Non-parametric estimation of ROC curve

Jiezhun Gu*

Department of Statistics, North Carolina State University, Raleigh, NC 27695

*email: jgu@unity.ncsu.edu

Subhashis Ghosal

Department of Statistics, North Carolina State University, Raleigh, NC 27695

Anindya Roy

Department of Mathematics and Statistics University of Maryland Baltimore County,

Baltimore, MD 21250

Summary

Receiver operating characteristic (ROC) curve is widely applied in measuring discriminatory ability of diagnostic or prognostic tests. This makes ROC analysis one of the most actively research areas in medical statistics. Many parametric and nonparametric estimation methods have been proposed for estimating the ROC curve and its functionals. In this paper, we introduce a non-parametric method based on the Bayesian bootstrap technique to estimate ROC curves for continuous diagnostic variables based on independent observations. The area under the ROC curve (AUC) is used to measure the accuracy of different diagnostic methods. The accuracy of the estimate of the ROC curve in the simulation studies is examined by the integrated absolute error (IAE). In comparison with other existing curve estimation methods, the Bayesian bootstrap method compares favorably in terms of accuracy, robustness and simplicity.

Keywords: Area under the curve (AUC); Bayesian bootstrap; Integrated absolute error; U-statistics.
1 Introduction

Since its introduction in the context of electronic signal detection (Green and Swets, 1966), Receiver Operating Characteristic (ROC) curve has become the method of choice for quantification of accuracy of medical diagnostics tests. ROC analysis provides a concrete way of evaluating accuracy of different diagnostics modalities relative to either a gold standard or other diagnostic tests. The primary reason for the popularity of ROC analysis in medical statistics is its mathematical framework under which the trade-off between true positive fractions and false positive fractions at different decision threshold for different diagnostic tests can be computed and compared. The ROC curve is a plot of the true positive fraction as a function of the false positive fraction (sensitivity versus one minus specificity) and is obtained by varying the threshold criterion for deciding between positive and negative diagnoses. The ROC plot thus displays full information at all of the system operating points (sensitivity and one minus specificity pairs). We assume that the diagnostic variables (X for the group without disease and Y for those with disease) have continuous distributions. Such an assumption seems more reasonable with the rapid growth of measuring capabilities of sophisticated diagnostic tools, especially in radiology reading studies. A particular benefit of the method is that it gives a performance assessment technique for the technology that is independent of practice-of-medicine issues. It is also invariant under monotone increasing transformation of the diagnostic variables, emphasizing the relationship between the distributions of the diagnostic variables rather than distributions themselves and providing decision threshold values based on the objective of the study. Moreover, the most commonly used index of the ROC curve, called the area under the curve (AUC), can be interpreted as the probability of Y greater than X, which can be easily understandable to a wider audience. Other ROC functionals of interest such as the partial AUC (pAUC) also have similar intu-
itive explanations. These features make the ROC analysis extremely popular in diagnostics research. In practice, the distributions of the diagnostics variables will be unknown. Hence an important component of the ROC analysis is accurate estimation of the ROC curve and its associated summary measures such as AUC and pAUC. Often the information regarding the distribution of the diagnostic variables is so limited that non-parametric estimation of ROC seems to be more natural.

The primary objective of this paper is to propose the use of the Bayesian bootstrap (BB) method (Rubin, 1981) for estimating and building confidence intervals for the different components of the ROC analysis. In this paper, we compare the accuracy of the proposed Bayesian bootstrap method with popular methods based on parametric models and semi-parametric models in the context of estimation of the AUC and other characteristics of the ROC curve. We also compare the accuracy of confidence intervals for the AUC produced by our method and three other existing methods. The BB estimator retains the flexibility and robustness of general non-parametric estimators, yet demonstrates high efficiency comparable to the parametric methods when the assumed parametric models are correct. We also detail the steps involved in the computation of the BB estimator.

The literature for ROC analysis is extensive. Hanley (1989) gave a comprehensive review on ROC methodology. Pepe (2003) covered a general framework of ROC in the senses of both clinical treatment and statistical theory. As a parametric approach, binormal model (Green and Swets, 1966) based on the signal detection theory is widely used involving two parameters, namely intercept and slope. There are several methods (Metz et al., 1998; Pepe, 2000; Zou and Hall, 2000) available to estimate the intercept and slope. Pepe (2003) and Zhou et al. (2002) provide excellent reviews of existing methods for bionormal ROC curve estimation in the context of continuous data. For ordinal data, maximum likelihood estimation was discussed in Ogilvie and Creelman (1968) and Dorfman and Alf (1969). Ordinal
regression model with the continuous covariates was studied by Tosteson and Begg (1988) focusing on the ROC curve rather than the indices of ROC. Direct regression modeling (Pepe, 1997) allows more general settings. Some semiparametric and nonparametric approaches have also been considered. DeLong et al. (1988) used a U-statistic to compare the AUC for correlated data. Lloyd (1998) presented a kernel estimate of the distribution of diagnostic variable. Ishwaran and Gatsonis (2000) proposed a hierarchical Bayesian method for a ROC model using the probit link. Other semiparametric methods are suggested by Pepe (2000), Alonzo and Pepe (1999), Zou and Hall (2000) and Cai and Moskowitz (2004). Zhou et al. (2005) studied the situation in the absence of a gold standard using non-parametric methods.

Our methodology is explained in Section 2. Results from a simulation study are displayed in Section 3 and real data analyses are given in Section 4. We end with a general discussion in Section 5.

2 Methodology

The purpose of using the BB method is to get a curve estimate as well as a confidence band for the ROC curve valid for a large class of functional forms of the ROC. The methodology is flexible enough and can be easily extended to compute estimates and to produce confidence intervals for ROC curve and other summary measures such as AUC, pAUC when the \((X, Y)\) are obtained as paired observations. In this paper we discuss the estimation and construction of confidence bands for ROC curves when the \(X\) data are independent of the \(Y\) data.

Let \(X\) and \(Y\) be two independent continuous variables, for instance, two diagnostic variables coming from two populations, one without disease and one with disease, respectively. By varying the decision threshold value \(z\) and plotting the true positive rate (sensitivity) versus the false positive rate (one minus specificity), the ROC curve is obtained: \(\{(P(X > z), \)}
\( P(Y > z) \) : \( X \sim F \) without disease, \( Y \sim G \) with disease, \( z \in \mathbb{R} \).

Mathematically, we can write the functional form of ROC curve as follows:

\[
R(t) = \bar{G}(\bar{F}^{-1}(t)), \quad 0 \leq t \leq 1, \tag{1}
\]

where \( \bar{F}(z) = P(X > z) \) and \( \bar{G}(z) = P(Y > z) \) are survival functions of \( X \) and \( Y \), respectively. A commonly used index to compare the accuracy of the modalities is the area under the curve (AUC) defined as \( A = \int_0^1 R(t) dt \) and an estimate of \( A \) is given by

\[
\hat{A} = \int_0^1 \hat{R}(t) dt, \tag{2}
\]

where \( \hat{R}(t) \) is an estimate of \( R(t) \) computed based on the data \( X_1, \ldots, X_m \) and \( Y_1, \ldots, Y_n \). Here the \( X_i \)'s and \( Y_j \)'s are independently and identically distributed as \( F \) and \( G \), respectively.

The accuracy of estimation for the entire ROC curve can be measured by the integrated absolute error (IAE) (Moise et al., 1988):

\[
IAE = \int_0^1 |\hat{R}(t) - R(t)| dt. \tag{3}
\]

Clearly \( |\hat{A} - A| \leq IAE \). To construct a uniform confidence band for ROC, it is advantageous to map the domain to the real line via a transformation \( \psi \), such as the logistic transformation \( \psi(x) = \log(x/(1 - x)) \), \( x \in (0, 1) \). The maximum possible estimation error in the \( \psi \)-scale is \( \epsilon(\psi, R, \hat{R}) = \sup\{|\psi(\hat{R}(t)) - \psi(R(t))| : t \in (0, 1)\} \). The width of a uniform 100(1 - \( \alpha \))% confidence band for the transformed ROC, denoted by \( d_\alpha(\psi, R) \), is given by:

\[
d_\alpha = d_\alpha(\psi, R) = 100(1 - \alpha)% \text{ percentile of the distribution of } \epsilon(\psi, R, \hat{R}). \tag{4}
\]
It follows that \( \psi^{-1}(\psi(\hat{R}_k(t)) - d_\alpha) \leq R(t) \leq \psi^{-1}(\psi(\hat{R}_k(t)) + d_\alpha) \). The transformation-retransformation automatically ensures that the confidence band lies within the unit area. In practice, \( d_\alpha \) has to be estimated from the data, usually by some resampling technique. In this paper, we only consider the logistic transformation.

### 2.1 Some existing methods

- Binormal model is one of the most popular models in ROC study. The binormal model follows from normality assumption on monotone transformation of the diagnostic variables. Specifically, the binormal model assumes that \( h(X) \sim \text{Normal}(u_D, \sigma_D^2) \) and \( h(Y) \sim \text{Normal}(u_D, \sigma_D^2) \) where \( h(x) \) is some monotone increasing function of \( x \). The binormal ROC curve is given by

\[
R(t) = \Phi(a + b\Phi^{-1}(t)),
\]

where \( a = (u_D - u_D)/\sigma_D, \ b = \sigma_D/\sigma_D \) and \( \Phi(x) \) is the cdf of the standard normal distribution. The corresponding AUC has a closed parametric form and is given by \( AUC = \Phi(a/\sqrt{1+b^2}) \). To estimate the parameters \( a \) and \( b \), the following two methods have been considered in the literature.

1. **Box-Cox Transformation method (BN-T).** Zou and Hall (2000) used Box-Cox transformation to estimate the monotone increasing transformation \( h(\cdot) \). The estimating procedure includes: (1) Transforming the observed values of the diagnostic variables to positive quantities by location change and exponentiation; (2) Apply Box-Cox transformation \( x(\lambda) = (x^\lambda - 1)/\lambda \) if \( \lambda > 0 \), and \( \log x \) if \( \lambda = 0 \) to obtain transformed observations \( X'_i(\lambda) \) and \( Y'_j(\lambda) \) where \( \lambda \) is the optimum power transformation parameter in the Box-Cox transformation. (3)
Assuming $X'_i(\lambda) \sim \text{iid } N(\mu, \sigma^2)$ and $Y'_j(\lambda) \sim \text{iid } N(v, \tau^2)$, get MLE of $\lambda$, $\mu$, $\sigma$, $v$ and $\tau$; (4) ROC curve estimate by Box-Cox transformation method is given by

$$\hat{R}(t) = \Phi((\hat{v} - \hat{\mu})/\hat{\tau} + \hat{\sigma}/\hat{\tau}\Phi^{-1}(t)).$$

2. **Generalized linear model (GLM) method (BN-G).** Using the relationship

$$E(1(Y > X)|\bar{F}(X) = t) = P(Y > X|X = \bar{F}^{-1}(t)) = \bar{G}(\bar{F}^{-1}(t)) = R(t)$$

Pepe (2000) suggested a GLM method for estimating the ROC curve. By conditioning on $t_i = \bar{F}(X_i)$, we have $E(U_{ji}) = \bar{R}(t_i)$, where $U_{ji} = 1(Y_j > X_i)$. Thus, the following probit regression model:

$$\Phi^{-1}(U_{ji}) = a + b \Phi^{-1}(\hat{t}_i), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n,$$

can be used for estimating $a$ and $b$, where $\hat{t}_i = \hat{\bar{F}}_m(x_i)$ are obtained by plugging in the empirical survival function of $X$. However, $U_{ji}$ are dependent and $\hat{t}_i$ are random; hence the applicability of the GLM estimation method is not fully justified in this context.

- **Semiparametric method (SP).** Semiparametric location-scale model (Pepe, 2003) assumes $X = u_D + \sigma_D\epsilon$ and $Y = u_D + \sigma_D\epsilon$ where $\mu$’s and $\sigma$’s are the location and scale parameters and $\epsilon$ has some unspecified survival function $S_0$ with mean 0 and variance 1. The functional form of the ROC is given by $R(t) = S_0(-a + bS_0^{-1}(t))$. The empirical survival function of the $\epsilon$ is then defined as

$$\hat{S}_0(y) = \frac{1}{m + n} \left\{ \sum_{i=1}^{m} 1 \left( \frac{X_i - \hat{u}_D}{\hat{\sigma}_D} > y \right) + \sum_{j=1}^{n} 1 \left( \frac{Y_j - \hat{u}_D}{\hat{\sigma}_D} > y \right) \right\},$$
where \( \hat{u}_D, \hat{\sigma}_D, \hat{\mu}_D, \hat{\sigma}_D \) are the sample means and the sample standard deviations of \( X \)'s and \( Y \)'s, respectively. By plugging in the empirical survival function \( \hat{S}_0 \) and estimates of location scale parameters into the expression for \( R(t) \), we get the semiparametric ROC estimate: \( \hat{R}(t) = \hat{S}_0(-\hat{a} + \hat{b}\hat{S}_0^{-1}(t)) \), where \( \hat{a} = (\hat{u}_D - \hat{\mu}_D)/\hat{\sigma}_D, \hat{b} = \hat{\sigma}_D/\hat{\sigma}_D \).

- **DeLong method (DL).** DeLong et al. (1988) used a U-statistics approach to estimate \( \theta = \text{AUC} \) by \( \hat{\theta} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \gamma(X_i, Y_j) \), where \( \gamma(x, y) = 1\{x < y\} + \frac{1}{2}1\{x = y\} \). They also provided consistent estimators of \( \text{Var}(\hat{\theta}) \) given by \( V_x^2/(m(m-1)) + V_y^2/(n(n-1)) \), where \( V_x(X_i) = \frac{1}{n} \sum_{j=1}^{n} \gamma(X_i, Y_j) - \hat{\theta}, V_y(Y_j) = \frac{1}{m} \sum_{i=1}^{m} \gamma(X_i, Y_j) - \hat{\theta} \) and \( V_x = \sum_{i=1}^{m} V_x^2(X_i), V_y = \sum_{j=1}^{n} V_y^2(Y_j) \).

In practice, the population distributions are never fully specified. The parametric models can often be very poor approximation of the true population model. Thus, nonparametric estimators of ROC curve seem appealing in many situations where limited or no information is available about the population distribution of the diagnostic variables. Empirical ROC estimators (cf. Pepe, 2003) are easily obtained by plugging in the empirical distribution functions of \( X \) and \( Y \) into the functional form of the ROC curve: \( \hat{R}_{m,n}(t) = \hat{G}_n(\hat{F}_m^{-1}(t)) \), where \( \hat{G}_n(t) \) and \( \hat{F}_m(t) \) are empirical survival functions of \( Y \) and \( X \), respectively. In order to have continuous estimators of the ROC curve the jumps in the empirical cdf can be interpolated linearly. However, this method can only produce an estimator of the ROC curve but not a confidence band for the curve. In order to get the error of the curve estimate, bootstrap method (Efron, 1979) can be used by drawing resamples with replacement from the given sample. Based on the bootstrap samples, estimates of ROC curve and the corresponding indices are obtained, as well as the errors in those estimates. In bootstrap realization, the ordinates are necessarily multiples of \( n^{-1} \) and the abscissas are necessarily multiples of \( m^{-1} \), which leads to some inherent discreteness. The BB method proposed below assigns
the Dirichlet distribution to the weights on the value of the ordinates and abscissas, and hence provides a smoother version of the bootstrap. Figure 1 gives an illustration of these differences. We now describe the BB estimator in detail.

2.2 The Bayesian bootstrap (BB) estimator

The Bayesian bootstrap estimator of the ROC curve and its associated summary measures can be computed by the following steps:

1. Step 1. (ROC curve as a cdf of auxiliary variables.) First we note that the ROC curve can be viewed as a cdf on the unit interval. Because $R(t) = \Pr(Y > F^{-1}(t)) = \Pr(F(Y) \leq t)$, where $X \sim F$ and $Y \sim G$ independently, so we have that $Z_j = F(Y_j)$, $j = 1, \ldots, n$, are independently and identically distributed with cdf $R$.

2. Step 2. (BB imputation of the auxiliary variables.) If $F$ were known, we could have obtained the values of $Z_j$’s. As $F$ is unknown, we can impute $F$ from its BB distribution based on the independent observations $X_1, \ldots, X_m$. That is, obtain a random realization of $\bar{F}$, denoted as $\bar{F}^*$, by

$$\bar{F}^*(u) = \sum_{i=1}^{m} p_i 1(X_i > u), \quad (6)$$

where $(p_1, \ldots, p_m) \sim Dir(m; 1, \ldots, 1)$ and are independent of other variables; here and below $Dir(m; 1, \ldots, m)$ stands for the $m$-dimensional Dirichlet distribution with each parameter 1, or equivalently the uniform distribution on the $m$ dimensional unit simplex. Put $Z_j = \bar{F}^*(Y_j)$. A convenient method for generating Dirichlet distributed
random vector \((p_1, \ldots, p_m)\) is to generate \(w_1, \ldots, w_m \sim \text{iid exp} \ (1)\) and put \(p_i = w_i / \sum_{j=1}^{m} w_j, i = 1, \ldots, m\).

3. Step 3. (Generating random realizations of the ROC curve.) By using the imputed conditionally independent samples of \(Z_j\)’s, a random realization of their cdf \(R\) may be obtained from the corresponding BB distribution. That is, a random realization of the ROC curve is given by

\[
R^*(t) = R_{m,n}^*(t) = \sum_{j=1}^{n} q_j 1(Z_j \leq t),
\]

where \((q_1, \ldots, q_n) \sim \text{Dir}(n; 1, \ldots, 1)\) and are independent of other variables.

4. Step 4. (BB estimator of the ROC curve.) After carrying out \(N\) many BB simulations, the BB estimate, denoted by \(\hat{R}_{m,n}^{BB}(t)\) is obtained by averaging the random realizations of \(R\):

\[
\hat{R}_{m,n}^{BB}(t) = \text{mean}(R_{m,n}^*(t)), \quad 0 \leq t \leq 1.
\]

By plugging in \(\hat{R}_{m,n}^{BB}(t)\) in (2), we obtain the BB estimate for the AUC \(A\).

In order to compute error estimates for the BB estimators of the ROC curve and associated indices, the above steps need to be repeated \(K\) times (where \(K\) is a reasonably large number). For example, the BB standard error of \(\hat{A}\) is given by

\[
s = \sqrt{\frac{1}{N-1} \sum_{l=1}^{N} (A^{(l)} - \hat{A})^2},
\]

where \(A^{(l)}\) is the randomly realized value of AUC based on the \(l^{th}\) set of BB samples. Also, \(100(1 - \alpha)\%\) BB credible interval for \(A\) can be obtained from the percentiles of \(\{A^{(l)}, l = 1, \ldots, N\}\). To obtain a credible band for \(R\) based on BB samples, we may estimate \(d_{\alpha}\) by
the $100(1-\alpha)$% percentile of the sample sup\{$|\psi(R^{(l)}(t)) - \psi(\hat{R}(t))| : t \in (0, 1), l = 1, \ldots, N$\}, and substitute $\hat{d}_\alpha$ in (4). In practice, however, in order to estimate $d_\alpha$, it is easier to draw fresh BB samples than to save every realization till the BB estimate has been computed.

Similar ideas may be used to estimate the IAE.

3 Simulation study

In this section, we compare accuracies of the estimates of ROC curve and the AUC functional obtained by the BB, BN-G, BN-T and SP methods. For BN-G, BN-T and SP methods, bootstrap is used to estimate standard errors or construct confidence intervals. Table 1 and 2 report the coverage probabilities and the mean lengths of the 90% confidence intervals for each method and parameter combinations. The DeLong method does not give an estimate of the entire ROC curve, and hence we use it only to compute AUC estimates in the real data examples given in Section 4.

Four different parametric models are used for the $(F,G)$ pair. The models are lognormal, location-scale exponential, gamma and beta and they are abbreviated as $A, B, C, D$, respectively, in Table 1 and Table 2. In each model, we fix the $X$ dataset with one combination of the parameters and vary the $Y$ datasets with 6 different parameter combinations. In the lognormal model, $X$ datasets are generated from the lognormal distribution with corresponding normal parameters $u_x = 0$ and $\sigma_x = 1$ while the $Y$ datasets are generated from the lognormal distribution with normal parameters $(u_y, \sigma_y) = (0.33, 0.33), (1, 0.33), (0.33, 1), (1, 1), (1, 3), (3, 3)$. In the location-scale exponential model, all $X$’s and $Y$’s are generated from the exponential distribution with exponential parameter $\alpha = 2$ (the density function is $f(x \mid \alpha) = \alpha^{-1}e^{-x/\alpha}$) and the location and scale parameters for the $X$ datasets are $u_x = 0$ and $\sigma_x = 1$ and those for the $Y$ datasets are chosen from $(u_y, \sigma_y) = (0.33, 0.33)$,
In the gamma model, $X$ datasets are generated from the gamma distribution with mean and standard error $u_x = 1$ and $\sigma_x = 1$, and the $Y$ datasets are generated from the gamma distribution with $(u_y, \sigma_y) = (1.33, 0.33), (2, 0.33), (1.33, 1), (2, 1), (2, 3), (5, 3)$. In the beta model, $X$ dataset are generated from the beta distribution with mean and standard error $u_x = 0.15$ and $\sigma_x = 0.15$, and $Y$ datasets are generated from the beta distribution with $(u_y, \sigma_y) = (0.20, 0.01), (0.20, 0.15), (0.40, 0.15), (0.20, 0.30), (0.40, 0.30), (0.50, 0.45)$. We choose the resample size $N = 1000$ and the number of independent simulations $K = 1000$. We consider 90% credible intervals in our simulation and the results are reported in Table 1. We also examine the IAE for some specific cases: lognormal $((u_x, \sigma_x) = (0, 1), (u_y, \sigma_y) = (1, 1))$, location-scale exponential $((u_x, \sigma_x) = (0, 1), (u_y, \sigma_y) = (1, 1))$, gamma $((u_x, \sigma_x) = (1, 1), (u_y, \sigma_y) = (2, 1))$, beta $((u_x, \sigma_x) = (0.15, 0.15), (u_y, \sigma_y) = (0.20, 0.30))$, the results are displayed in Figure 2.

From simulation results Table 1, Table 2 and Figure 2, we can observe that the proposed BB method performs favorably in view of accuracy and robustness. BN-G method gives considerable larger IAE in lognormal, location-scale exponential and gamma datasets. Moreover, in location-scale exponential dataset with $u_x = 0, \sigma_x = 1, u_y = 3, \sigma_y = 3$, the coverage of the true AUC abnormally decreases from 0.857 to 0.772 when the sample size increases from 15 to 50. The coverage of the true AUC from the BN-T method decreases from 0.867 to 0.774 when the sample size increases from 15 to 50 in beta datasets with $u_x = 0.15, \sigma_x = 0.15, u_y = 0.20, \sigma_y = 0.30$. Similar pattern occurs in beta datasets with $u_x = 0.15, \sigma_x = 0.15, u_y = 0.50, \sigma_y = 0.45$. Moreover, the mean length of the 90% CI of AUC increases from 0.443 to 0.482 when the sample size increases from 15 to 50 in beta datasets with $u_x = 0.15, \sigma_x = 0.15, u_y = 0.40, \sigma_y = 0.15$. BN-T method gives the highest mean length of the 90% CI of AUC in the six datasets. Thus, the BN-T method is quite unstable for models which can not be reduced to normal models via monotone transformation. The SP
method behaves unusually in that the coverage probability decreases from 0.820 to 0.150 in the lognormal dataset with $u_x = 0, \sigma_x = 1, u_y = 3, \sigma_y = 3$ and also decreases from 0.834 to 0.662 in the beta dataset with $u_x = 0.15, \sigma_x = 0.15, u_y = 0.50, \sigma_y = 0.45$. In general, the coverage probability increases a little bit when the sample size increases from 15 to 50, while the the mean lengths of 90% CI of AUC, IAE and their variations decrease significantly.

[Insert Table 1 here]

[Insert Table 2 here]

[Insert Figure 2 here]

4 Real data analyses

We shall illustrate the BB method for AUC estimation using two well studied data and compare with the estimates obtained by the DL method.

The first data is from a study to evaluate the discriminatory ability of three preoperative measured indices with respect to the benefit from the surgery (DeLong, et al., 1988). These correlated indices were used as Krebs-Goplerud score (K-G), Albumin (ALB), total protein (TP) on 49 ovarian cancer patients who were treated for correction of intestinal obstruction. The question of interest is to determine which index best discriminates the benefit from a surgery. Estimates were computed and compared with DL’s seen Table 3 upper part. By using pairwise deletion of unobserved data, the results using DL method are slightly different from those in the original paper (DL, et al., 1988). We see the estimated variance of the estimate of AUC is smaller for the BB method for ALB and TP which are continuous variables, while DL’s method gives smaller estimated variance for K-G, which is discrete. The estimated ROC curves obtained by the BB method are displayed left in Figure 3 for each of the three variables ALB, TP and K-G.
Another data set was published by Wieand et al. (1989). This study was based on 51 patients as control group diagnosed as pancreatitis and 90 patients as cases group diagnosed as pancreatic cancer by two biomarkers, which were a cancer antigen (CA 125) and a carbohydrate antigen (CA 19-9). The purpose was to decide which marker would better distinguish case group from control group. Using the DL method and the BB method, the AUC is estimated. The two estimates have the same values, but the estimated variance for the BB method is slightly smaller (Table 3 lower part). By examining the BB estimate of ROC curve (right in Figure 3 ), we can conclude that roughly speaking, the marker CA 19-9 has better discriminatory ability especially on the range (0, 0.2) of false positive rate.

5 Discussion

We have proposed an estimation method for the ROC curve and its functionals based on the Bayesian bootstrap technique. One of the appealing features of our methodology is that it readily produces standard errors, confidence intervals and confidence bands for the entire ROC curve as well as some associated summary measures. The BB method closely resembles a non-parametric Bayesian analysis using Dirichlet process priors, but substantial simplification is obtained by avoiding the inversion of $\tilde{F}$ through the use of the pseudo-samples $Z_j$’s in (7). This simplification is possible because the BB corresponds to a “non-informative” posterior for which the prior base measure in a Dirichlet process has been chosen to be the null measure, and generating samples from the posterior reduces to finite dimensional random variate generations; see Rubin (1981) for details.

The Bayesian interpretation of our approach means that the method is not based on
large sample techniques, and in principle applies to any sample size. We have compared our method with other methods available in the literature by means of a simulation study. Considering about the running time of the simulations, BN-T method takes extremely longer time to run, while BB method runs so enjoyably that you may even ignore the running time. SP method is fast, but slower than BB method. BN-G method is also very slow. From the simulation results, the BN-G method gives rise to large IAE for the estimate of the ROC curve and the BN-T method seems to be very unstable as the estimate depends a lot on the initial guess of the Box-Cox index and the actual family of distribution of the diagnostic variables. All three methods (BN-G, BN-T, and SP) suffer from extremely low coverage probability for some data generating mechanism or other. In contrast, the BB method gives rise to a conceptually simple, completely nonparametric and easily implementable method with remarkable accuracy and robust coverage probability. It is straightforward to extend the BB method to the multivariate situation to estimate the ROC function componentwise for each pair of components of two random vectors. The methodology immediately extends to paired observations \( (X_i, Y_i), i = 1, \ldots, m(= n) \) by choosing the BB weights \((q_1, \ldots, q_m)\) to be same as the weights \((p_1, \ldots, p_m)\). Moreover, the methodology can produce estimates and confidence bands for difference of two ROC curves arising from independent experiments or dependent experiments. These and other issues are being considered for future research.

If desired, the BB method can be smoothed out using a kernel and a bandwidth. The smoothed version of the BB survival function is defined as:

\[
\tilde{F}(u) = \sum_{i=1}^{m} p_i \bar{\Phi}(\frac{u - X_i}{h_m}); \quad Z_j = \tilde{F}_1(Y_j), j = 1, \ldots, n; \quad R_{m,n}^*(t) = \sum_{j=1}^{n} q_j \Phi(\frac{t - Z_j}{h_m}) \tag{10}
\]

where \( m \) is the sample size of \( X, (p_1, \ldots, p_m) \sim Dir(m; 1, \ldots, 1) \) independently of \( X, h_m \rightarrow 0, \bar{\Phi}(x) \) is the survival function of standard normal distribution. Because we are primarily
interested in the cumulative cdf, smaller bandwidth such as \( h_m \sim m^{-1/3} \) may be considered rather than the more commonly used \( h_m \sim m^{-1/5} \) for estimating densities. The smoothed out method can then be carried out in a similar fashion. From a limited simulation study not shown here, we conclude that the smoothed BB produces almost identical results. Thus the extra computation in equation (10) compared to equation (6) and subsequently at later steps cannot be strongly justified.

Acknowledgements

Research of the first two authors is partially supported by NSF grant number DMS-0349111.

References


Figure 1: Comparison of Empirical and the BB’s estimates of ROC with the True (Simulation dataset: $X \sim \text{iid } \mathcal{N}(1,1)$, $Y \sim \text{iid } \mathcal{N}(1.5,1)$, $m = n = 50$, $N = 1000$)

Table 3: ROC AUC estimates

<table>
<thead>
<tr>
<th>Indices</th>
<th>DeLong method / variance</th>
<th>BB method / variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALB</td>
<td>0.737/0.0148</td>
<td>0.745/0.0075</td>
</tr>
<tr>
<td>TP</td>
<td>0.648/0.0152</td>
<td>0.661/0.0089</td>
</tr>
<tr>
<td>K-G</td>
<td>0.726/0.0066</td>
<td>0.793/0.0080</td>
</tr>
<tr>
<td>CA 19-9</td>
<td>0.861/0.00094</td>
<td>0.861/0.00088</td>
</tr>
<tr>
<td>CA 125</td>
<td>0.706/0.00219</td>
<td>0.706/0.00197</td>
</tr>
</tbody>
</table>
Table 1: Simulation: Coverage of AUC and related mean lengths of the 90% CI in lognormal, location-scale exponential datasets.

<table>
<thead>
<tr>
<th>Data</th>
<th>$n = 15$</th>
<th></th>
<th></th>
<th></th>
<th>$n = 50$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BB</td>
<td>BN-G</td>
<td>BN-T</td>
<td>SP</td>
<td>BB</td>
<td>BN-G</td>
<td>BN-T</td>
<td>SP</td>
</tr>
<tr>
<td>$A: u_y=0.33$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_y=(0.33)$</td>
<td>0.881</td>
<td>0.876</td>
<td>0.894</td>
<td>0.842</td>
<td>0.892</td>
<td>0.899</td>
<td>0.933</td>
<td>0.833</td>
</tr>
<tr>
<td>1.00</td>
<td>(0.346)</td>
<td>(0.327)</td>
<td>(0.357)</td>
<td>(0.358)</td>
<td>(0.197)</td>
<td>(0.180)</td>
<td>(0.188)</td>
<td>(0.194)</td>
</tr>
<tr>
<td>0.33</td>
<td>(0.241)</td>
<td>(0.230)</td>
<td>(0.209)</td>
<td>(0.256)</td>
<td>(0.145)</td>
<td>(0.135)</td>
<td>(0.128)</td>
<td>(0.170)</td>
</tr>
<tr>
<td>1.00</td>
<td>(0.322)</td>
<td>(0.310)</td>
<td>(0.331)</td>
<td>(0.339)</td>
<td>(0.182)</td>
<td>(0.169)</td>
<td>(0.179)</td>
<td>(0.201)</td>
</tr>
<tr>
<td>0.33</td>
<td>0.899</td>
<td>0.886</td>
<td>0.866</td>
<td>0.923</td>
<td>0.893</td>
<td>0.871</td>
<td>0.882</td>
<td>0.900</td>
</tr>
<tr>
<td>1.00</td>
<td>(0.262)</td>
<td>(0.254)</td>
<td>(0.250)</td>
<td>(0.276)</td>
<td>(0.149)</td>
<td>(0.141)</td>
<td>(0.143)</td>
<td>(0.176)</td>
</tr>
<tr>
<td>0.33</td>
<td>0.901</td>
<td>0.899</td>
<td>0.883</td>
<td>0.954</td>
<td>0.900</td>
<td>0.898</td>
<td>0.905</td>
<td>0.695</td>
</tr>
<tr>
<td>1.00</td>
<td>(0.322)</td>
<td>(0.323)</td>
<td>(0.336)</td>
<td>(0.314)</td>
<td>(0.182)</td>
<td>(0.178)</td>
<td>(0.176)</td>
<td>(0.178)</td>
</tr>
<tr>
<td>3.00</td>
<td>0.861</td>
<td>0.859</td>
<td>0.801</td>
<td>0.820</td>
<td>0.886</td>
<td>0.890</td>
<td>0.873</td>
<td>0.150</td>
</tr>
<tr>
<td>1.00</td>
<td>(0.230)</td>
<td>(0.233)</td>
<td>(0.205)</td>
<td>(0.275)</td>
<td>(0.136)</td>
<td>(0.134)</td>
<td>(0.122)</td>
<td>(0.162)</td>
</tr>
<tr>
<td>$B: u_y=0.33$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_y=(0.33)$</td>
<td>0.885</td>
<td>0.887</td>
<td>0.889</td>
<td>0.883</td>
<td>0.897</td>
<td>0.897</td>
<td>0.903</td>
<td>0.903</td>
</tr>
<tr>
<td>1.00</td>
<td>(0.331)</td>
<td>(0.320)</td>
<td>(0.358)</td>
<td>(0.343)</td>
<td>(0.189)</td>
<td>(0.173)</td>
<td>(0.206)</td>
<td>(0.199)</td>
</tr>
<tr>
<td>0.33</td>
<td>0.906</td>
<td>0.895</td>
<td>0.893</td>
<td>0.916</td>
<td>0.903</td>
<td>0.901</td>
<td>0.865</td>
<td>0.907</td>
</tr>
<tr>
<td>1.00</td>
<td>(0.331)</td>
<td>(0.316)</td>
<td>(0.339)</td>
<td>(0.345)</td>
<td>(0.185)</td>
<td>(0.171)</td>
<td>(0.181)</td>
<td>(0.192)</td>
</tr>
<tr>
<td>1.00</td>
<td>(0.305)</td>
<td>(0.288)</td>
<td>(0.285)</td>
<td>(0.304)</td>
<td>(0.172)</td>
<td>(0.159)</td>
<td>(0.158)</td>
<td>(0.167)</td>
</tr>
<tr>
<td>3.00</td>
<td>0.884</td>
<td>0.886</td>
<td>0.852</td>
<td>0.917</td>
<td>0.912</td>
<td>0.878</td>
<td>0.877</td>
<td>0.912</td>
</tr>
<tr>
<td>1.00</td>
<td>(0.200)</td>
<td>(0.195)</td>
<td>(0.167)</td>
<td>(0.207)</td>
<td>(0.113)</td>
<td>(0.109)</td>
<td>(0.098)</td>
<td>(0.131)</td>
</tr>
<tr>
<td>3.00</td>
<td>0.854</td>
<td>0.857</td>
<td>0.910</td>
<td>0.930</td>
<td>0.902</td>
<td>0.772</td>
<td>0.925</td>
<td>0.919</td>
</tr>
<tr>
<td>(3.00)</td>
<td>(0.098)</td>
<td>(0.097)</td>
<td>(0.076)</td>
<td>(0.110)</td>
<td>(0.057)</td>
<td>(0.062)</td>
<td>(0.045)</td>
<td>(0.065)</td>
</tr>
</tbody>
</table>
Table 2: Simulation: Coverage of AUC and related mean lengths of the 90% CI in gamma and beta datasets.

<table>
<thead>
<tr>
<th>Data</th>
<th>BB</th>
<th>BN-G</th>
<th>BN-T</th>
<th>SP</th>
<th>BB</th>
<th>BN-G</th>
<th>BN-T</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1.33</td>
<td>0.867</td>
<td>0.899</td>
<td>0.864</td>
<td>0.888</td>
<td>0.883</td>
<td>0.939</td>
<td>0.871</td>
</tr>
<tr>
<td>n</td>
<td>15</td>
<td>(0.312)</td>
<td>(0.316)</td>
<td>(0.329)</td>
<td>(0.182)</td>
<td>(0.167)</td>
<td>(0.217)</td>
<td>(0.194)</td>
</tr>
<tr>
<td>σ &lt;sub&gt;y&lt;/sub&gt; = (0.33)</td>
<td>2.00</td>
<td>0.865</td>
<td>0.836</td>
<td>0.868</td>
<td>0.847</td>
<td>0.915</td>
<td>0.813</td>
<td>0.956</td>
</tr>
<tr>
<td></td>
<td>(0.223)</td>
<td>0.209</td>
<td>0.199</td>
<td>0.224</td>
<td>(0.138)</td>
<td>(0.131)</td>
<td>(0.156)</td>
<td>(0.151)</td>
</tr>
<tr>
<td></td>
<td>1.33</td>
<td>0.918</td>
<td>0.886</td>
<td>0.900</td>
<td>0.911</td>
<td>0.898</td>
<td>0.901</td>
<td>0.883</td>
</tr>
<tr>
<td></td>
<td>(1.00)</td>
<td>0.320</td>
<td>0.306</td>
<td>0.328</td>
<td>0.329</td>
<td>(0.179)</td>
<td>(0.166)</td>
<td>(0.181)</td>
</tr>
<tr>
<td></td>
<td>2.00</td>
<td>0.886</td>
<td>0.876</td>
<td>0.840</td>
<td>0.873</td>
<td>0.886</td>
<td>0.851</td>
<td>0.860</td>
</tr>
<tr>
<td></td>
<td>(1.00)</td>
<td>0.244</td>
<td>0.234</td>
<td>0.223</td>
<td>0.244</td>
<td>(0.140)</td>
<td>(0.132)</td>
<td>(0.133)</td>
</tr>
<tr>
<td></td>
<td>2.00</td>
<td>0.896</td>
<td>0.898</td>
<td>0.880</td>
<td>0.897</td>
<td>0.890</td>
<td>0.896</td>
<td>0.878</td>
</tr>
<tr>
<td></td>
<td>(3.00)</td>
<td>0.327</td>
<td>0.325</td>
<td>0.349</td>
<td>0.334</td>
<td>(0.184)</td>
<td>(0.177)</td>
<td>(0.187)</td>
</tr>
<tr>
<td></td>
<td>5.00</td>
<td>0.852</td>
<td>0.861</td>
<td>0.858</td>
<td>0.882</td>
<td>0.906</td>
<td>0.837</td>
<td>0.886</td>
</tr>
<tr>
<td></td>
<td>(3.00)</td>
<td>(0.101)</td>
<td>(0.105)</td>
<td>(0.084)</td>
<td>(0.108)</td>
<td>(0.059)</td>
<td>(0.061)</td>
<td>(0.051)</td>
</tr>
<tr>
<td>D</td>
<td>0.20</td>
<td>0.868</td>
<td>0.859</td>
<td>0.897</td>
<td>0.884</td>
<td>0.896</td>
<td>0.900</td>
<td>0.872</td>
</tr>
<tr>
<td>σ &lt;sub&gt;y&lt;/sub&gt; = (0.01)</td>
<td>(0.336)</td>
<td>(0.323)</td>
<td>(0.507)</td>
<td>(0.367)</td>
<td>(0.203)</td>
<td>(0.185)</td>
<td>(0.472)</td>
<td>(0.214)</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.888</td>
<td>0.884</td>
<td>0.928</td>
<td>0.883</td>
<td>0.907</td>
<td>0.894</td>
<td>0.940</td>
</tr>
<tr>
<td></td>
<td>(0.321)</td>
<td>(0.306)</td>
<td>(0.411)</td>
<td>(0.333)</td>
<td>(0.181)</td>
<td>(0.167)</td>
<td>(0.298)</td>
<td>(0.181)</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.873</td>
<td>0.866</td>
<td>0.968</td>
<td>0.868</td>
<td>0.917</td>
<td>0.799</td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>(0.177)</td>
<td>(0.170)</td>
<td>(0.443)</td>
<td>(0.173)</td>
<td>(0.102)</td>
<td>(0.101)</td>
<td>(0.482)</td>
<td>(0.109)</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.891</td>
<td>0.878</td>
<td>0.867</td>
<td>0.865</td>
<td>0.898</td>
<td>0.892</td>
<td>0.774</td>
</tr>
<tr>
<td></td>
<td>(0.321)</td>
<td>(0.323)</td>
<td>(0.336)</td>
<td>(0.348)</td>
<td>(0.182)</td>
<td>(0.177)</td>
<td>(0.212)</td>
<td>(0.184)</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.876</td>
<td>0.873</td>
<td>0.892</td>
<td>0.878</td>
<td>0.906</td>
<td>0.895</td>
<td>0.893</td>
</tr>
<tr>
<td></td>
<td>(0.269)</td>
<td>(0.262)</td>
<td>(0.299)</td>
<td>(0.273)</td>
<td>(0.152)</td>
<td>(0.145)</td>
<td>(0.166)</td>
<td>(0.159)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.896</td>
<td>0.876</td>
<td>0.874</td>
<td>0.834</td>
<td>0.886</td>
<td>0.912</td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td>(0.333)</td>
<td>(0.332)</td>
<td>(0.440)</td>
<td>(0.341)</td>
<td>(0.189)</td>
<td>(0.187)</td>
<td>(0.266)</td>
<td>(0.211)</td>
</tr>
</tbody>
</table>
Figure 2: From left to right and from top to bottom, the boxplots of IAE by the lognormal datasets, location-scale exponential datasets, gamma and beta datasets, with sample size $A$ stands for $n = 15$, $B$ stands for $n = 50$; Index 1–BB method, 2–BN-G, 3–BN-T, 4–SP.
Figure 3: From left to right, BB estimates of ROC curves for three indices (K-G, ALB, TP); BB estimates of ROC curves for diagnostic markers CA 19-9 and CA 125.