Bayesian bootstrap inference for the ROC surface

Vanda Inácio de Carvalho*
Pontificia Universidad Católica de Chile, Chile, icalhau@mat.puc.cl

Miguel de Carvalho
Pontificia Universidad Católica de Chile, Chile, mdecarvalho@mat.puc.cl

Abstract

The receiver operating characteristic (ROC) surface is a popular tool for evaluating the accuracy of medical diagnostic tests that classify individuals into one of three ordered classes. We propose a fully nonparametric method based on the Bayesian bootstrap for conducting inferences for the ROC surface and its functionals, such as the volume under the surface. The proposed estimator is based on a simple, yet interesting, representation of the ROC surface in terms of placement variables. As an illustration of the proposed methods, we analyze data concerning the diagnosis of neurological impairment in HIV patients.

Keywords: Bayesian bootstrap; HIV data; ROC surface; volume under the surface.

1. Introduction

Medical diagnostic tests play a crucial role in health care and its accuracy must be rigorously assessed through statistical analysis before the test is approved for use in practice. The receiver operating characteristic (ROC) curve is a popular tool for evaluating the accuracy of a diagnostic test that classifies subjects into two populations: diseased and non-diseased. However, in practice, clinicians often face situations that require a decision among three or more diagnostic alternatives. For example, cognitive function declines from normal function to mild impairment, to severe impairment and/or dementia. The ROC surface has been proposed to assess the diagnostic accuracy in ordered three-class classification problems as a direct generalization of ROC curves (Nakas & Yiannoutsos, 2004). The volume under the ROC surface (VUS) has also been proposed as a summary measure of diagnostic accuracy, and it is the analogous for the three-class case of the area under the ROC curve in the two-class setting.

There is vast literature on parametric, semiparametric, and nonparametric ROC curve analysis; see Zhou et al. (2011) for an overview. The amount of existing work on ROC surface analysis is, by comparison, limited. Relevant works for estimating the surface include Li & Zhou (2009) who developed a frequentist nonparametric and semiparametric approach, Inácio et al. (2011) who proposed an estimator based on mixtures of finite Polya trees, and Zan & Wang (2013) who proposed an estimator based on polytomous logistic regression procedures.

In this short paper, we develop a smooth, flexible, and computationally appealing estimator for the ROC surface based on Bayesian bootstrap (BB). The BB (Rubin, 1981) is the Bayesian analog of the frequentist bootstrap (Efron, 1979) but gives smoother choices of weights; more details on the BB are given in Section 2.2. Our estimator can be regarded as an extension to the three-class setting the estimator proposed by Gu et al. (2008) for the ROC curve.

2. Methodology

2.1. ROC surface and placement variables

Let Y be a continuous random variable denoting the test outcome; without loss of generality assume that subjects from class 3 tend to have higher values of Y than subjects in class 2, and the latter tend to have higher values of Y than class 1 subjects. Further, let Y_1, Y_2, and Y_3 denote the test outcomes in classes 1, 2, and 3, with F_1, F_2, and F_3 being the respective distribution functions. For any pair of ordered cutoff values, (c_1, c_2) with $c_1 < c_2$, an individual is classified into class 1 if $Y \leqslant c_1$, into class 2 if $c_1 < Y \leqslant c_2$, and into class 3 otherwise. Thus, the probabilities of correct classification into each class are as follows

$$\begin{aligned}
p_1(c_1, c_2) &= \Pr(Y_1 \leqslant c_1) = F_1(c_1),\\
p_2(c_1, c_2) &= \Pr(c_1 < Y_2 \leqslant c_2) = F_2(c_2) - F_2(c_1),\\
p_3(c_1, c_2) &= \Pr(Y_3 > c_2) = 1 - F_3(c_2).
\end{aligned}$$

The ROC surface is then the three-dimensional plot in the unit cube depicting the probabilities of correct classification into each class

$$\{(F_1(c_1), F_2(c_2) - F_2(c_1), 1 - F_3(c_2)) : (c_1, c_2) \in \mathbb{R}^2, c_1 < c_2\}.$$

For the sake of simplicity, hereafter we drop the dependence of p_1, p_2, and p_3, on c_1 and c_2. By writing $c_1 = F_1^{-1}(p_1)$ and $c_2 = F_3^{-1}(1 - p_3)$, we obtain the functional form of the ROC surface

$$\mathrm{ROCS}(p_1, p_3) = \begin{cases} F_2(F_3^{-1}(1-p_3)) - F_2(F_1^{-1}(p_1)), & \text{if } F_1^{-1}(p_1) < F_3^{-1}(1-p_3),\\ 0, & \text{otherwise.}\end{cases}$$

The volume under the ROC surface is a summary measure of the overall diagnostic accuracy and it is defined as

$$\mathrm{VUS} = \int_0^1 \int_0^1 \mathrm{ROCS}(p_1, p_3)\mathrm{d}p_1\mathrm{d}p_3 = \Pr(Y_1 < Y_2 < Y_3).$$

When the three distributions completely overlap, and thus the test has no discriminatory ability, the VUS takes the value 1/6, whereas a VUS of 1 corresponds to a perfect test.

Our estimator is motivated by a simple, yet interesting, representation of the ROC surface which is based on the notion of placement variable (Pepe & Cai, 2004). Specifically, note that if $F_1^{-1}(p_1) < F_3^{-1}(1 - p_3)$,

$$\begin{aligned}
\mathrm{ROCS}(p_1, p_3) &= F_2(F_3^{-1}(1-p_3)) - F_2(F_1^{-1}(p_1)) \qquad (1)\\
&= \Pr(Y_2 \leqslant F_3^{-1}(1-p_3)) - \Pr(Y_2 \leqslant F_1^{-1}(p_1))\\
&= \Pr(1 - F_3(Y_2) \geqslant p_3) - \Pr(F_1(Y_2) \leqslant p_1)\\
&= \Pr(Z \geqslant p_3) - \Pr(U \leqslant p_1),
\end{aligned}$$

where $U = F_1(Y_2)$ and $Z = 1 - F_3(Y_2)$. Thus, the ROC surface turns out to be the difference between the survival function of Z and the distribution function of U. The variable U is the proportion of class 1 subjects with test outcomes smaller than Y_2 and the variable Z is the proportion of class 3 subjects with test outcomes larger than Y_2. The variables U and Z quantify the degree of separation of the test outcomes in the three classes of patients. Specifically, U quantifies the degree of separation between the test outcomes in classes 1 and 2, whereas Z quantifies the degree of separation between classes 2 and 3. For instance, if the test outcomes in the three classes are highly separated, the placement of most class 2 subjects is at the upper tail of the class 1 distribution and at the lower tail of the class 3 distribution, so that most class 2 subjects will have large U and Z values. On the other hand, when the three distributions of test outcomes totally overlap, both U and Z will have an Uniform$(0, 1)$ distribution.

2.1. Bayesian bootstrap inference for the ROC surface

We start by recalling how the BB works in the one-population setting. For example, let $(Y_1, \ldots, Y_n)$ be a random sample from an unknown distribution F and suppose that the parameter of interest is F itself. In Efron's frequentist bootstrap (Efron, 1979) inference about F is obtained by repeatedly generating bootstrap samples, where each sample is drawn with replacement from the data. In the bth bootstrap replicate, $F^{(b)}$ is computed as

$$F^{(b)}(\cdot) = \sum_{i=1}^{n} \pi_i^{(b)} \delta_{Y_i}(\cdot), \qquad (2)$$

where $\pi_i^{(b)}$ is the proportion of times Y_i appears in the bth bootstrap sample, with $\pi_i^{(b)}$ taking on values in $\{0, 1/n, \ldots, n/n\}$. By opposition, in Rubin's BB (Rubin, 1981) the weights $\pi_i^{(b)}$ in (2) are generated from a

Dirichlet$(n; 1, \ldots, 1)$ distribution; note that in the BB the data should be regarded as fixed, so that we do not resample from it. The BB has connections with the Dirichlet Process (Ferguson, 1974); specifically, it can be regarded as a non-informative version of the Dirichlet Process (Gasparini, 1995, Theorem 2).

Now, let $(Y_{11}, \ldots, Y_{1n_1})$, $(Y_{21}, \ldots, Y_{2n_2})$, and $(Y_{31}, \ldots, Y_{3n_3})$ be random samples from classes 1, 2, and 3, respectively. The result in (1) provides the rationale for the following algorithm.

Bayesian bootstrap algorithm

For $b = 1, \ldots, B$:

Step 1. **Compute the placement variables based on the BB resampling**:
For $j = 1, \ldots, n_2$, compute

$$U_j^{(b)} = F_1^{(b)}(Y_{2j}) = \sum_{i=1}^{n_1} q_i^{(b)} I(Y_{1i} \leqslant Y_{2j}), \quad (q_1, \ldots, q_{n_1}) \sim \text{Dirichlet}(n_1; 1, \ldots, 1),$$

and

$$Z_j^{(b)} = 1 - F_3^{(b)}(Y_{2j}) = \sum_{l=1}^{n_3} r_l^{(b)} I(Y_{3l} \geqslant Y_{2j}), \quad (r_1, \ldots, r_{n_3}) \sim \text{Dirichlet}(n_3; 1, \ldots, 1).$$

Step 2. **Generate a random realization of the ROC surface**:
Based on (1) generate a realization of $\text{ROCS}^{(b)}(p_1, p_3)$, i.e.,

$$\text{ROCS}^{(b)}(p_1, p_3) = \sum_{j=1}^{n_2} v_j^{(b)} I(Z_j^{(b)} \geqslant p_3) - \sum_{j=1}^{n_2} w_j^{(b)} I(U_j^{(b)} \leqslant p_1),$$

$(v_1, \ldots, v_{n_2}) \sim \text{Dirichlet}(n_2; 1, \ldots, 1)$, and $(w_1, \ldots, w_{n_2}) \sim \text{Dirichlet}(n_2; 1, \ldots, 1)$. Compute the VUS associated to $\text{ROCS}^{(b)}(p_1, p_3)$, $\text{VUS}^{(b)}$, using numerical integration.

The BB estimate of the ROC surface, denoted as $\widehat{\text{ROCS}}(p_1, p_3)$, is then obtained by averaging the random realizations of the ROC surfaces, i.e,

$$\widehat{\text{ROCS}}(p_1, p_3) = \frac{1}{B} \sum_{b=1}^{B} \text{ROCS}^{(b)}(p_1, p_3).$$

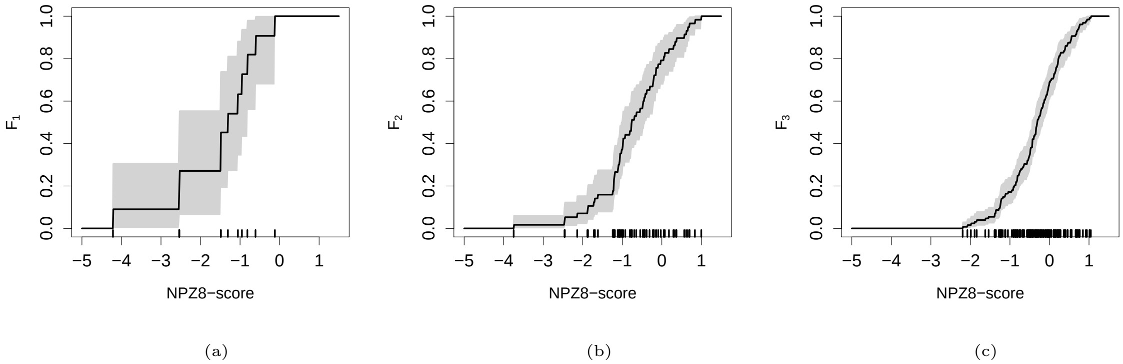
Similarly,

$$\widehat{\text{VUS}} = \frac{1}{B} \sum_{b=1}^{B} \text{VUS}^{(b)}.$$

A 95% probability interval for the VUS can be obtained from the percentiles of $\{\text{VUS}^{(1)}, \ldots, \text{VUS}^{(B)}\}$.

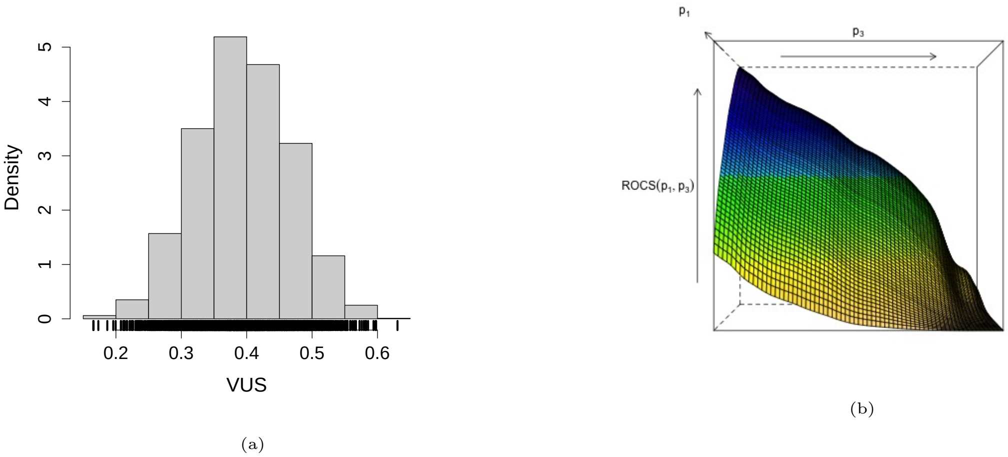
3. Application to HIV data

AIDS dementia complex (ADC; also known as HIV-dementia) is a common neurological disorder associated with HIV infection and typically occurs after years of HIV infection. Its essential features are disabling cognitive impairment accompanied by motor dysfunction, speech problems, and behavioral change. A patient with ADC can be in one of six stages: 0-normal, 0.5-subclinical, 1-mild, 2-moderate, 3-severe, 4-end stage. A possible clinical assessment of ADC is performed through a neuropsychological battery of eight age-normalized tests; the average index of the eight tests is termed NPZ-8 and has been validated as a reproducible measure of HIV-related neurological impairment (Selnes & Miller, 1994). The data analyzed here are from a study aiming to assess the accuracy of NPZ-8 in discriminating between patients exhibiting clinical symptoms of ADC (combined stages 1–3), subjects exhibiting minor neurological symptoms of ADC (ADC stage 0.5), and neurologically unimpaired individuals (ADC stage 0) (Nakas & Yiannoutsos, 2004). ADC patients in stages



(a) (b) (c)

Figure 1: BB estimate and 95% pointwise probability band of (a) F_1, (b) F_2, and (c) F_3, along with a rug representation of the raw data in each group.



(a)

(b)

Figure 2: (a) Histogram and rug representation of the 2000 BB sampled VUS. (b) BB estimate of the ROC surface.

1–3 are expected to have lower NPZ-8 scores than 0.5-stage subjects. They, in turn, are expected to have lower age-adjusted NPZ-8 scores compared to ADC stage 0 individuals. Complete data were available on 197 subjects; 129 neurologically normal, 57 with ADC stage 0.5 and 11 with ADC stages 1–3.

The BB estimates F_1, F_2, and F_3 are presented in Figure 1 along with their 95% pointwise probability bands. In Figure 2 (b), it is shown the BB estimate of the ROC surface which it is based on 2000 resamples, and where p_1 and p_3 being fine grids on $[0, 1]$. As can be observed, the produced ROC surface has the appealing feature of being smooth, thus allowing for useful interpretation of diagnostic performance at all thresholds. In Figure 2 (a) we also show an histogram of the 2000 BB sampled VUS and the BB estimate (95% probability interval) of the VUS is 0.393 $(0.248, 0.534)$, which lead us to conclude that the NPZ-8 score has a quite good accuracy in discriminating the aforementioned ADC stages.

References

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1–26.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615–629.

Gasparini, M. (1995). Exact multivariate Bayesian bootstrap distributions of moments. *Annals of Statistics*,

23, 762–768.

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

Inácio, V., Turkman, A. A., Nakas, C. T., & Alonzo, T. A. (2011). Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface. *Biometrical Journal*, **53**, 1011–1024.

Li, J., & Zhou, X. H. (2009). Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *Journal of Statistical Planning and Inference*, **139**, 4133–4142.

Nakas, C. T., & Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*, **23**, 3437–3449.

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

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics*, **9**, 130–134.

Selnes, O. A., & Miller, E. N. (1994). Development of a screening battery for HIV related cognitive impairment: the MACS experience. In *Neuropsychology of HIV infection. Current Research and Future Directions*, Grant, I., Martin, A. (Eds). Oxford UK: Oxford University Press, 176–187.

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2011). *Statistical Methods in Diagnostic Medicine.* 2nd Ed., New-York: Wiley.

Wan, S., & Zhang, B. (2013). Semiparametric ROC surface estimation for continuous diagnostic tests via polytomous logistic regression procedures. *Journal of Statistical Computation and Simulation*, **83**, 2195–2205.