Walsh (1997) clarifies these arguments. Robustness, in Swets (1986) and Hanley (1988), is understood as the ability of the bi-normal estimator to fit a ROC curve that 'looks right' in comparison either with the theoretical ROC curve or with the observed rating method. But this author goes further, discussing the ability of the bi-normal estimator to produce valid inferences in circumstances in which the data does not satisfy the normality assumption. A simulation study to analyze the impact of data coming from a bi-logistic model combined with bi-normal estimator was developed to study: (i) the AUC estimator, (ii) the performance of the statistical test to compare AUC from two ROC curves, and (iii) the impact on size and power of this statistical test. The choice of the bi-logistic distributions to model departures from bi-normal assumption relies on the difficulty to distinguish these models, since the logistic model was considered one of the possible hardest scenarios to detect departures from the normality assumption. In his simulation study, Walsh also considers the effect of different sets of decision thresholds, and concludes that the bi-normal estimator is sensitive to model misspecification and to the location of the decision thresholds.

The problem of robustness has deserved the attention of other authors. Greco and Ventura (2011) develop an M-estimator for the $P(Y > X)$ in the context of a stress-strength model, that has direct application in AUC estimation. Recently, Devlin *et al.* (2013) discuss the impact of model misspecification in three estimators resulting from modeling the parametric form of the ROC curve directly.

3.2. Nonparametric estimation of the ROC curve

3.2.1. Empirical estimator and variants

The simplest nonparametric method is the empirical estimator, which is based plugging in empirical estimates into (2.1). Specifically, the empirical estimate of the ROC curve is given by

$$\widetilde{\mathrm{ROC}}(t) = 1 - \widetilde{G}\big(\widetilde{F}^{-1}(1-t)\big)\,, \tag{3.3}$$

where $\widetilde{F}^{-1}$ and $\widetilde{G}$ respectively denote the empirical quantile function and the empirical distribution function associated to healthy and diseased populations; roughly speaking, the empirical distribution function is defined, for any given value t, as the percentage of sample points smaller or equal to t.

The empirical ROC curve preserves many properties of the empirical distribution function and it is uniformly convergent to the theoretical curve (Hsieh and Turnbull, 1996). Nevertheless, the estimator has some drawbacks, and it may suffer from large variability, particularly for small sample sizes (Lloyd, 1998; Lloyd and Yong, 1999; Jokiel-Rokita and Pulit, 2013). While this is not a major problem in machine learning, data mining, and finance—where large samples are common—in medicine this may be inadequate, as small samples are commonplace in clinical practice. In addition to all this, the estimated ROC curve is not continuous, and thus its interpretation becomes more complex (Jokiel-Rokita and Pulit, 2013).

Other methods have been explored to obtain smooth ROC curve estimates, either through kernel smoothing (Lloyd, 1998; Lloyd and Yong, 1999) or through smooth versions of the empirical distribution function (Jokiel-Rokita and Pulit, 2013).

3.2.2. Kernel estimator

To overcome the lack of smoothness of the empirical estimator, Zou *et al.* (1997) used kernel methods to estimate the ROC curve, which were later improved by Lloyd (1998). Kernel density estimators are known to be simple, versatile, with good theoretical and practical properties (Silverman, 1986; Tenreiro, 2010), merits that the corresponding ROC curve estimator inherit.

Let $(x_1, ..., x_n)$ and $(y_1, ..., y_m)$ be two independent samples from X and Y, respectively. The kernel density estimators of f and g, the probability density functions associated with F and G, are:

$$\widehat{f}(x) = \frac{1}{nh_0} \sum_{i=1}^{n} K_0\left(\frac{x - x_i}{h_0}\right), \qquad \widehat{g}(y) = \frac{1}{mh_1} \sum_{i=1}^{m} K_1\left(\frac{y - y_i}{h_1}\right).$$

Here the $h_i > 0$ are bandwidths, which are used to control the amount of smoothness, and the K_i are kernel functions, that obey (i) $\int_{\mathbb{R}} K_i(x)\,\mathrm{d}x = 1$, (ii) $\int_{\mathbb{R}} x\,K_i(x)\,\mathrm{d}x = 0$, and (iii) $\int_{\mathbb{R}} x^2 K_i(x)\,\mathrm{d}x > 0$, for $i = 0, 1$. Using these estimators, the cumulative distribution functions can be estimated as

(3.4)

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{x} \frac{1}{h_0} K_0\left(\frac{u - x_i}{h_0}\right) \mathrm{d}u\,, \quad \widehat{G}(y) = \frac{1}{m} \sum_{i=1}^{m} \int_{-\infty}^{y} \frac{1}{h_1} K_1\left(\frac{v - y_i}{h_1}\right) \mathrm{d}v\,.$$

These integrals can be evaluated numerically. The choice of the kernels K_0 and K_1 among the available proposals is not problematic, since they all give comparable results, as was pointed out by Krzanowski and Hand (2009) and Jokiel-Rokita and Pulit (2013). This justifies the pragmatic option of using equal kernels, and

a popular option is the Gaussian kernel (Sheather, 2004; Hong *et al.*, 2007; Zhou *et al.*, 2011; Fabsic, 2012), and in this case Equation (3.4) can be written as

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{x-x_i}{h_0}\right), \qquad \widehat{G}(x) = \frac{1}{m}\sum_{i=1}^{m}\Phi\left(\frac{y-y_i}{h_1}\right). \tag{3.5}$$

Plugging-in (3.4) into (2.1) leads to the kernel-based ROC curve estimator:

$$\widehat{\mathrm{ROC}}(t) = 1 - \widehat{G}\big(\widehat{F}^{-1}(1-t)\big). \tag{3.6}$$

The most sensitive aspect of the kernel-based ROC curve estimator in (3.6) is the choice of the 'optimal' bandwidth (Zhou and Harezlak, 2002; Hall and Hyndmann, 2003; Zhou *et al.*, 2011; Jokiel-Rokita and Pulit, 2013). This, combined with the selection of K determines the properties of the estimator. Zou *et al.* (1997) used bandwidths that are asymptotically optimal for estimating f and g. Lloyd (1998) improved the previous proposal by choosing bandwidths that are asymptotically optimal for estimating F and G, since the ROC curve depends directly on these cumulative distribution functions. Lloyd and Yong (1999) showed how kernel density estimators overcome the empirical ones. Qiu and Le (2001) proposed a ROC curve estimator based on a kernel distribution function estimator to G and a local smoothing quantile function estimator to F^{-1}. Peng and Zhou (2004) introduced another kernel estimator involving only one bandwidth, estimated in an optimal asymptotical way, that has better performance near the boundary of the support of X and Y. Koláček and Karunamuni (2009) proposed a related kernel-based estimator for the ROC curve that removes the boundary effects. Contrasting with these approaches, Jokiel-Rokita and Pulit (2013) proposed a strongly consistent estimator based on a smoothed version of the empirical ROC curve that, according to a simulation study, outperformed the empirical and a kernel estimator for small sample sizes.

Kernel-based estimators can also be used for estimating the AUC. For example, using the estimators proposed by Lloyd (1998) and a Gaussian kernel, yields the following estimator

$$\widehat{\mathrm{AUC}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\Phi\left(\frac{y_j - x_i}{\sqrt{h_0^2 + h_1^2}}\right). \tag{3.7}$$

See Fabsic (2012), for a simulation study comparing several parametric and non-parametric methods.

4. BAYESIAN METHODS

4.1. Introduction

Bayesian methods are introduced in ROC curve estimation as an alternative to maximum likelihood methods. Bayesian approaches enable the introduction of prior information into the estimation process, which reduces the uncertainty of the inferences. This point is specially important when a gold standard test, which correctly classifies all subjects as healthy or diseased, is unavailable, either because there is no gold standard for the disease or because the procedure is costly, technically demanding, harmful or even life-threatening. In this framework, the true state of the individuals is unknown and the modeling process may benefit from including existing information about the problem under study through the use of prior distributions.

The Bayesian framework enables obtaining credibility intervals for the ROC curve and for other summary measures, such as the AUC. As it is known, one of the benefits of the Bayesian methodology is the capability of producing regions in terms of the posterior distributions of the parameters. These regions, contrarily to confidence intervals resulting from frequentist analysis, allow for probabilistic interpretations of the inferences. Additionally, predictive probabilities of the health status of future individuals can be obtained through the predictive distribution. Furthermore, the Bayesian perspective is specially suited to model complex designs, namely through the use of hierarchical structures (Ishwaran and Gatsonis, 2000; O'Malley and Zou, 2006; Johnson and Johnson, 2006).

It is well known that the ability of a diagnostic test to discriminate between diseased and healthy populations, may be influenced by various factors (Pepe, 2003). Moreover, assessing the covariate impact may provide useful information regarding the test adequacy towards different populations and conditions (de Carvalho *et al.*, 2013). On the contrary, neglecting covariate effects may lead to biased inferences about the test performance. Covariate effects on the ROC curves are addressed in several works (*e.g.* Peng and Hall, 1996; Branscum *et al.*, 2008; de Carvalho *et al.*, 2013).

Traditionally, in a Bayesian framework, ROC curve estimation has been explored in a parametric manner. More recently, semiparametric and nonparametric methodologies have also been developed. In the next subsections some of these approaches will be described.

4.2. Parametric approaches

Some of the first accounts of using a Bayesian methodology in ROC curve estimation are based on regression models (Peng and Hall, 1996; Hellmich *et al.*, 1998). Probit-linked generalized linear regression models are applied to ordinal test results, leading to Bayesian inferences for ROC curves and functionals such as the AUC. In particular, the approach adopted by Peng and Hall (1996) admits latent bi-normal distributions for diseased and nondiseased populations, even though other parametric distributions could be considered. The authors use data augmentation techniques to impute unobserved continuous data from the latent distribution, thus allowing to overcome the difficulties due to the ordinal nature of the observations. Noninformative priors are applied. This ordinal regression model can explain modifications observed in the ROC curves caused by changing the value of a single covariate.

As mentioned earlier, some regression approaches to ROC curve analysis consider hierarchical structures (O'Malley and Zou, 2006; Johnson and Johnson, 2006). A Bayesian multivariate hierarchical transformation model is developed by O'Malley and Zou (2006) based on clustered continuous diagnostic test data with covariates. This approach is useful in the context of multilevel data with clustered responses, like, for example, radiologic data collected from patients (individual level) nested in different hospitals (clusters). The authors aim to model the diagnostic test accuracy and define a composite diagnostic test. The authors remark that a cluster-specific transformation of the outcomes is applied to handle the heterogeneity between the clusters and that multiple correlated outcomes may be used. The methodology is applied to prostate cancer biopsy data gathered from a multi-center clinical trial.

Johnson and Johnson (2006) address a situation frequently observed in radiology, in which several radiologists rate, in an ordinal scale, multiple exams collected from the same individual. A Bayesian hierarchical latent variable model for analyzing multirater correlated ordinal data is proposed. The three sources of variation (differences in patients characteristics, in diagnostic exams and in raters) are explicitly modeled, each one corresponding to a different level of the model hierarchy. Simulation studies show that this model is more efficient than the most widely used model for multirater correlated data analysis (Dorfman *et al.*, 1992).

4.3. Semiparametric and nonparametric approaches

Bayesian semiparametric and nonparametric approaches have been used for ROC curve estimation in the last few years (Erkanli *et al.*, 2006; Wang *et al.*, 2007; Gu *et al.*, 2008; Branscum *et al.*, 2013). These methodologies are still being developed and constitute a very active line of research.

Nonparametric Bayesian methods are meant to overcome the restrictions imposed by considering a fixed parametric model and the consequent difficulties in capturing nonstandard data features, such as multimodality and skewness. Contrarily to the traditional parametric framework, the nonparametric framework enables a more flexible modeling of the data, in the sense that no specific parametric family of distributions is considered.

The nonparametric approach entails a modeling framework that requires specifying a prior distribution over the space of all probability measures. As pointed out by Inácio (2012), this does not mean an absence of parameters in the model, on the contrary it involves an (possibly) infinite number of parameters. In this framework, Dirichlet processes, mixtures of Dirichlet processes, Polya trees, and mixtures of Polya trees are frequently used priors; for further details on this see Inácio (2012), and references therein.

A Bayesian semiparametric approach for ROC curve estimation method, based on mixtures of Dirichlet processes, was developed by Erkanli *et al.* (2006). A Gibbs sampling framework is used to obtain posterior distributions of the mixtures of Dirichlet processes model, thus providing posterior predictive estimates of sensitivity, specificity, ROC curves and AUC. The authors show that, even when a gold standard diagnostic test is not available, the results still stand. Moreover, it closely parallels the kernel density estimation approach, previously referred to in this paper.

A nonparametric Bayesian method reported by Hanson *et al.* (2008) uses Dirichlet process mixtures and mixtures of Polya trees for analyzing continuous serologic data. A novelty of this approach is the inclusion of a stochastic ordering constraint for the serologic values distributions of the infected and noninfected populations. This is a biologically reasonable assumption, since the serologic scores tend to be higher for the infected individuals than for the noninfected ones. According to the authors, the approach has the benefit of guaranteeing that the AUC is always larger than 0.5, meaning that the ROC curve never goes below the main diagonal. The two models are applied to Johne's disease data observed in dairy cattle. Qualitatively similar inferences are obtained and the same conclusions, regarding the accuracy of the serologic tests, can be drawn from both applications.

In the Bayesian nonparametric context, few works study the effect of covariates in ROC curve estimation. This issue is explored by de Carvalho *et al.* (2013). The model is based on dependent Dirichlet processes and allows the entire distribution in each group to smoothly change as a function of the covariates. This approach can accommodate multiple continuous and categorical predictors. An approximated version of the general model, based on B-splines, was compared with the semiparametric approach of Pepe (1998), with an extension of the previous approach that uses a B-splines trend and with the nonparametric kernel estimator of Rodríguez-Álvarez *et al.* (2011). The proposed model outperforms its competitors for nonlinear scenarios and small sample sizes. An application of the model to diabetes diagnosis is presented.

As explained by Inácio (2012), ROC surfaces have been proposed for the evaluation of the diagnostic accuracy in ordered three-class problems as a direct generalization of the ROC curve. A flexible Bayesian nonparametric approach based on mixtures of finite Polya trees priors is described by Inácio (2012).

The bootstrap has been used to ROC curve estimation by Gu *et al.* (2008). The authors also present estimation credible intervals of the ROC curve and apply the approach for testing the validity of the bi-normal assumption.

4.4. Absence of a gold standard

Imperfect diagnostic tests are widely used in medicine and, as we pointed out earlier, the Bayesian methodology is particularly suited for problems of this nature (Krzanowski and Hand, 2009).

Returning to the previously mentioned work of Erkanli *et al.* (2006), an extension of the nonparametric model to the case of imperfect reference test is given, in which a binary latent variable is introduced to express the true but unknown disease status. Extensive literature exists on the use of latent class models to evaluate the performance of binary diagnostic tests in the absence of a gold standard, either using maximum likelihood or Bayesian estimation methods (see Gonçalves *et al.*, 2012, and references therein).

Again, in the context of no gold standard data analysis, Choi *et al.* (2006) develop a parametric Bayesian methodology that admits two diagnostic tests applied to the same individuals. The data are modeled under the bi-normal assumption; this assumption may require a suitable transformation, which can be difficult to find in the absence of a gold standard test. Training data or previous studies with a gold standard could suggest an adequate transformation. The method is initially formulated for the gold standard case and slightly modified to address the gold standard absence. A latent variable indicating the true disease

state is introduced, resembling Erkanli *et al.* (2006). The method has difficulty in assigning the correct disease status when the overlap of diseased and nondiseased groups is too large.

Wang *et al.* (2007) explore the problem of estimating the ROC curve of a new ordinal or continuous scale diagnostic test by comparison with an imperfect binary reference test, assuming conditional independence between the two tests. Identifiability problems require data from at least two populations with different prevalences. The method is based on a multinomial model and no assumptions are needed concerning the shape of the distributions corresponding to the test values. Care is taken in guaranteeing the monotonicity of the ROC curve.

Both Choi *et al.* (2006) and Wang *et al.* (2007) illustrate their methods using different datasets from Johne's disease in cattle.

A group of Bayesian latent class models for mixed continuous and discrete diagnostic test data is explored by Weichenthal *et al.* (2010). These models are used to determine the probability of asbestos exposure from lung fiber count data. The model admits correlations between repeated measurements of the same test within individuals.

Branscum *et al.* (2008) propose Bayesian nonparametric and semiparametric approaches to ROC analysis and disease diagnosis in the absence of a gold standard. A nonparametric model using mixtures of Polya trees is proposed to estimate probabilities of disease risk and the ROC curve. Semiparametric extensions of this model are also proposed. These semiparametric models incorporate additional information regarding the disease status. Two types of information are used: standard covariate information and information from additional binary diagnostic tests. Such additional information improves the discriminatory ability to correctly classify subjects as healthy or diseased, leading to a modeling process in between the gold standard case and the nonparametric modeling in the absence of a gold standard. This is a very flexible approach that allows combining in a single framework available information on risk factors and additional diagnostic tests outcomes to enhance diagnostic predictive accuracy.

Nonparametric Bayesian analysis involving Polya tree priors is also dealt with in Branscum *et al.* (2013). The usefulness of the discussed flexible models over a standard parametric method is shown in an application to a lung cancer biomarker.

5. FINAL REMARKS

Statistical modeling of ROC curves is a vast topic and offers several future research lines. The use of flexible models that accommodate covariates and prior information is an active field of research. If proper ROC curves are desired in many applications, in Bioinformatics not proper ROC curves have been increasingly used as new tools for the analysis of differentially expressed genes in microarray experiments (*e.g.* Parodi *et al.*, 2008; Silva-Fortes *et al.*, 2012). A particularly relevant issue in this setting is robustness, but further research is definitely required on this.

ACKNOWLEDGMENTS

We would like to express our gratitude to the Editors for the invitation to present this paper. This work has been supported by *Fundação para a Ciência e Tecnologia* (FCT), Portugal, through PTDC/MAT/118335/2010, PEst-OE/MAT/UI0006/2014 and PEst-OE/MAT/UI0822/2014. A. Subtil acknowledges financial support from the FCT grant SFRH/BD/69793/2010.

REFERENCES

Bamber, P. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *Journal of Mathematical Psychology*, **12**, 387–415.

Branscum, A. J.; Johnson, W. O. and Baron, A.T. (2013). Robust medical test evaluation using flexible bayesian semiparametric regression models, *Epidemiology Research International*, ID 131232.

Branscum, A. J.; Johnson, W. O.; Hanson, T. E. and Gardner, I. A. (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis, *Statistics in Medicine*, **27**, 2474–2496.

Choi, Y.-K.; Johnson, W. O.; Collins, M. T. and Gardner, I. (2006). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard, *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 210–229.

Collison, P. (1998). Of bombers, radiologists, and cardiologists: time to ROC, *Heart*, **80**, 215–217.

de Carvalho, V. I.; Jara, A.; Hanson, T. E. and de Carvalho, M. (2013). Bayesian nonparametric ROC regression modeling, *Bayesian Analysis*, **8**, 623–646.

Devlin, S. A.; Thomas, E. G. and Emerson, S. S. (2013). Robustness of approaches to ROC curve modeling under misspecification of the underlying probability model, *Communications in Statistics—Theory and Methods*, **42**, 3655–3664.

Dorfman, D. D.; Berbaum, K. S. and Metz, C. E. (1992). Receive operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method, *Investigative Radiology*, **27**, 723–731.

Dorfman, D. D.; Berbaum, K. S.; Metz, C. E.; Lenth, R. V.; Hanley, J. A. and Dagga, H. A. (1996). Proper receiver operating characteristic analysis: the bigamma model, *Academic Radiology*, **4**, 138–149.

Erkanli, A.; Sung, M.; Costello, J. E. and Angold, A. (2006). Bayesian semi-parametric ROC analysis, *Statistics in Medicine*, **25**, 3905–3928.

Fabsic, P. (2012). *Comparing the Accuracy of ROC Curve Estimation Methods*, Master Thesis, Swiss Federal Institute of Technology, Zurich.

Farcomeni, A. and Ventura, L. (2010). An overview of robust methods in medical research, *Statistical Methods in Medical Research*, **21**, 111–133.

Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve, *Statistics in Medicine*, **21**, 3093–3096.

Gonçalves, L.; Subtil, A.; Oliveira, M. R.; Rosário, V.; Lee, P. and Shaio, M.-F. (2012). Bayesian latent class models in malaria diagnosis, *PLoS ONE*, **7**, e40630.

Greco, L. and Ventura, L. (2011). Robust inference for the stress-strength reliability, *Statistical Papers*, **52**, 773–788.

Gu, J.; Ghosal, S. and Roy, A. (2008). Bayesian bootstrap estimation of ROC curve, *Statistics in Medicine*, **27**, 5407–5420.

Hajian-Tilaki, K. O.; Hanley, J. A.; Joseph, L. and Collet, J. P. (1997). A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests, *Medical Decision Making*, **17**, 94–102.

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

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Machine Learning*, **77**, 103–123.

Hanley, J. A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves, *Medical Decision Making*, **8**, 197–203.

Hanson, T. E.; Kottas, A. and Branscum, A. J. (2008). Modelling stochastic order in the analysis of Receiver Operating Characteristic data: Bayesian nonparametric approaches, *Journal of the Royal Statistical Society. Ser.* C, **57**, 207–225.

Hellmich, M.; Abrams, K. R., Jones, D. R. and Lambert, P. C. (1998). A Bayesian approach to a general regression model for ROC curves, *Medical Decision Making*, **18**, 436–443.

Heritier, S.; Cantoni, E.; Copt, S. and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*, John Wiley & Sons, Chichester.

Hong, X.; Chen, S. and Harris, C. J. (2007). A kernel-based two-class classifier for imbalanced data sets, *IEEE Transactions on Neural Networks*, **18**, 28–41.

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

Huang, Y. and Pepe, M. S. (2009). A parametric ROC model based approach for evaluating the predictiveness of continuous markers in case-control studies, *Biometrics*, **65**, 1133–1144.

Hughes, G. and Bhattacharya, B. (2013). Symmetry properties of binormal and bi-gamma receiver operating characteristic curves are described by Kullback–Leibler divergences, *Entropy*, **15**, 1342–1356.

Hussain, E. (2012). The bi-gamma ROC curve in a straightforward manner, *Journal of Basic & Applied Sciences*, **8**, 309-314.

Inácio, V. (2012). *Semiparametric and Nonparametric Modeling of Diagnostic Data*, PhD Thesis, University of Lisbon, Portugal.

Johnson, T. D. and Johnson, V. E. (2006). A Bayesian hierarchical approach to multirater correlated ROC analysis, *Statistics in Medicine*, **25**, 1858–1871.

Jokiel-Rokita, A. and Pulit, M. (2013). Nonparametric estimation of the ROC curve based on smoothed empirical distribution functions, *Statistics and Computing*, **23**, 703–712.

Ishwaran, H. and Gatsonis, A. C. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis, *The Canadian Journal of Statistics*, **28**, 731–750.

Koláček, J. and Karunamuni, R. J. (2009). On boundary correction in kernel estimation of ROC curves, *Austrian Journal of Statistics*, **38**, 17–32.

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

Lasko, T. A.; Bhagwat, J. G.; Zou, K. H. and Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics, *Journal of Biomedical Informatics*, **38**, 404–415.

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.

Lusted, L. B. (1960). Logical analysis in roentgen diagnosis, *Radiology*, **74**, 178–193.

Martinez, E. Z.; Louzada-Neto, F. and Pereira, B. B. (2003). A curva ROC para testes diagnósticos (Document in Portuguese), *Cadernos Saúde Coletiva*, **11**, 7–31.

Marzaban, C. (2004). The ROC curve and the area under it as performance measures, *Weather and Forecasting*, **19**, 1106–1114.

Morrison, A. M.; Coughlin, K.; Shine, J. P.; Coull, B. A. and Rex A. C. (2003). Receiver operating characteristic curve analysis of beach water quality indicator variables, *Applied and Enviromental Microbiology*, **69**, 6405–6411.

O'Malley, A. J. and Zou, K. H. (2006). Bayesian multivariate hierarchical transformation models for ROC analysis, *Statistics in Medicine*, **25**, 459–479.

Parodi, S.; Pistoia, V. and Muselli, M. (2008). Not proper ROC curves as new tool for the analysis of differentially expressed genes in microarray experiments, *BMC Bioinformatics*, **8**, 410.

Peng, F. and Hall, W. J. (1996). Bayesian analysis of ROC curves using Markov-chain Monte Carlo methods, *Medical Decision Making*, **16**, 404–411.

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

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

Pundir, S. and Amala, R. (2012). A study on the bi-Rayleigh ROC curve model, *Bonfring International Journal of Data Mining*, **2**, 42–47.

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

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

Sheather, S. J. (2004). Density estimation, *Statistical Science*, **19**, 588–597.

Silva-Fortes, C.; Amaral Turkman, M. A. and Sousa L. (2012). Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups, *BMC Bioinformatics*, **13**, 147.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance, *Psycological Bulletin*, **99**, 181–198.

Swets, J. A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*, Lawrence Erlbaum Associates, New Jersey.

Tang, L.; Du, P. and Wu, C. (2010). Compare diagnostic tests using transformation-invariant smoothed ROC curves, *Journal of Statistical Planning and Inference*, **140**, 3540–3551.

Tang, L. L. and Balakrishnan, N. (2011). A random-sum Wilcoxon statistic and its application to analysis of ROC and LROC data, *Journal of Statistical Planning and Inference*, **141**, 335–344.

Tenreiro, C. (2010). Uma introdução à estimação não-paramétrica da densidade (Document in Portuguese), *Edições SPE*, Lisboa.

Walsh, S. J. (1997). Limitations to the robustness of binormal ROC curves: effects of model misspecification and location of decision thresholds on bias, precision, size and power, *Statistics in Medicine*, **16**, 669–679.

Wang, C.; Turnbull, B. W.; Gröhn, Y. T. and Nielsen, S. S. (2007). Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown, *Journal of Agricultural, Biological, and Environmental Statistics*, **12**, 128–146.

Weichenthal, S.; Joseph, L.; B é lisle, P. and Dufresne, A. (2010). Bayesian estimation of the probability of asbestos exposure from lung fiber counts, *Biometrics*, **66**, 603–612.

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

Zhou, X-H.; Obuchowski, N. A. and McClisk, D. K. (2011). *Statistical Methods in Diagnostic Medicine*, Second Edition, John Wiley & Sons, New Jersey.

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.