International Journal of Statistics and Probability; Vol. 6, No. 1; January 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Estimating the Area under the ROC Curve with Modified Profile Likelihoods

Giuliana Cortese[1]

[1] Department of Statistical Sciences, University of Padua, Italy

Correspondence: Giuliana Cortese, Department of Statistical Sciences, University of Padua, Via Cesare Battisti 241, 35121 Padua, Italy. Tel: 39-49-827-4159. E-mail: gcortese@stat.unipd.it

Received: October 11, 2016 Accepted: November 11, 2016 Online Published: November 18, 2016

doi:10.5539/ijsp.v6n1p1 URL: https://doi.org/10.5539/ijsp.v6n1p1

Abstract

Receiver operating characteristic (ROC) curves are a frequent tool to study the discriminating ability of a certain characteristic. The area under the ROC curve (AUC) is a widely used measure of statistical accuracy of continuous markers for diagnostic tests, and has the advantage of providing a single summary index of overall performance of the test. Recent studies have shown some critical issues related to traditional point and interval estimates for the AUC, especially for small samples, more complex models, unbalanced samples or values near the boundary of the parameter space, i.e., when the AUC approaches the values 0.5 or 1. Parametric models for the AUC have shown to be powerful when the underlying distributional assumptions are not misspecified. However, in the above circumstances parametric inference may be not accurate, sometimes yielding misleading conclusions. The objective of the paper is to propose an alternative inferential approach based on modified profile likelihoods, which provides more accurate statistical results in any parametric settings, including the above circumstances. The proposed method is illustrated for the binormal model, but can potentially be used in any other complex model and for any other parametric distribution. We report simulation studies to show the improved performance of the proposed approach, when compared to classical first-order likelihood theory. An application to real-life data in a small sample setting is also discussed, to provide practical guidelines.

Keywords: area under the ROC curve, binormal model, continuous diagnostic marker, modified profile likelihood, ROC curve, stress-strength model.

1. Introduction

Receiver operating characteristic (ROC) curves are frequently used to study the ability of a certain characteristic in discriminating and classifying units under study. One of the most popular summary measures based on the ROC curve is the area under the curve (AUC) (Krzanowski & Hand, 2009), which was originally developed in radar signal detection (Bamber, 1975), and later it has been used in a broad range of applied contexts such as radiology, psychiatry, reliability theory and industrial inspection systems, earthquake resistance.

The AUC is also widely applied in medicine as a measure of statistical accuracy of continuous markers for diagnostic tests (Faraggi & Reiser, 2002; Pepe, 2003; Zhou, McClish, & Obuchowski, 2009). A diagnostic test based on a continuous marker provides usually a response about the possible clinical status of subjects, identifying them as diseased (test positive) or non-diseased (test negative) patients. Such test requires that a certain cut-off point t is chosen. The probabilities that the test correctly classifies subjects as diseased and non-diseased, are called, respectively, the sensitivity and specificity of the test associated with t.

To formalize the problem more generally, denote with $\bar{D}$ and D, respectively, the true negative and true positive status of units in the population of interest (e.g., real condition of being non-diseased or diseased). Let us define two continuous random variables Y and X that describe a continuous characteristic of interest in the two distinct groups $\bar{D}$ and D, respectively. Let $F_Y(\cdot)$ and $F_X(\cdot)$ be the corresponding cumulative distribution functions, and $f_Y(t)$ and $f_X(t)$ the associated probability density functions. Consider a classification rule based on a certain cut-off point t (e.g., a diagnostic test that classifies subjects as 'non-diseased' if the observed value of the characteristic is below t, and as 'diseased' if the observed value is above t). The probability that a unit with true status $\bar{D}$ is correctly classified by the diagnostic test (test negative) is called 'specificity' and defined as $p(t) = F_Y(t)$, while the probability that a unit with true status D is correctly classified by the test (test positive) is called 'sensitivity' and defined as $q(t) = 1 - F_X(t)$. Sensitivity and specificity vary when different choices of t are made over the continuous scale of the characteristic. The ROC curve is then obtained by plotting $p(t)$ versus $1 - q(t)$ for all possible values of t.

The AUC has the advantage of providing a single index that summarizes the overall performance of the test (or rule) based on the continuous characteristic, rather than an entire curve, and it is particularly useful for comparisons under different

populations or different tests. The aim is often to minimize the error $1 - q(t)$ committed by the test, and simultaneously increase the efficacy in discovering units from the D population. Therefore, values of the AUC close to 1 indicate very high accuracy of the test, while very low accuracy corresponds to values closer to 0.5. Bamber (1975) showed that the AUC based on continuous distributions is a probabilistic measure that is equal to

$$A = P(Y \leq X) = \int_{-\infty}^{\infty} F_Y(t)\, dF_X(t) = \int_0^1 q \circ \bar{p}^{-1}(z)\, dz, \tag{1}$$

where $\bar{p}(t) = 1 - p(t)$. The quantity A can also be interpreted as the probability that, in a randomly selected pair of $\bar{D}$ and D subjects, the test value is higher for a subject from the D population. In more general contexts, the AUC is used as a measure of difference between distributions (Wolfe & Hogg, 1971). It is often used in engineering and reliability theory with the name of stress-strength model (Johnson, 1988; Kotz, Lumelskii, & Pensky, 2003). When X represents the strength of a certain component and Y is the applied stress, then A measures the probability that a component would not fail if it is put under a systematic stress.

Inference for the AUC has been studied under different modeling assumptions, following mainly a nonparametric, a parametric or a Bayesian approach. In practical applications, it has been suggested that all these approaches are useful and the comparison of their results may provide additional information on the consistency among them. Moreover, the AUC has been also investigated under various relevant settings, such as presence of explanatory variables, measurement errors and clustered data (Pardo-Fernández, Rodríguez-Álvarez, Van Keilegom, et al., 2013; Reiser, 2000; Zou, Carlsson, & Yu, 2012).

Recently, a special attention has been devoted to interval estimation of A and some related critical issues have been widely discussed in the literature (Feng, Cortese, & Baumgartner, 2015). Some of these issues concern a bad performance of confidence intervals for the AUC especially for small samples, more complex models, unbalanced samples or values near the boundary of the parameter space (i.e, A approaching 0.5 or 1). In particular, classical parametric approaches have the general problem that the smaller the sample size and the higher the number of parameters, less accurate they are in the interval and point estimation. On the other hand, nonparametric methods tend also to perform poorly when the sample size is small. Moreover, in general the parametric methods seem to outperform the nonparametric ones when the underlying distributional assumptions are not misspecified, and in presence of samples that show a nearly perfect separation between subjects in the two groups $\bar{D}$ and D (Obuchowski & Lieber, 2002).

In the current papers we restrict our attention to the parametric framework for inference on the AUC. For the binormal model, where Y and X are assumed to follow normal distribution with different means and variances, Reiser and Guttman (1986) and Reiser and Faraggi (1997) proposed a method for the construction of confidence intervals based on a standard approximate t of Student solution. Although their procedure appears to work well also for unbalanced or small samples, it is not extendible to different parametric models for Y and X, such as Weibull, Gamma or any other more general parametric distributions not in the location-scale family, or to e.g. mixture model in presence of bimodal distributions. Moreover, it is not clear how to handle presence of explanatory variables or clustered data. Classical asymptotic methods based on parametric likelihood theory can easily be applied for constructing confidence intervals or test of hypothesis for the AUC for any type of assumed parametric model. However, it is well known from the general likelihood theory that the resulting Wald type statistic and likelihood ratio statistic do not show a good performance in all situations, especially in the coverage probability of 95% confidence intervals (Severini, 2000). A recent parametric approach was based on higher-order asymptotic likelihood theory (Cortese & Ventura, 2013). However, it has been shown that such method has some limitations: it may easily fail in presence of very small or unbalanced samples or when the samples produce a nearly perfect observed discrimination, it is computationally unstable near the maximum likelihood estimate of A. Some of these problems have been underlined in Feng et al. (2015).

To overcome these drawbacks, the current paper addresses the problem of inaccurate parametric inference in case of small or unbalanced sample sizes, with special attention to confidence intervals and test of hypothesis for the AUC. Also the problem of correct inference near the limit values 0.5 and 1, which represent the situations of, respectively, lowest and maximal accuracy of the continuous characteristic under study, is investigated. In regard of these objectives, we present inference for the AUC based on a modified version of the profile likelihood function, denoted in the literature as 'modified profile likelihood' (Cox & Reid, 1992). In this setting, the parameter identifying the AUC is treated as parameter of interest, whereas the remaining parameters related to the underling parametric distributions of Y and X are treated as nuisance parameter. The proposed approach is very general, applicable to any type of parametric distribution assumptions and to any data setting, such as clustered data or additional data on explanatory variables (Sartori, 2003).

It has been widely studied that standard likelihood inference for a parameter of interest could be misleading in presence of relatively many nuisance parameters, with respect to the sample size, or for small samples. The classical approach for

making inference on a parameter of interest in presence of nuisance parameters is based on profile likelihoods. The profile likelihood function is the likelihood in which the nuisance parameters are maximized out, for every fixed value of the parameter of interest. This likelihood is not a proper likelihood and therefore, the derived score function is biased (Severini, 2000). Consequently, this bias may increase with the dimension of the nuisance parameter and produce inaccurate estimation. The modified profile likelihoods are an interesting alternative to the profile likelihoods, since they correct for the presence of nuisance parameters (Cox & Barndorff-Nielsen, 1994; Cox & Reid, 1992) showing an improved performance.

The scope of the paper is to investigate the performance of modified profile likelihoods for inference on the AUC based on a general parametric model. The inferential procedure is presented in the general setting. Then, the methodological aspects are illustrated for the binormal model. In order to show how to obtain point estimates, confidence intervals and test of hypothesis based on the modified profile likelihood, we consider an application to real data in a setting of small samples.

The paper is organized as follows. Section 2 provides the general notation and introduces the inferential problem in parametric models for the AUC. Here the classical approach and the proposed approach based on modified profile likelihoods are described. In Section 3, the theory is applied to the specific case of a binormal model and computations are illustrated. Section 4 reports simulation studies comparing the different methods and Section 5 shows the application to real-life data on imaging for detecting brain tumor. Finally, conclusions and future directions are given in Section 6.

2. Notation and the Inferential Problem

In this section we consider a generic parametric model for the AUC, where the Y and X components are assumed to follow the parametric distributions $F_Y(t;\theta_Y)$ and $F_X(t;\theta_X)$, identified by the finite-dimensional parameter vectors θ_Y and θ_X, respectively. Let us define $\theta = (\theta_Y, \theta_X)$ be the entire parameter vector of the model of dimension p, with $\theta \in \Theta \subseteq \mathbb{R}^p$. The AUC is then obtained as

$$A = \int_{-\infty}^{\infty} F_Y(t;\theta_Y)\, dF_X(t;\theta_X) \equiv g(F_Y(t;\theta_Y), F_X(t;\theta_X)), \tag{2}$$

where the functional relation between A and $(F_Y(\cdot), F_X(\cdot))$ is defined with $g(\cdot)$, for ease of notation.

With the scope of making inference on the AUC, let $y = (y_1, \ldots, y_{n_1})$ be a random sample of size n_1 of i.i.d. observations drawn from Y, and $x = (x_1, \ldots, x_{n_2})$ be a random sample of size n_2 of i.i.d. observations drawn from X. Assume also that Y and X are independent. Let $f_Y(y;\theta_Y)$ and $f_X(x;\theta_X)$ be the probability density functions associated to Y and X, respectively. The log-likelihood function for θ is defined as $\ell(\theta) = \ell(\theta; y, x) = \sum_{i=1}^{n_1} \log f_Y(y_i;\theta_Y) + \sum_{i=1}^{n_2} \log f_X(x_i;\theta_X)$, and under broad conditions, $\hat{\theta}$ is the maximum likelihood estimator (MLE) obtained as unique solution to the score equation $\ell_\theta(\theta) = \partial\ell(\theta)/\partial\theta = 0$. The MLE of the AUC can be directly obtained as $\hat{A} = g(\hat{\theta})$, due to the likelihood invariance property.

In the proposed approach, we intend to treat the parameter A as a scalar parameter of interest, while the remaining parameters that identify the parametric distributions of Y and X are considered as nuisance parameter. Then, the original model needs to be reparameterized so that $\psi = \psi(\theta) = A$ is the parameter of interest, as defined in (2), and $\lambda = \lambda(\theta)$ is a nuisance parameter vector of length $(p-1)$, obtained by a transformation of the original parameter θ. Therefore, we can write the likelihood function for the new parameters (ψ, λ) as

$$\ell(\psi, \lambda) = \sum_{i=1}^{n_1} \log f_Y(y_i;\psi,\lambda) + \sum_{i=1}^{n_2} \log f_X(x_i;\psi,\lambda)\,.$$

The MLEs $\hat{A} = \hat{\psi}$ and $\hat{\lambda}$ are the unique solutions to, respectively, the score equations $\ell_\psi(\psi,\lambda) = \partial\ell(\psi,\lambda)/\partial\psi = 0$ and $\ell_\lambda(\psi,\lambda) = \partial\ell(\psi,\lambda)/\partial\lambda = 0$.

2.1 Inference Based on the Profile Likelihood

From $\ell(\psi,\lambda)$, classical likelihood inference for the parameter of interest $\psi = A$ in presence of nuisance parameters, can be based on profile likelihood procedures, which require to eliminate the nuisance parameter λ by replacing it by the constrained MLE, $\hat{\lambda}_\psi$, obtained by maximizing $\ell(\psi,\lambda)$ with respect to λ for fixed ψ. This method is based on the profile log-likelihood $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$, which can then be easily maximized to get the estimated AUC, $\hat{\psi} = \hat{A}$. The related standard error can be computed as $(J_p(\hat{\psi}))^{-1/2}$, where $J_p(\psi) = -\partial^2\ell_p(\psi)/\partial\psi^2$ is the corresponding profile observed Fisher information.

Confidence intervals and test of hypothesis can rely on first-order approximations. Specifically, inference on A can be based on the Wald statistic

$$W_p(\psi) = J_p(\hat{\psi})^{1/2}(\hat{\psi} - \psi)\,, \tag{3}$$

or on the signed log-likelihood ratio statistic

$$R_p(\psi) = \text{sign}(\hat{\psi} - \psi)\left(2(\ell_p(\hat{\psi}) - \ell_p(\psi))\right)^{1/2} , \tag{4}$$

which have asymptotic standard normal distributions.

A $100(1-\alpha)\%$ confidence interval for ψ based on the Wald statistic is given as $[\hat{\psi} - z_{1-\alpha/2}\ j_p(\hat{\psi})^{-1/2},\ \hat{\psi} + z_{1-\alpha/2}\ j_p(\hat{\psi})^{-1/2}]$, where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution. Alternatively, a $100(1-\alpha)\%$ confidence interval for ψ can be constructed from the $R_p(\psi)$ statistic, and can be written as $\{\psi : |R_p(\psi)| \leq z_{1-\alpha/2}\}$. The Wald-type confidence interval is often preferred because it is very simple and immediate to be computed, as compared to the likelihood ratio confidence interval, which typically requires a numerical solution. However, it is well-known that in general inferential procedures based on the Wald statistics have a general poor performance and are less accurate than the procedures based on the signed log-likelihood ratio statistic, especially at the boundaries of the parameter space (Severini, 2000).

2.2 Inference Based on the Modified Profile Likelihood

The profile likelihood is a standard method for inference in large-sample situations, and does not always perform well in small-sample problems. When the focus of the inferential interest is a parameter ψ, while the remaining parameters are not of central concern (nuisance), an interesting alternative approach is based on the modified profile likelihoods. With the scope to improve inferences, these likelihoods consist of an adjustment to the classical profile likelihoods by the inclusion of a penalization term for the possible presence of nuisance parameters. The amount of the penalization depends on the information available for λ, and increases when this information is large. Modified profile likelihoods have also the appealing property of being invariant to interest-preserving reparametrizations. This last property means that inferential results obtained for (ψ, λ) are also valid for $(\eta(\psi), \xi(\psi, \lambda))$, where η and ξ are one-to-one transformations.

The general expression for a modified profile log-likelihood (Severini, 2000) is

$$\ell_{mp}(\psi) = \ell_p(\psi) + M(\psi), \tag{5}$$

where $\ell_p(\psi)$ is the profile log-likelihood and $M(\psi)$ is the modification term. For this term, a high degree of accuracy is obtained when it has the expression

$$M(\psi) = \frac{\left|J_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\right|^{1/2}}{\left|\ell_{\lambda;\hat{\lambda}}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda})\right|}, \tag{6}$$

where

$$J_{\lambda\lambda}(\psi, \lambda) = -\ell_{\lambda\lambda}(\psi, \lambda) = -\partial^2\ell(\psi, \lambda)/\partial\lambda\partial\lambda^T, \qquad \ell_{\lambda;\hat{\lambda}}(\psi, \lambda; \hat{\psi}, \hat{\lambda}) = -\partial^2\ell(\psi, \lambda; \hat{\psi}, \hat{\lambda})/\partial\lambda\partial\hat{\lambda}^T .$$

In practice, the first term $J_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ is easily computed numerically or analytically by differentiation of $\ell_{\lambda\lambda}(\psi, \lambda)$. When the log-likelihood can be written in terms of the MLE, $\hat{\psi}$ and $\hat{\lambda}$, and an ancillary statistic a, i.e., as $\ell(\psi, \lambda; y, x) = \ell(\psi, \lambda; \hat{\psi}, \hat{\lambda}, a)$, computation of the term $\ell_{\lambda;\hat{\lambda}}(\psi, \lambda; \hat{\psi}, \hat{\lambda})$ is also straightforward. When differentiating with respect to $\hat{\lambda}$, the quantities ψ, $\hat{\psi}$ and a need to be held fixed. However, we have here omitted the conditioning to the ancillary a because it is not needed explicitly for computations and, in our context of parametric models for the AUC, in most of the cases the modification term in (6) can be obtained without specifying a.

Inference for ψ can be easily performed by treating (5) as a standard log-likelihood for ψ, without the burden of dealing with nuisance parameters. The solution to the maximization of $\ell_{mp}(\psi)$ provides a maximum modified profile likelihood estimate (MMLE), defined as $\hat{\psi}_{mp}$. In particular, the standard error associated to $\hat{\psi}_{mp}$ is computed as $(J_{mp}(\hat{\psi}_{mp}))^{-1/2}$, where $J_{mp}(\psi) = -\partial^2\ell_{mp}(\psi)/\partial\psi^2$. Therefore, using the normal approximation, it is possible to use a Wald-type confidence interval, e.g., $[\hat{\psi}_{mp} - z_{1-\alpha/2}\ J_{mp}(\hat{\psi}_{mp})^{-1/2},\ \hat{\psi}_{mp} + z_{1-\alpha/2}\ J_p(\hat{\psi}_{mp})^{-1/2}]$.

Moreover, the resulting signed modified log-likelihood ratio statistic, defined as

$$R_{mp}(\psi) = \text{sign}(\hat{\psi}_{mp} - \psi)\left(2(\ell_{mp}(\hat{\psi}_{mp}) - \ell_{mp}(\psi))\right)^{1/2} , \tag{7}$$

has asymptotic standard normal distribution, and has properties that are superior to those of the usual signed likelihood ratio statistic (Sartori, 2003). The statistic $R_{mp}(\psi)$ is then preferred, with respect to the Wald-type statistic, for construction of confidence intervals and test of hypothesis. In practice, a $100(1-\alpha)\%$ confidence interval based on $R_{mp}(\psi)$ is given as $\{\psi : |R_{mp}(\psi)| \leq z_{1-\alpha/2}\}$. A one-sided statistical test with null hypothesis $H_0 : \psi = \psi_0$ can be performed using the test-statistic $R_{mp}(\psi_0)$.

3. An Important Example: the Binormal Model

The main example about possible applications of the theory described in Subsections 2.2, is given for the popular binormal model, where Y and X are normally distributed with different means and variances, e.g., $Y \sim N(\mu_Y, \sigma_Y^2)$ and $X \sim N(\mu_X, \sigma_X^2)$. Under this assumption, it is known (Kotz et al., 2003) that the AUC can be written as

$$A = \Phi(\delta) = \Phi\left(\frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right), \tag{8}$$

where $\Phi(\cdot)$ is the cumulative probability function of the standard normal distribution. Denote with $\delta = (\mu_X - \mu_Y)/\sqrt{\sigma_X^2 + \sigma_Y^2}$ the quantile of the standard normal which provide an area equal to A. Here, there are two possible interesting choices for the parameter of interest ψ. We may have either $\psi = A$ or $\psi = \delta$. These two choices are equivalent in terms of inferential results because both the profile likelihood and the modified profile likelihood are invariant for interest-preserving reparameterizations, and thus for the transformation $A = \Psi(\delta)$. In the current paper, for practical reasons, we illustrate the procedures for the second choice $\psi = \delta$, since this case is relatively simpler to implement. Moreover, in this case, convergence in the corresponding parameter space $\Psi = \mathbb{R}$ is always obtained, whereas the choice $\psi = A$ with parameter space $\Psi = [0, 1]$ may yield computational problems on the boundaries.

We study the parameter of interest $\psi = \delta$, while the nuisance parameter can be chosen to be, e.g., $\lambda = (\lambda_1, \lambda_2, \lambda_3)$, with $\lambda_1 = \mu_Y$, $\lambda_2 = \sqrt{\sigma_Y^2}$, and $\lambda_3 = \sqrt{\sigma_Y^2 + \sigma_X^2}$. Other choices are also possible, where the parameter space is $\Psi \times \Lambda$, and thus the range of λ is independent of the range of ψ.

Given the MLE $\hat{\theta}$ computed from the original likelihood $\ell(\theta)$, by the invariance property, the MLE for the AUC is

$$\hat{A} = \Phi\left(\hat{\delta}\right) = \Phi\left(\frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\hat{\sigma}_Y^2 + \hat{\sigma}_X^2}}\right),$$

where $\hat{\mu}_Y = \sum_i y_i/n_1$, $\hat{\mu}_X = \sum_i x_i/n_2$ and $\hat{\sigma}_Y^2 = \sum_i (y_i - \hat{\mu}_Y)^2/n_1$, $\hat{\sigma}_X^2 = \sum_i (x_i - \hat{\mu}_X)^2/n_2$.

Consider now the likelihood function for the new parameters (ψ, λ),

$$\ell(\psi, \lambda) = -\frac{1}{2}\left[n_1 \log \lambda_2^2 + n_2 \log(\lambda_3^2 - \lambda_2^2)\right] - \frac{n_1\left[\hat{\lambda}_2^2 + (\hat{\lambda}_1 - \lambda_1)^2\right]}{2\lambda_2^2} - \frac{n_2\left[\hat{\lambda}_3^2 - \hat{\lambda}_2^2 + (\hat{\lambda}_1 + \hat{\psi}\hat{\lambda}_3 - \lambda_1 - \psi\lambda_3)^2\right]}{2(\lambda_3^2 - \lambda_2^2)}, \tag{9}$$

and observe that it is a function only of the unknown parameters and the minimal sufficient statistic $(\hat{\psi}, \hat{\lambda})$, where $\hat{\psi} = \hat{\delta}$, and $\hat{\lambda}_1 = \hat{\mu}_Y$, $\hat{\lambda}_2 = \sqrt{\hat{\sigma}_Y^2}$ and $\hat{\lambda}_3 = \sqrt{\hat{\sigma}_Y^2 + \hat{\sigma}_X^2}$, and thus, depends on the data only through the MLEs.

The constrained MLE $\hat{\lambda}_\psi = (\hat{\lambda}_{1\psi}, \hat{\lambda}_{2\psi}, \hat{\lambda}_{3\psi})$ for fixed ψ is found by numerical procedures as solution to the system of score equations $\ell_{\lambda_i}(\psi, \lambda) = \partial\ell(\psi, \lambda)/\partial\lambda_i = 0$, for $i = 1, 2, 3$. Their analytic expressions is given in the Appendix. The profile log-likelihood $\ell_p(\psi, \hat{\lambda}_\psi)$ is then obtained by replacing λ with $\hat{\lambda}_\psi$ in (9).

For the binormal model, computation of the signed log-likelihood ratio statistic $R_p(\psi)$ given in (4) is then straightforward. The Wald statistic $W_p(\psi)$ in (3) requires to find the observed information $J_p(\hat{\psi})$, which can be computed analytically or by a numerical procedure, for example by using the function `hessian` of package `numDeriv` in the `R` software.

The key parameter of interest is the AUC, therefore we can easily obtain inferential conclusions on A from those obtained from ψ. For example, the Delta method can be applied to find the standard error of $\hat{A} = \Phi(\hat{\psi})$, which is then equal to $\hat{s}_A = \Phi'(\hat{\psi})(J_p(\hat{\psi}))^{-1/2} = f_Z((\hat{\psi})(J_p(\hat{\psi}))^{-1/2}$, with $f_Z(\cdot)$ being the p.d.f. of the standard normal. Therefore, a Wald-type confidence interval for A is given as $[\hat{A} - z_{1-\alpha/2}\ \hat{s}_A,\ \hat{A} + z_{1-\alpha/2}\ \hat{s}_A]$, and a hypothesis testing concerning A can be based on the test-statistic $(\hat{A} - A_0)/\hat{s}_A$.

In addition, to specify the modified profile log-likelihood in (5) and (6), we need to compute the modification term $M(\psi)$. In doing so, the block of the observed information matrix, $J_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ is equal to minus the Hessian matrix, which can be easily obtained by numerical procedures in the `R` software, as above. The analytic expressions of the sample space derivatives $\ell_{\lambda;\hat{\lambda}}(\psi, \lambda; \hat{\psi}, \hat{\lambda})$ for the binormal model are provided in the Appendix. The signed modified log-likelihood ratio statistic $R_{mp}(\psi)$ given in (7) can then be constructed to solve test of hypothesis concerning key values of the AUC, such as e.g. $A = 0.5$ or $A = 1$. For example, the one-sided test with hypotheses $H_0 : \psi = \psi_0 = 0$ versus $H_0 : \psi > 0$ is equivalent to testing whether the AUC is significantly higher than 0.5, and can be performed using the test-statistic

$R_{mp}(\psi_0) = \text{sign}(\hat{\psi}^0_{mp} - \psi_0)\left(2(\ell_{mp}(\hat{\psi}^0_{mp}) - \ell_{mp}(\psi_0))\right)^{1/2}$, where $\hat{\psi}^0_{mp}$ denotes the maximum modified likelihood estimate of ψ in the parameter space $\Psi_0 = \{\psi \in \Psi : \psi > \psi_0\}$.

4. Simulation Studies

The performance of the proposed method for constructing confidence intervals and point estimates for A is illustrated through a simulation study, based on 5000 Monte Carlo trials. We considered different values of ψ ($\psi = 0.6, 0.8, 0.95, 0.99$) and many different combinations of sample sizes $(n_1, n_2) = (5, 5)$, $(10, 10)$, $(20, 20)$, $(30, 30)$, $(15, 5)$, $(5, 15)$, $(30, 5)$, $(5, 30)$, $(80, 10)$, $(10, 80)$. Note that the case of very unbalanced samples is also taken into account.

First, the simulation studies investigated the coverage probabilities of 95% confidence intervals based on the signed profile log-likelihood ratio statistic $R_p(\psi)$, the signed modified profile log-likelihood ratio statistic $R_{mp}(\psi)$, and the Wald statistic $W_p(\psi)$. All these statistics are asymptotically distributed as standard normal, and the approximation is often more accurate for $R_{mp}(\psi)$. Results in Table 1 show that $R_{mp}(\psi)$ is more accurate than $R_p(\psi)$ and $W_p(\psi)$, in terms of both central coverage probability and symmetry of the error rates, for all the considered AUC values and sample sizes. Of course, for all methods, we observe a less accurate coverage when the sample sizes are very small ($(n_1, n_2) = (5, 5), (10, 10)$), which then increases for higher sample sizes. However, the $R_{mp}(\psi)$ coverage is observed to reach nearly the 95% nominal level, being slightly affected by low values of sample sizes (see, e.g., for $A = 0.95, 0.99$), in contrast to the $W_p(\psi)$ and $R_p(\psi)$ that provide seriously poor performance for small samples. Very interestingly, this poor performance becomes even worse for higher values of the AUC, such as $A = 0.95, 0.99$. On the contrary, the good performance of the $R_{mp}(\psi)$ seems to be very stable for all values of the AUC.

An important result is observed for unbalanced samples: $W_p(\psi)$ and $R_p(\psi)$ seem to be negatively affected by the sample unbalance, since their coverage decreases even more with respect to the nominal level, whereas, the $R_{mp}(\psi)$ coverage keeps stable and enough accurate in all the unbalanced settings. In particular, we note that the coverages are lower for samples with high n_1 and low n_2 (e.g., $(n_1, n_2) = (30, 5), (80, 10)$) as compared to the inverse case of low n_1 and high n_2. This fact may depend on the reparameterization chosen for the nuisance parameters λ, since we have that the MLEs $\hat{\lambda}_2$ and $\hat{\lambda}_3$ are both affected by the small sample size n_2 and then would be poorly estimated.

The very poor performance shown by the Wald statistic for high values of the AUC is expected. It is well known that when the profile log-likelihood is not quadratic around the MLE, as it happens in our AUC study (see Figure 1 of data example in Section 5), the Wald statistic may lead to very asymmetric confidence intervals. In fact, in Table 1 we observe a nearly null empirical lower error and a higher empirical upper error than expected. Asymmetric errors are also seen for the $R_p(\psi)$ statistic, although the discrepancy from the expected errors is negligible.

Simulation studies were also used to evaluate the properties of the $R_{mp}(\psi)$-based estimator of A, in comparison with the MLE $\hat{\psi}$. The two estimators are compared in terms of median bias and results are shown in Table 2, where estimated standard errors and simulations-based (empirical) median absolute deviation (MAD) are also reported. The choice of a median-bias criteria is due to the median unbiasedness property of the $\hat{\psi}_{mp}$-based estimator, and it is more robust under model misspecification. It can be noted that the estimator based on modified likelihood, $\hat{\psi}_{mp}$, is preferable to the MLE in terms of the considered criteria, since it is less median-biased than the MLE, in particular for small sample sizes and unbalanced samples. Estimates seem to be more biased for unbalanced samples with high n_1 and low n_2. However, this problem is attenuated when the AUC value increases, and the bias of $\hat{\psi}_{mp}$ reduces to about the half of the bias of the MLEs.

5. A Worked Data Example

In this section, an application of the inferential approaches discussed in the current paper to real-life data is presented. We consider data from imaging studies used for brain tumor grading. The data have been originally collected in Tsuchida, Takeuchi, Okazawa, Tsujikawa, and Fujibayashi (2008), and were also discussed in the paper by Feng et al. (2015). This data are also available in the `R` package `auRoc` (Feng, 2015). The objective of the study was to evaluate the clinical significance of 1-11C-acetate (ACE) positron emission tomography (PET) in 10 patients with brain glioma, in comparison with 18F-fluorodeoxyglucose (FDG) PET. FDG and ACE are two different imaging techniques for detecting brain glioma. The aim of this section is to examine again the diagnostic accuracy of both techniques in discriminating between patients with low grade (grades I or II) and patients with high grade (grades III and IV). Patients grading was previously determined by magnetic resonance imaging, a gold-standard method used to classify patients with brain glioma in low and high grade classes. Five patients were characterized as low grade and the other five patients as high grade. All patients underwent FDG and ACE diagnostic measurements and the standard uptake value (SUV) was calculated for the same regions of interest in the brain. These SUV values were compared between low grade and high grade patients. The diagnostic accuracy of FDG and ACE was investigated by estimating the area under the ROC curve. Point estimates, confidence intervals and test of hypothesis were performed following the three approached presented in the paper.

Table 1. Two-sided empirical coverage of confidence intervals with 95% nominal levels for A based on the Wald statistic $W_p(\psi)$, the profile log-likelihood ratio statistic $R_p(\psi)$ and the modified profile log-likelihood ratio statistic $R_{mp}(\psi)$, under the binormal model. The central coverage probabilities and the non-coverage probabilities on the left and right tails, which represent, respectively, the lower and upper errors, are reported.

		$W_p(\psi)$			$R_p(\psi)$			$R_{mp}(\psi)$		
A	(n_1, n_2)	coverage	lower	upper	coverage	lower	upper	coverage	lower	upper
	(5, 5)	0.841	0.049	0.110	0.920	0.032	0.049	0.940	0.027	0.033
	(10, 10)	0.894	0.030	0.075	0.933	0.026	0.041	0.942	0.024	0.034
	(20, 20)	0.920	0.032	0.048	0.938	0.031	0.031	0.942	0.031	0.027
	(30, 30)	0.936	0.023	0.040	0.947	0.024	0.029	0.950	0.023	0.027
0.6	(15, 5)	0.830	0.057	0.114	0.908	0.038	0.054	0.938	0.029	0.033
	(5, 15)	0.890	0.037	0.073	0.926	0.031	0.042	0.946	0.025	0.029
	(30, 5)	0.826	0.057	0.116	0.908	0.040	0.052	0.938	0.031	0.030
	(5, 30)	0.882	0.047	0.071	0.919	0.040	0.041	0.946	0.028	0.026
	(80, 10)	0.891	0.040	0.069	0.928	0.034	0.038	0.943	0.029	0.028
	(10, 80)	0.920	0.031	0.050	0.939	0.028	0.033	0.953	0.022	0.025
	(5, 5)	0.776	0.018	0.206	0.912	0.026	0.062	0.935	0.026	0.039
	(10, 10)	0.863	0.011	0.127	0.934	0.021	0.045	0.944	0.022	0.034
	(20, 20)	0.905	0.011	0.084	0.945	0.020	0.036	0.948	0.022	0.030
	(30, 30)	0.912	0.011	0.077	0.944	0.020	0.036	0.947	0.022	0.031
0.8	(15, 5)	0.765	0.013	0.221	0.910	0.023	0.068	0.938	0.023	0.039
	(5, 15)	0.881	0.013	0.106	0.934	0.023	0.042	0.948	0.023	0.028
	(30, 5)	0.771	0.022	0.207	0.896	0.031	0.072	0.931	0.028	0.041
	(5, 30)	0.907	0.014	0.079	0.939	0.024	0.037	0.949	0.023	0.027
	(80, 10)	0.863	0.014	0.123	0.928	0.027	0.046	0.938	0.027	0.035
	(10, 80)	0.935	0.014	0.051	0.951	0.021	0.029	0.956	0.020	0.023
	(5, 5)	0.648	0.001	0.351	0.905	0.013	0.082	0.941	0.017	0.042
	(10, 10)	0.764	0.002	0.235	0.924	0.017	0.059	0.942	0.021	0.037
	(20, 20)	0.840	0.001	0.159	0.939	0.019	0.042	0.946	0.023	0.032
	(30, 30)	0.874	0.002	0.124	0.945	0.021	0.034	0.950	0.024	0.026
0.95	(15, 5)	0.656	0.001	0.343	0.896	0.018	0.086	0.936	0.021	0.043
	(5, 15)	0.809	0.001	0.190	0.939	0.015	0.046	0.954	0.019	0.027
	(30, 5)	0.653	0.002	0.345	0.896	0.016	0.088	0.936	0.018	0.046
	(5, 30)	0.857	0.002	0.141	0.939	0.020	0.040	0.958	0.018	0.024
	(80, 10)	0.770	0.001	0.229	0.926	0.015	0.059	0.942	0.018	0.040
	(10, 80)	0.908	0.006	0.086	0.944	0.023	0.033	0.949	0.025	0.026
	(5, 5)	0.562	0.000	0.438	0.898	0.011	0.091	0.944	0.017	0.038
	(10, 10)	0.687	0.000	0.313	0.922	0.014	0.064	0.940	0.019	0.041
	(20, 20)	0.774	0.000	0.226	0.939	0.017	0.044	0.947	0.021	0.032
	(30, 30)	0.824	0.000	0.176	0.945	0.020	0.035	0.953	0.023	0.024
0.99	(15, 5)	0.749	0.000	0.251	0.937	0.015	0.048	0.948	0.018	0.034
	(5, 15)	0.562	0.000	0.438	0.898	0.011	0.091	0.947	0.015	0.039
	(30, 5)	0.817	0.000	0.183	0.942	0.016	0.042	0.946	0.021	0.032
	(5, 30)	0.563	0.000	0.437	0.898	0.010	0.092	0.948	0.014	0.038
	(80, 10)	0.890	0.001	0.109	0.942	0.022	0.037	0.944	0.024	0.032
	(10, 80)	0.674	0.000	0.326	0.922	0.014	0.064	0.941	0.019	0.040

Table 2. Empirical median biases (Bias), median absolute deviations (MAD) and estimated standard errors (SE) of the estimators for the AUC obtained from the profile likelihood ($\hat{A} = \Phi(\hat{\psi})$) and from the modified profile likelihood ($\hat{A}_{mp} = \Phi(\hat{\psi}_{mp})$), in the binormal model.

		$\hat{A} = \Phi(\hat{\psi})$			$\hat{A}_{mp} = \Phi(\hat{\psi}_{mp})$		
A	(n_1, n_2)	Bias	MAD	SE	Bias	MAD	SE
	(5, 5)	0.0170	0.1938	0.1561	0.0099	0.1811	0.1574
	(10, 10)	0.0121	0.1323	0.1173	0.0083	0.1281	0.1178
	(20, 20)	0.0029	0.0903	0.0853	0.0010	0.0890	0.0854
	(30, 30)	0.0014	0.0720	0.0702	0.0002	0.0713	0.0703
0.6	(15, 5)	0.0131	0.1757	0.1400	0.0082	0.1672	0.1437
	(5, 15)	0.0115	0.1314	0.1144	0.0075	0.1267	0.1165
	(30, 5)	0.0131	0.1763	0.1343	0.0094	0.1693	0.1404
	(5, 30)	0.0060	0.1131	0.0969	0.0039	0.1103	0.1018
	(80, 10)	0.0078	0.1140	0.1023	0.0058	0.1122	0.1046
	(10, 80)	0.0019	0.0728	0.0684	0.0007	0.0722	0.0709
	(5, 5)	0.0378	0.1450	0.1177	0.0203	0.1432	0.1234
	(10, 10)	0.0193	0.1017	0.0917	0.0112	0.1001	0.0941
	(20, 20)	0.0090	0.0725	0.0682	0.0050	0.0721	0.0690
	(30, 30)	0.0075	0.0572	0.0565	0.0048	0.0570	0.0569
0.8	(15, 5)	0.0374	0.1444	0.0812	0.0222	0.1433	0.1201
	(5, 15)	0.0132	0.0873	0.1130	0.0040	0.0861	0.0829
	(30, 5)	0.0390	0.1401	0.1125	0.0249	0.1391	0.1191
	(5, 30)	0.0071	0.0682	0.0631	-0.0034	0.0687	0.0643
	(80, 10)	0.0140	0.1004	0.0895	0.0072	0.0992	0.0922
	(10, 80)	0.0025	0.0416	0.0414	-0.0002	0.0417	0.0420
	(5, 5)	0.0263	0.0340	0.0511	0.0152	0.0469	0.0610
	(10, 10)	0.0119	0.0399	0.0416	0.0055	0.0436	0.0458
	(20, 20)	0.0063	0.0321	0.0315	0.0030	0.0333	0.0331
	(30, 30)	0.0049	0.0257	0.0263	0.0027	0.0262	0.0272
0.95	(15, 5)	0.0242	0.0367	0.0514	0.0138	0.0487	0.0610
	(5, 15)	0.0085	0.0362	0.0368	0.0019	0.0407	0.0398
	(30, 5)	0.0239	0.0373	0.0515	0.0136	0.0494	0.0609
	(5, 30)	0.0052	0.0283	0.0282	0.0008	0.0295	0.0299
	(80, 10)	0.0116	0.0396	0.0410	0.0059	0.0431	0.0449
	(10, 80)	0.0019	0.0192	0.0186	0.0004	0.0195	0.0192
	(5, 5)	0.0075	0.0037	0.0197	0.0045	0.0082	0.0272
	(10, 10)	0.0043	0.0078	0.0150	0.0022	0.0104	0.0181
	(20, 20)	0.0025	0.0081	0.0108	0.0013	0.0091	0.0120
	(30, 30)	0.0016	0.0074	0.0089	0.0008	0.0079	0.0095
0.99	(15, 5)	0.0030	0.0083	0.0124	0.0014	0.0098	0.0142
	(5, 15)	0.0076	0.0036	0.0198	0.0047	0.0078	0.0277
	(30, 5)	0.0014	0.0077	0.0089	0.0006	0.0082	0.0096
	(5, 30)	0.0074	0.0038	0.0200	0.0045	0.0082	0.0282
	(80, 10)	0.0006	0.0050	0.0054	0.0003	0.0051	0.0056
	(10, 80)	0.0047	0.0074	0.0148	0.0027	0.0099	0.0179

Table 3. Point estimates (estimated A) and 95% confidence intervals (95% CI) based on the Wald, profile log-likelihood and modified profile log-likelihood statistics, for the FDG and ACE imaging techniques.

	FDG			ACE		
Method	Estimated A	SE	95% CI	Estimated A	SE	95% CI
Wald	0.726	0.160	(0.413, 1)	0.897	0.096	(0.709,1)
Profile lik.	0.726	0.160	(0.368, 0,939)	0.897	0.096	(0.585,0.990)
Modified profile lik.	0.714	0.163	(0.355, 0.934)	0.879	0.107	(0.548,0.986)

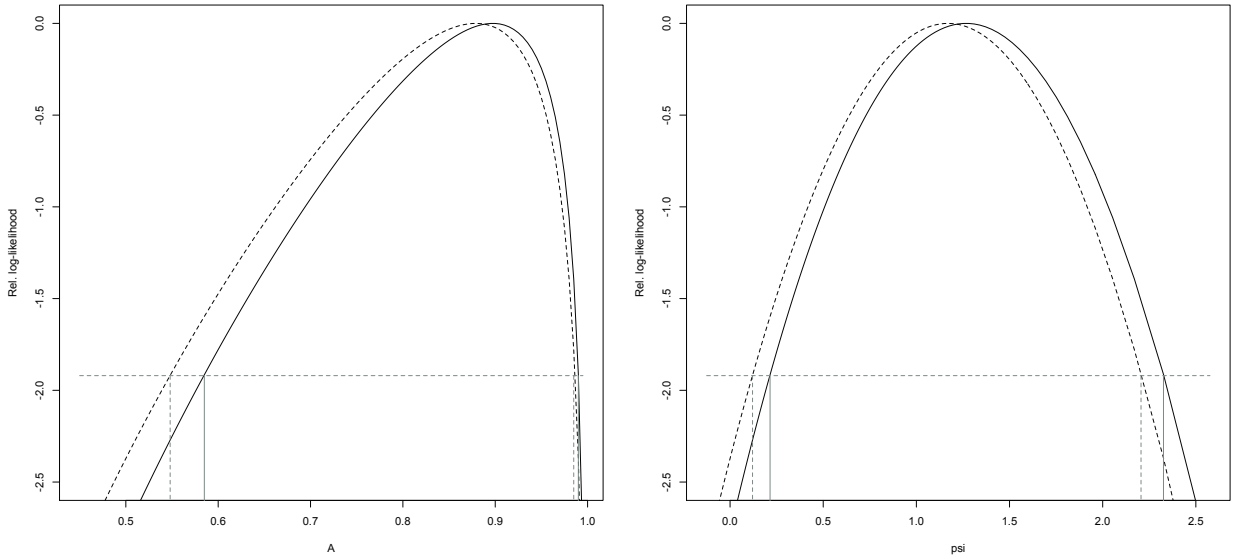

Figure 1. Plot of r_p (thick solid line) and r_p^* (thick dashed line) for a range of values of the parameter R. Vertical lines are drawn to identify confidence intervals for R based on r_p (thin solid line) and r_p^* (thin dashed line)

Assumption of normality in the distributions of SUVs from FDG and ACE in the low-grade and high-grade patients has been shown not to be violated (Feng et al., 2015). For FDG, the average SUV values in the low and high groups were, respectively, 4.714 and 7.124, while for ACE, the average SUV values in the low and high groups were, respectively, 1.850 and 2.626. Lower SUV values are associated to the low grade patients. Therefore, here the random variable Y represents the FDG SUV values in the low grade population, while the random variable X represents the FDG SUV values in the high grade population.

Table 3 summaries the main inferential results for the AUC computed, separately, for the FDG SUV values and the ACE SUV values. For the FDG, the different statistical methods gave very similar estimates of the area under the ROC curve, equal to ~0.7, showing that the FDG has poor discrimination accuracy between the low and high grade populations. The standard errors are also very similar, whereas the Wald confidence interval equal to $(0.413, 1)$ is right-shifted as compared to the confidence intervals based on the $R_p(\psi)$ and $R_{mp}(\psi)$ statistics, which are virtually identical. In addition, we performed a test for the null hypothesis $H_0 : A = 0.5$ versus the alternative $H_1 : A > 0.5$, and found that the $R_p(\psi)$ and $R_{mp}(\psi)$ statistics produce similar non significant p-values ($p = 0.105$ and $p = 0.119$, respectively), then suggesting that there is no evidence of any discriminatory power in the FDG technique.

From Table 3, we observe that also for the ACE, the different statistical methods gave similar estimates of the area under the ROC curve, equal to ~0.9. Thus, it was found that the ACE is much more accurate at discriminating. The Wald confidence interval equal to $(0.709, 1)$ is extremely and erroneously right-shifted, and thus it deviates from the other confidence intervals based on the $R_p(\psi)$ and $R_{mp}(\psi)$ statistics. These latter two differ in particular at the lower limit, as also illustrated in Figure 1. This fact is due to the skewed shape of the profile log-likelihood. In the case of ACE, the test of hypothesis with null $H_0 : A = 0.5$ gave significant results, which differ between the two $R_p(\psi)$- and $R_{mp}(\psi)$-based approaches ($p = 0.009$ and $p = 0.015$, respectively). This result indicates that the ACE technique has the ability to discriminate. When testing the null $H_0 : A = 0.6$, the resulting p-values ($p = 0.030$ and $p = 0.043$, respectively) provide

evidence of a discrimination accuracy above 0.6, but the more correct $R_{mp}(\psi)$-based approach gives less evidence for this conclusion. Note that here the power of the tests is low due to very small sample sizes.

Figure 1 reports the relative log-likelihoods, defined as $\ell(\theta) - \ell(\hat{\theta})$, for both the parameters A and ψ. It is noted that relative modified profile log-likelihoods are shifted to the left with respect to the relative profile log-likelihoods, due the adjustment term $M(\psi)$. Moreover, we observe that the $\Phi(\cdot)$ reparameterization on the parameter of interest has the natural effect to make the quadratic functions for $\psi = \delta$ become skewed to the left.

6. Discussion

The paper has presented the performance of a new inferential approach in parametric models for the AUC, which was shown to be useful and easy to implement. The proposed method was applied to make inference for the binormal model, and can immediately be adapted to any other parametric distribution. Alternatively, when the normality assumption is violated, a Box-Cox type power transformation to the original data can also be applied (Box & Cox, 1964; Faraggi & Reiser, 2002). The additional unknown parameters concerning the Box-Cox transformation may be either treated within the entire model as nuisance parameters, or one may, first, apply the appropriate transformation to the original data, and then use inference for the normal theory presented in this paper. We note that the presence of additional nuisance parameters to the model is not expected to affect the accuracy of the inferential results when a modified profile likelihood approach is adopted.

Profile likelihoods have a biased score function of order $O(1)$, which does not typically disappear asymptotically. Modified profile likelihoods have properties very similar to those of usual full likelihoods, and their adjustment term reduces the bias to order $O(n^{-1})$. Consequently, the signed likelihood ratio statistic based on the modified profile likelihood has properties that are superior to those of the usual signed likelihood ratio statistic (Cox & Barndorff-Nielsen, 1994).

The results from simulation studies show that inference based on the modified profile log-likelihood approach has superior performance compared to the classical profile log-likelihood approach, in terms of central coverage probability, symmetry of error rates, and median bias. Wald statistics can lead to seriously misleading inferential conclusions for small or unbalanced samples, especially at the boundaries of the parameter space (Molenberghs & Verbeke, 2007). Moreover, Wald-type tests of hypothesis may lead to erroneous significant results. For example, in the real data application for the ACE technique, it was found that $A_0 = 0.7$ falls outside the Wald confidence interval, and a test of hypothesis of the null $H_0 : A = 0.7$ yields a significant p-value of 0.02, in contrast to the profile likelihood approach that shows no evidence for a discrimination ability above 0.7.

The proposed approach has the potential to be applicable to any general parametric setting, for example, in settings where Y and X follow two different parametric distributions, or in models with mixture distributions, which are often used when the empirical distribution shows a bimodal behaviour. Moreover, future developments concerning the modified likelihood approach could be very relevant in the context of AUC estimation, especially when the data are stratified, or the interest of the inquiry is on modeling different stratum-specific AUC, for example by including stratum-specific fixed effects as nuisance parameters. Another important case of application of the proposed approach may be when the two random variables X and Y depend on covariates, or in general when the AUC models rely on many nuisance parameters (Sartori, 2003).

Acknowledgements

The author was supported by 'Progetto di Ateneo 2015' (CPDA153257), University of Padua.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*(4), 387–415. https:/doi.org/10.1016/0022-2496(75)90001-2

Box, G. EP., & Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*, 211–252.

Cortese, G., & Ventura, L. (2013). Accurate higher-order likelihood inference on $P(Y < X)$. *Computational statistics*, *28*(3), 1035–1059. https:/doi.org/10.1007/s00180-012-0343-z

Cox, D., & Barndorff-Nielsen, O. (1994). *Inference and asymptotics* (Vol. 52). CRC Press.

Cox, D., & Reid, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika*, *79*(2), 408–411. https:/doi.org/10.1093/biomet/79.2.408

Faraggi, D., & Reiser, B. (2002). Estimation of the area under the roc curve. *Statistics in Medicine*, *21*(20), 3093–3106. https:/doi.org/10.1002/sim.1228

Feng, D. (2015). auRoc: Various methods to estimate the AUC [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=auRoc` (R package version 0.1-0)

Feng, D., Cortese, G., & Baumgartner, R. (2015). A comparison of confidence/credible interval methods for the area under the ROC curve for continuous diagnostic tests with small sample size. *Statistical Methods in Medical Research*, Published online before print. https:/doi.org/10.1177/0962280215602040

Johnson, R. A. (1988). Stress-strength models for reliability. *Handbook of Statistics*, *7*, 27–54. https:/doi.org/10.1016/S0169-7161(88)07005-1

Kotz, S., Lumelskii, Y., & Pensky, M. (2003). The stress-strength model and its generalizations. *Theory and Applications. Singapore: World Scientific*, *43*, 44.

Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. CRC Press.

Molenberghs, G., & Verbeke, G. (2007). Likelihood ratio, score, and wald tests in a constrained parameter space. *The American Statistician*, *61*(1), 22–27. https:/doi.org/10.1198/000313007X171322

Obuchowski, N. A., & Lieber, M. L. (2002). Confidence bounds when the estimated ROC area is 1.0. *Academic Radiology*, *9*(5), 526–530. https:/doi.org/10.1016/s1076-6332(03)80329-x

Pardo-Fernández, J. C., Rodríguez-Álvarez, M. X., Van Keilegom, I., et al. (2013). *A review on ROC curves in the presence of covariates* (Tech. Rep.). UCL.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

Reiser, B. (2000). Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Statistics in Medicine*, *19*(16), 2115–2129.

Reiser, B., & Faraggi, D. (1997). Confidence intervals for the generalized ROC criterion. *Biometrics*, 644–652. https:/doi.org/10.2307/2533964

Reiser, B., & Guttman, I. (1986). Statistical inference for $Pr(Y < X)$: the normal case. *Technometrics*, *28*(3), 253–257.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, *90*(3), 533–549. https:/doi.org/10.1093/biomet/90.3.533

Severini, T. A. (2000). *Likelihood methods in statistics*. Oxford University Press.

Tsuchida, T., Takeuchi, H., Okazawa, H., Tsujikawa, T., & Fujibayashi, Y. (2008). Grading of brain glioma with 1-11 C-acetate PET: comparison with 18 F-FDG PET. *Nuclear Medicine and Biology*, *35*(2), 171–176. https:/doi.org/10.1016/j.nucmedbio.2007.11.004

Wolfe, D. A., & Hogg, R. V. (1971). On constructing statistics and reporting data. *The American Statistician*, *25*(4), 27–30.

Zhou, X.-H., McClish, D. K., & Obuchowski, N. A. (2009). *Statistical methods in diagnostic medicine* (Vol. 569). John Wiley & Sons.

Zou, K. H., Carlsson, M. O., & Yu, C.-R. (2012). Comparison of adjustment methods for stratified two-sample tests in the context of ROC analysis. *Biometrical Journal*, *54*(2), 249–263. https:/doi.org/10.1002/bimj.201000251

Appendix

The analytic expressions of the partial derivatives of $\ell(\psi, \lambda)$ with respect to the nuisance parameters for the binormal model are given hereafter:

$$\ell_{\lambda_1}(\psi,\lambda) = \frac{n_1(\hat{\lambda}_1 - \lambda_1)}{\lambda_2^2} + \frac{n_2(\hat{\lambda}_1 + \hat{\psi}\hat{\lambda}_3 - \lambda_1 - \psi\lambda_3)}{\lambda_3^2 - \lambda_2^2}$$

$$\ell_{\lambda_2}(\psi,\lambda) = -\frac{n_1}{\lambda_2} + \frac{n_2\lambda_2}{\lambda_3^2 - \lambda_2^2} - \frac{n_2\lambda_2\left[\hat{\lambda}_3^2 - \hat{\lambda}_2^2 + (\hat{\lambda}_1 + \hat{\psi}\hat{\lambda}_3 - \lambda_1 - \psi\lambda_3)^2\right]}{(\lambda_3^2 - \lambda_2^2)^2} + \frac{n_1\left[\hat{\lambda}_2^2 + (\hat{\lambda}_1 - \lambda_1)^2\right]}{\lambda_2^3}$$

$$\ell_{\lambda_3}(\psi,\lambda) = -\frac{n_2\lambda_3}{\lambda_3^2 - \lambda_2^2} + \frac{n_2\lambda_3\left[\hat{\lambda}_3^2 - \hat{\lambda}_2^2 + (\hat{\lambda}_1 + \hat{\psi}\hat{\lambda}_3 - \lambda_1 - \psi\lambda_3)^2\right]}{(\lambda_3^2 - \lambda_2^2)^2} + \frac{n_2\psi(\hat{\lambda}_1 + \hat{\psi}\hat{\lambda}_3 - \lambda_1 - \psi\lambda_3)}{\lambda_3^2 - \lambda_2^2}$$

The analytic expressions of the sample space derivatives $\ell_{\lambda;\hat{\lambda}}(\psi,\lambda)$ for the binormal model follow hereafter:

$$
\begin{aligned}
\ell_{\lambda_1;\hat{\lambda}_1}(\psi,\lambda) =& \frac{n_1}{\lambda_2^2} + \frac{n_2}{\lambda_3^2-\lambda_2^2} \\
\ell_{\lambda_1;\hat{\lambda}_2}(\psi,\lambda) =& 0 \\
\ell_{\lambda_1;\hat{\lambda}_3}(\psi,\lambda) =& \frac{n_2\hat{\psi}}{\lambda_3^2-\lambda_2^2} \\
\ell_{\lambda_2;\hat{\lambda}_1}(\psi,\lambda) =& -\frac{2n_2\lambda_2(\hat{\lambda}_1+\hat{\psi}\hat{\lambda}_3-\lambda_1-\psi\lambda_3)}{(\lambda_3^2-\lambda_2^2)^2} + \frac{2n_1(\hat{\lambda}_1-\lambda_1)}{\lambda_2^3} \\
\ell_{\lambda_2;\hat{\lambda}_2}(\psi,\lambda) =& \frac{2n_2\lambda_2\hat{\lambda}_2}{(\lambda_3^2-\lambda_2^2)^2} + \frac{2n_1\hat{\lambda}_2}{\lambda_2^3} \\
\ell_{\lambda_2;\hat{\lambda}_3}(\psi,\lambda) =& -\frac{2n_2\lambda_2\left[\hat{\lambda}_3+\hat{\psi}(\hat{\lambda}_1+\hat{\psi}\hat{\lambda}_3-\lambda_1-\psi\lambda_3)\right]}{(\lambda_3^2-\lambda_2^2)^2} \\
\ell_{\lambda_3;\hat{\lambda}_1}(\psi,\lambda) =& \frac{2n_2\lambda_3(\hat{\lambda}_1+\hat{\psi}\hat{\lambda}_3-\lambda_1-\psi\lambda_3)}{(\lambda_3^2-\lambda_2^2)^2} + \frac{n_2\psi}{\lambda_3^2-\lambda_2^2} \\
\ell_{\lambda_3;\hat{\lambda}_2}(\psi,\lambda) =& -\frac{2n_2\lambda_3\hat{\lambda}_2}{(\lambda_3^2-\lambda_2^2)^2}
\end{aligned}
$$