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

Institute of Statistics Mimeo Series # 2592

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 active research areas in medical statistics. Many parametric and semiparametric estimation methods have been proposed for estimating the ROC curve and its functionals. In this paper, we propose a fully nonparametric Bayesian bootstrap (BB) estimation method for ROC curve and its functionals. The BB method gives a bandwidth free automatically smooth estimate. The area under the 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, BB method performs well in terms of accuracy, robustness and simplicity. We also propose a procedure based on the BB approach to test the binormality assumption.

Keywords: Area under the curve (AUC); Bayesian bootstrap; Integrated absolute error; ROC curve; Testing binormality.

1 Introduction

Since its introduction in the context of electronic signal detection (Green and Swets [1]), Receiver Operating Characteristic (ROC) curve has become the method of choice for quantification of accuracy of medical diagnostics tests. The diagnostic variables $X \sim F$ for the group without disease and $Y \sim G$ for those with disease are well defined. The ROC curve is a plot of the true positive fraction (TPF) as a function of the false positive fraction (FPF), or sensitivity versus one minus specificity, and is obtained by varying the threshold criterion for deciding between positive and negative diagnosis. Some features such as invariance property and interpretation of the area under the curve (AUC) as $\Pr(Y > X)$ make the ROC analysis extremely popular in diagnostics research.

The literature for ROC analysis for continuous diagnostic variables based on independent observations is extensive. Semiparametric methods for ROC analysis have particularly been popular since the presence of nonparametric components make these models considerably flexible, yet they can incorporate specific parametric features. Under the semiparametric framework, there are several different approaches: The simplest one is binormal (Green and Swets [1]) model, which assumes the binormality of the diagnostic test variables after some monotone increasing transformation. The intercept and slope in the binormal model can be estimated by several methods, such as by Hsieh and Turnbull [2], Metz et al. [3], Zou and Hall [4], Pepe [5], [6], Cai and Moskowitz [7], among others. However, see Goddard and Hinberg [8] for some examples where binormality fails. Similarly, bi-gamma (Dorfman et al. [9]) and bi-beta (Zou et al. [10]) models have also been considered. Li et al. [11] proposed a model where F and G are nonparametrically and parametrically specified. Qin and Zhang [12] modeled the functional form of the likelihood ratio to estimate the parameters. Normal mixture model was studied by Hall and Zhou [13]. Among completely nonparametric

methods, kernel estimate of F and G was discussed by Zou et al. [14], Lloyd [15], among others. Because AUC is an important index for ROC curve, its estimation method was well discussed by Bamber [16], Brownie et al. [17], DeLong et al. [18], Qin and Zhou [19], among others.

Within the nonparametric framework, the empirical estimate of ROC was studied by (Hsieh and Turnbull [2]), along with its asymptotic property. Li et al. [20] obtained the weak convergence theory for the ROC estimator under censoring by plugging in the product-limit estimators. Motivated by the empirical counterpart of the ROC curve, based on the Bayesian bootstrap (Rubin [21]) resampling distributions, we propose a completely nonparametric modeling for estimating and building credible intervals for the ROC curves. The BB method leads to smooth estimates without requiring to choose a bandwidth. Our simulations show that these bands and intervals have approximate frequentist validity even for very small sample sizes. This phenomenon is theoretically explained by strong approximation theory (Gu and Ghosal [22]). In comparison with other existing curve estimation methods, our BB method performs well in terms of accuracy, robustness, simplicity and smoothness. We also propose a procedure to test the binormality assumption as an application of our BB method.

Our methodology is explained in Section 2. Testing binormality procedure is presented in Section 3. Results from simulation studies are displayed in Section 4 and real data analyses are given in Section 5.

2 Methodology

The purpose of using the BB method is to get a curve estimate as well as a credible band for the ROC curve valid for all pairs of distribution functions.

Let $X \sim F$ and $Y \sim G$ 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 c_t (if $X > c_t$ or $Y > c_t$, false or true positive event occurs) and plotting the true positive fraction (sensitivity) versus the false positive fraction (one minus specificity), the ROC curve is obtained: $\{(\mathrm{P}(X > c_t), \mathrm{P}(Y > c_t))\} = \{(t, R(t))\}$, where $c_t \in \mathbb{R},\ t = P(X > c_t)$ is called the false positive fraction. Mathematically, we can write the functional form of ROC curve (Pepe [6], page 106) as follows:

$$R(t) = P(Y > c_t) = P(Y > \bar{F}^{-1}(t)) = \bar{G}(\bar{F}^{-1}(t)) = P(\bar{F}(Y) \leq t) \tag{1}$$

where $\bar{F}(u) = P(X > u)$ and $\bar{G}(u) = P(Y > u)$ 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) A and its estimate $\hat{A}$ are defined as

$$A = \int_0^1 R(t)dt \qquad \text{and} \qquad \hat{A} = \int_0^1 \hat{R}(t)dt, \tag{2}$$

where $\hat{R}(t)$ is some estimate of $R(t)$.

The accuracy of estimation for the entire ROC curve can be measured by the integrated absolute error (IAE) (Moise et al. [23]): IAE= $\int_0^1 |\hat{R}(t) - R(t)|dt$. Clearly $|\hat{A} - A| \leq IAE$. To construct a uniform credible band for ROC, it is advantageous to map the domain to the real line via a transformation ψ, such as the logistic transformation $\psi(x) = \log(x/(1-x)),\ x \in (0,1)$. The maximum possible estimation error in the ψ-scale is $\epsilon(\psi, R, \hat{R}) = \sup\{|\psi(\hat{R}(t)) - \psi(R(t))| : t \in (0,1)\}$. The width of a uniform $100(1-\alpha)\%$ credible band for the transformed ROC, denoted by $d_\alpha(\psi, R)$, is given by:

$$d_\alpha = d_\alpha(\psi, R) = 100(1-\alpha)\% \text{ percentile of the distribution of } \epsilon(\psi, R, \hat{R}). \tag{3}$$

Thus the uniform $100(1-\alpha)\%$ credible band for the ROC curve can be constructed by

$$\psi^{-1}(\psi(\hat{R}(t)) - d_\alpha) \leq R(t) \leq \psi^{-1}(\psi(\hat{R}(t)) + d_\alpha). \tag{4}$$

The transformation-retransformation automatically ensures that the credible band lies within the unit square. In practice, d_α has to be estimated, usually by some resampling technique. If the uniform credible band for ROC on $t \in (0,1)$ is too wide, other alternatives such as pointwise $100(1-\alpha)\%$ credible band or uniform $100(1-\alpha)\%$ (or less) credible band restricted on a small subinterval of interest should be considered instead.

The motivation of our BB estimator is twofold as shown below. On the one hand, we can view this as a bandwidth free smoothing of the empirical estimate. On the other hand, we argue that it is a non-informative limit of a Bayesian estimate based on the Dirichlet process prior.

1. Empirical ROC estimators (Hsieh and Turnbull [2]) are easily obtained by plugging the empirical counterparts into the ROC functional form. In order to have continuous estimators of the ROC curve, the jumps in the empirical CDF can be interpolated linearly. Bootstrap method (Efron [24]) can be used to get the error of the curve estimate. However, inherent discreteness of the estimate is partially due to finite choice of the weights. The BB method proposed below assigns the Dirichlet distribution to the weights. By forming an ensemble of estimators and averaging, it provides a smoother version of the bootstrap. Figure 1 gives an illustration of these differences even when sample size is small.

   [Insert Figure 1 here]

2. To implement a Bayesian analysis, a natural choice of priors on F and G is Dirichlet process denoted as DP with certain pairs of precision M and center measure ξ, say $F \sim$

$\mathrm{DP}(M_1, \xi_1)$, $G \sim \mathrm{DP}(M_2, \xi_2)$. Conditional on data $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, the posterior of $(F|\text{data})$ is $\mathrm{DP}(M_1+m, \frac{M_1\xi_1+m\mathbb{F}_m}{M_1+m})$, $(R|F, G,\text{data})$ is $\mathrm{DP}(M_2+n, \frac{M_2\xi_2\circ\bar{F}^{-1}+n\mathbb{G}_n\circ\bar{F}^{-1}}{M_2+n})$. Unfortunately, the above posterior can be obtained only by the Sethuraman [25] representation of the Dirichlet process and function inversion. The Sethuraman representation involves generating an infinite collection of random variables which is computationally very intensive. We consider the non-informative limit of the Dirichlet processes by letting $M_1 \to 0$ and $M_2 \to 0$. This simplifies the procedure to an infinitely easier simulation problem. In fact, we do not even have to specify the center measures ξ_1 and ξ_2. We need only to generate from the uniform distribution over the simplex which can be done quite easily; see Remark (1). Our BB estimator is just based on this simplification, i.e., the posterior of $(F|\text{data})$ is $\mathrm{DP}(m, \mathbb{F}_m)$, $(R|F, G, \text{data})$ is DP $(n, \mathbb{G}_n \circ \bar{F}^{-1})$.

The BB estimator of the ROC curve and its associated summary measures can be computed as follows: Recall $R(t)= \Pr(\bar{F}(Y) \leq t)$, $t \in \mathcal{D} \subset [0, 1]$, where $\mathcal{D}$ denotes a prespecified set of FPF of interest. If we can impute the variable $Z = \bar{F}(Y)$ by plugging in the survival distribution of F based on the BB resampling distribution, then CDF of Z based on the BB resampling distribution is one realization of the ROC curve from the corresponding posterior distribution.

1. Step 1. (Imputing the placement variables based on the BB resampling distribution.) Let $Z_j = \bar{F}^{\#}(Y_j) = 1 - F^{\#}(Y_j)$, $F^{\#}(Y_j) = \sum_{i=1}^{m} p_i 1(X_i \leq u)$, $(p_1, \ldots, p_m) \sim$ Dirichlet $(m; 1, \ldots, 1)$ independent of others, equivalently to generate $\bar{F}^{\#} \sim$ $\mathrm{DP}(m, \mathbb{F}_m)$ evaluated at Y_j's as Z_j (Z_j is also called non-disease placement value (Pepe [6], page 105) evaluated at Y_j). The difference between ours and Pepe's lies in that we choose the survival function using the BB resampling distribution instead of the empirical one.

2. Step 2. (Generating one random realization of the ROC curve.) Generate one realization of $R^{\#}_{m,n}(t)$, i.e., CDF of $Z_1, ..., Z_n$, where $R^{\#}_{m,n}(t) = \sum_{j=1}^{n} q_j 1(Z_j \leq t)$, $(q_1, \ldots, q_n) \sim$ Dirichlet $(n; 1, \ldots, 1)$ independent of others, equivalent to generate $R^{\#}_{m,n}(t) \sim$ DP $(n, \mathbb{G}_n \circ \bar{F}^{-1})$ evaluated at $t \in \mathcal{D}$. Corresponding random realization of AUC is denoted as $A^{\#}$, plus some subscript as the index of the realization.

3. Step 3. (Averaging the ensemble of random ROC curves.) The BB estimate, denoted as $\hat{R}^{BB}_{m,n}(t)$, is obtained by averaging the random realizations of the ROC curves, i.e., $\hat{R}^{BB}_{m,n}(t) = \text{mean}(R^{\#}_{m,n}(t))$, $t \in \mathcal{D}$. Similarly, we obtain the BB estimate of AUC denoted as A^{BB} by plugging $\hat{R}^{BB}_{m,n}(t)$ into (2).

Because of two levels of random variations and averaging over them, the BB estimate is much smoother than the empirical one. Note that we do not need a kernel to smooth it out.

Remark (1): A convenient method for generating $(p_1, \ldots, p_m) \sim$ Dirichlet $(m; 1, \ldots, 1)$ is to generate $w_1, \ldots, w_m \sim$ i.i.d. exponential distribution with rate 1 and put $p_i = w_i / \sum_{j=1}^{m} w_j, i = 1, \ldots, m$.

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 A^{BB} is given by

$$s = \sqrt{\frac{1}{K-1} \sum_{l=1}^{K} (A^{\#}_l - A^{BB})^2}, \tag{5}$$

Also $100(1-\alpha)\%$ BB credible interval for A can be obtained from

$$\text{the percentiles of } \{A^{\#}_l,\ l = 1, \ldots, N\} \text{ at level } \alpha. \tag{6}$$

To obtain a uniform credible band for R based on BB samples, we may estimate d_α by

the 100(1-α)% percentile of the sample $\sup\{|\psi(R^{(l)}(t)) - \psi(\hat{R}(t))| : t \in (0,1), l = 1, \ldots, N\}$, where $R^{(l)}(t)$ and $\hat{R}(t)$ are l^{th} random realization and BB estimate of $R(t)$, respectively, and substitute $\hat{d}_\alpha$ in (4).

Similar ideas may be used to estimate the IAE.

3 Application to testing binormality

Because the binormal model is popularly used in practice, it is important to validate model assumption before using it. Several methods are available, such as one based on the linearity property of TPF and FPF on the "normal-deviate axes ", the graphic method was mentioned by Swets [26], Cai and Moskowitz [7] proposed a residual plot using bootstrap sampling method, Dorfman and Alf [27], Lin and Mudholkar [28], Bozdogan and Ramirez [29], respectively, proposed a goodness-of-fit test.

Our testing of binormality procedure is motivated as follows. In the binormal model, $H(X)$ ~Normal(0,1) and $H(Y) \sim$ Normal (μ, σ^2), with the convention $\mu > 0$, the continuous monotone increasing transformation is easily identifiable as $H(x) = \Phi^{-1}(F(x))$, hereafter, Φ^{-1} and Φ denote the quantile function and CDF of standard normal distribution, respectively. By plugging in the kernel smoothed empirical estimate $\tilde{F}_m$ of F, we can estimate $H(x)$ by $\hat{H}(x) = \Phi^{-1}(\tilde{F}_m(x))$, where $\tilde{F}_m = \Phi_{\sigma_m} * F_m$, "$*$ "stands for the convolution operation and σ_m is the bandwidth. See Zou et al. [14], Lloyd [15], Zhou and Harezlak [30], among others for a discussion about the choice of the bandwidth. Now μ and σ can be estimated by the sample mean and sample standard deviation of $\hat{H}(Y_1), \ldots, \hat{H}(Y_n)$ which are denoted by $\hat{\mu}$ and $\hat{\sigma}$, respectively. Under the null hypothesis that the binormal model is true, the ROC function is given by $R(t) = \Phi(a + b\Phi^{-1}(t))$, where $a = \mu/\sigma$, $b = 1/\sigma$.

Thus, we consider the test statistic $T = \sup_t |\hat{R}(t) - \Phi(\hat{a} + \hat{b}\Phi^{-1}(t))|$, where $\hat{R}(t)$ is

empirical (or the BB) estimate of $R(t)$, $\hat{a} = \hat{\mu}/\hat{\sigma}$ and $\hat{b} = 1/\hat{\sigma}$. We reject the null hypothesis of binormality for large value of T, that is at level α, we reject if $T \geq c_\alpha$. However, it is hard to analytically compute or approximate c_α, so we employ a resampling technique. Here the Bayesian bootstrap method is used. Because strong approximation ensures under fairly mild restrictions that asymptotically, the sampling distribution of any test statistic $\gamma_{m,n}(F_m, G_n)$ is equal to the BB resampling distribution of $\gamma_{m,n}(F_m^\#, G_n^\#)$ conditional on the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ (see definitions of $F^\#(u)$, p_i's and q_j's in Section 2, and define $G^\#(u) = \sum_{j=1}^n q_i 1(Y_i \leq u)$). In this case, we may define $T^\# = \gamma_{m,n}(F_m^\#, G_n^\#) = \sup_t |R_{m,n}^\#(t) - \Phi(a^\# + b^\# \Phi^{-1}(t))|$, where $R_{m,n}^\#(t)$ is defined in Section 2, $a^\#$ and $b^\#$ are obtained through $\mu^\#$ and $\sigma^\#$ which are estimated by the sample mean and sample standard deviation of $H^\#(Y_1), \ldots, H^\#(Y_n)$. Here we define $H^\#(x) = \Phi^{-1}(\tilde{F}_m^\#(x))$, where $\tilde{F}_m^\# = \Phi_{\sigma_m} * F_m^\#$. Then c_α can be estimated by $\hat{c}_\alpha = 100(1-\alpha)\%$ percentile of the BB distribution of $T^\#$ conditional on the given samples. Thus we reject the null hypothesis of binormality if $T \geq \hat{c}_\alpha$.

4 Simulation study

In order to check the performance of the BB estimator compared to some other existing alternatives, we conduct the simulations under various situations:

1. *Comparison with some semiparametric estimation methods based on binormality:*

   We compare accuracy of the estimates of ROC curve and the AUC functional obtained by our BB method with BN-G (ROC GLM method by Pepe [5]), BN-T Box-Cox (Zou and Hall [4]) and SP (semiparametric location-scale models by Pepe [6], page 112). Note that BN-G and BN-T assume a binormal model. For BN-G, BN-T and SP methods, the bootstrap is used to estimate standard errors or construct confidence intervals. The distributions of (F, G) used to generate the data are chosen as lognormal,

location-scale exponential, gamma and beta (abbreviated as A, B, C, D, respectively) shown in Table 1 for different combinations of the parameters. We replicate each simulation 1000 times and resamples 1000 times in each replication, compare coverage probabilities and average length of 90% credible intervals for AUC based on different estimation methods in Table 1 for various parameter combinations, and also examine IAEs (see Figure 2) to evaluate the fitness of the curve estimates by different methods for certain parameter combinations in Table 1.

From the simulation results shown in Table 1 and Figure 2, we can observe that the proposed BB method performs well in view of accuracy and robustness. The method BN-G gives larger IAE in some data sets, along with less coverage of AUC with larger sample size in one of the location-scale exponential data sets. The method BN-T gives the lower coverage probability in some cases. In general, the coverage probability increases when the sample size increases from 15 to 50, while the mean lengths of 90% CI of AUC, IAE and their variations decrease significantly.

[Insert Table 1 here]

[Insert Figure 2 here]

2. *Comparison with some nonparametric estimation methods:*

There are several nonparametric estimation methods available to estimate the area under the curve. Qin and Zhou [19] conducted extensive simulations to compare the accuracy and efficiency of the estimates using various methods, which are EL (Qin and Zhou [19]), MW (Mann-Whitney two-sample rank statistics), LT (logistic transformation by Pepe [6], page 107), PB (standard percentile bootstrap) and PTB (percentile-t bootstrap). Two out of three simulated data are used which are the same setting as proposed in Qin and Zhou [19]. They are distributed as normal with mean and stan-

dard deviation $(0, 1)$ and $(5^{1/2}\Phi^{-1}(AUC), 2)$ for F and G, respectively; exponential with rate=1 for F and rate=$(\frac{1}{AUC} - 1)$ for G, where pdf of exponential distribution with rate=λ is defined by $f_\lambda(x) = \lambda e^{-\lambda x}$. We will compare the performance of our BB estimator with these estimators. From simulation results (see Table 2), we can see BB estimator performs well, especially, the BB intervals tend to be shorter.

[Insert Table 2 here]

5 Real data analyses

We shall illustrate the BB method to construct credible band of curve estimate and credible interval for AUC estimate using the data set published by Wieand et al. [31]. 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). For the purpose of illustration, we only choose biomarker CA 19-9. The BB estimates are based on 5000 resamples and grid points with even interval length 0.01 on [0,1]. We only consider a pointwise 90% credible band in this case (see Figure 3).

[Insert Figure 3 here]

Our BB estimate (corresponding 95% credible interval) of AUC is 0.8542 which is similar to those of Qin and Zhang [12]'s and Wan and Zhang [32]'s, but the corresponding confidence interval (0.7834, 0.8995) is slightly shorter.

Using the testing binormality procedure shown in Section 3, we fail to reject the binormality assumption of biomarker CA 125 and CA 19-9 for both level 0.05 and 0.1, based on $T = 0.1469$ and $c_{0.05} = 0.2462$, $c_{0.1} = 0.2311$ for CA 125, and $T = 0.3992$ and $c_{0.05} = 0.4603$, $c_{0.1} = 0.4340$ for CA 19-9. These results are based on 1000 resamples and grid points with

even interval length 0.05 on [0,1].

Acknowledgements

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

Appendix

The Matlab code to implement our BB estimate of ROC curve is given as follows:

%–Given data: x, m observations from nondisease group

% y, n observations from disease group

% $grid$, the length of even interval of false positive fraction

% rep, resample size

% Define FPF and helper vectors, based on the information given before.

$t = [grid : grid : 1 - grid]$;% FPF vector

$ot = ones(length(t), 1)$;% vector of 1 with the same length as vector t

$onx = ones(m, 1); ony = ones(n, 1)$;% vectors of 1 with the same length as x and y

%AUC function (using Simpson's method):

function [auc] =auc(roctrue,grid); %input ROC curve vector as roctrue.

$auc = grid*(roctrue(1)+roctrue(length(roctrue))+2*sum(roctrue(2 : (length(roctrue)-1))) + 2 * sum(roctrue(2 : 2 : (length(roctrue) - 1))))/3$

%BB estimate of ROC, AUC.

`for` $r = 1 : rep$;

$p1 = exprnd(1, 1, n); p = p1/sum(p1); q1 = exprnd(1, 1, n); q = q1/sum(q1);$ % note: to generate Dirichlet weight vectors p and q.

$roc(r, :) = q * (z' * ot' < ony * t);$

$aucbb(r) = auc(roc(r, :), grid);$

`end`;

$rocbb = mean(roc)$; % BB estimate of ROC

$aucbb = auc(rocbb, grid)$; % BB estimate of AUC

Other sampling error information can be obtained easily.

References

1. Green DM, Swets JA. *Signal Detection Theory and Psychophysics.* John Wiley & Sons: New York, 1966.

2. Hsieh F, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics* 1996; **24**:25-40.

3. Metz CE, Herman BA, Shen J. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine* 1998; **17**:1033–1053.

4. Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics* 2000; **27**:621–631.

5. Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* 2000; **56**:352–359.

6. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford Statistical Science Series: Oxford University Press, 2003.

7. Cai T, Moskowitz C. Semiparametric estimation of the binormal ROC curve. *Biostatistics* 2004; **5**:573–586.

8. Goddard MJ, Hinberg I. Receiver operating characteristic (ROC) curves and non-normal data: An empirical study. *Statistics in Medicine* 1990; **9**:325–337.

9. Dorfman DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA, Dagga HA. Proper receiver operating characteristic analysis: The bigamma model. *Academic Radiology* 1997; **4**:138–149.

10. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, Wells WM 3rd, Jolesz FA, Kikinis R. Statistical validation of image segmentation quality based on a spatial overlap index. *Academic Radiology* 2004; **11**:178–189.

11. Li G, Tiwari RC, Wells MT. Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. *Biometrika* 1999; **86**:487-502.

12. Qin J, Zhang B. Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika* 2003; **90**:585–596.

13. Hall P, Zhou XH. Nonparametric estimation of component distributions in a multivariate mixture. *The annals of statistics* 2003; **31**:201–224.

14. Zou KH, Hall, WJ, Shapiro DE. Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1997; **16**:2143–2156.

15. Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association* 1998;

93:1356–1364.

16. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.

17. Brownie C, Simonoff JS, Hochberg Y, Reiser B. Estimating $\Pr(X < Y)$ in categorized data using ROC analysis. *Biometrics* 1988; **44**:615–621.

18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**:837–845.

19. Qin G, Zhou XH. Empirical likelihood inference for the area under the ROC curve. *Biometrics* 2006; **62**:613–622.

20. Li G, Tiwari RC, Wells MT. Quantile comparison functions in two-sample problems: With applications to comparisons of diagnostic markers. *Journal of the American Statistical Association* 1996; **91**:689–698.

21. Rubin DB. The Bayesian bootstrap. *The Annals of Statistics* 1981; **9**:130–134.

22. Gu J., Ghosal, S. Strong approximations for resample quantile processes and application to ROC methodology. *Institute of Statistics Mimeo Series North Carolina State University* 2006; **2593**:1–32.

23. Moise A, Clement B, Raissis M, Nanopoulos P. A test for crossing receiver operating characteristic (ROC) curves. *Communications in Statistics— Theory and Methods* 1988; **17**:1985–2003.

24. Efron B. Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics* 1979; **7**:1–26.

25. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**:639–650.

26. Swets JA. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin* 1986; **99**:181–198.

27. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal-detection theory–A diret solution. *Psychometricka* 1968; **33**:117–124.

28. Lin C, Mudholkar G. A simple test for normality against asymmetric alternatives. *Biometrika* 1980; **67**:455–461.

29. Bozdogan H, Ramirez DE. Testing for model fit: Assessing the Box-Cox transformation of multivariate data to near normality. *Computational Statistics Quarterly* 1986; **3**:127–150.

30. Zhou XH, Harezlak J. Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine* 2002; **21**:2045–2055.

31. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; **76**:585–592.

32. Wan S, Zhang B. Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. *Statistics in Medicine* 2007;**26**:2565–2586.

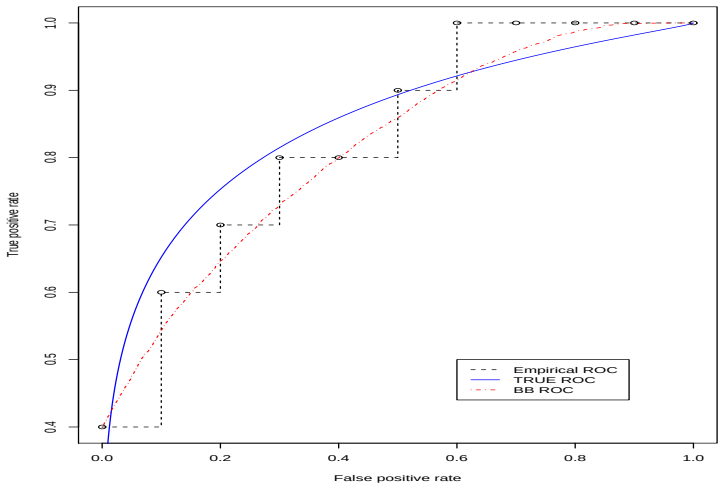

Figure 1: Comparison of empirical and the BB's estimate of ROC with the true (Simulation dataset: $X_1, \ldots, X_m \sim$ iid $N(0,1)$, $Y_1, \ldots, Y_n \sim$ iid $N(1.868, 1.5^2)$, m=n=10, 5000 resamples, grid points on [0,1] are chosen with equal interval length 0.001).

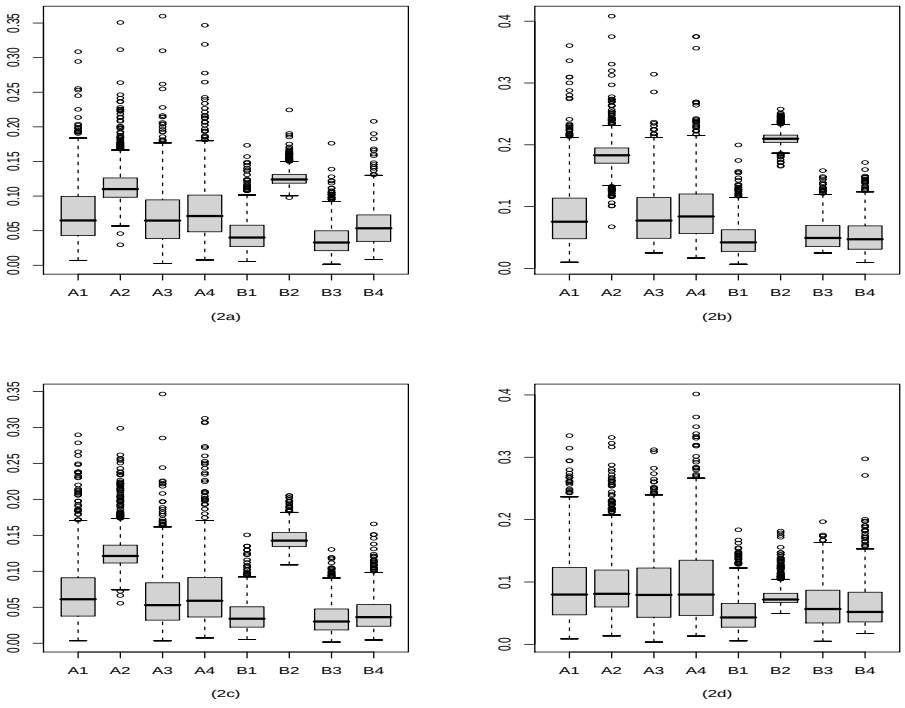

Figure 2: From (2a) to (2d), the boxplots of IAE are shown using the first data set of A,B,C,D, respectively in Table 1. Index A (or B)–m=n=15 (or m=n=50); Index 1–BB method, 2–BN-G, 3–BN-T, 4–SP.

Table 1: Coverage probabilities of AUC and corresponding average lengths of the 90% CI shown beneath in the parathesis. Our simulation results are based on 1000 simulated data sets and corresponding 1000 resamples. Data are generated by lognormal, location-scale exponential, gamma and beta distributions (abbreviated as A, B, C, D, respectively) with different combinations of the parameters (A: X and Y datesets are generated from the lognormal with corresponding normal parameters (u_x, σ_x) and (u_y, σ_y), respectively; B: X's and Y's are generated from the exponential distribution with rate=0.5 and the location and scale parameters (u_x, σ_x) and (u_y, σ_y), respectively; C: X's and Y's are generated from gamma distribution with mean and standard error (u_x, σ_x) and (u_y, σ_y), respectively; D: X's and Y's are generated from beta distribution with mean and standard error (u_x, σ_x) and (u_y, σ_y), respectively). The grid points on [0,1] is chosen with equal interval length 0.05.

Data		$m = n = 15$				$m = n = 50$			
u_x, σ_x	u_y, σ_y	BB	BN-G	BN-T	SP	BB	BN-G	BN-T	SP
A		.899	.886	.866	.923	.893	.871	.882	.900
0,1	1, 1	(.262)	(.254)	(.250)	(.276)	(.149)	(.141)	(.143)	(.176)
		.861	.859	.801	.820	.886	.890	.873	.150
0,1	3, 3	(.230)	(.233)	(.205)	(.275)	(.136)	(.134)	(.122)	(.162)
B		.875	.862	.860	.880	.886	.888	.856	.885
0,1	1, 1	(.305)	(.288)	(.285)	(.304)	(.172)	(.159)	(.158)	(.167)
		.854	.857	.910	.930	.902	.772	.925	.919
0,1	3, 3	(.098)	(.097)	(.076)	(.110)	(.057)	(.062)	(.045)	(.065)
C		.886	.876	.840	.873	.886	.851	.860	.867
1,1	2, 1	(.244)	(.234)	(.223)	(.244)	(.140)	(.132)	(.133)	(.146)
		.852	.861	.858	.882	.906	.837	.886	.847
1,1	5, 3	(.101)	(.105)	(.084)	(.108)	(.059)	(.061)	(.051)	(.057)
D		.891	.878	.867	.865	.898	.892	.774	.821
.15, .15	.2, .3	(.321)	(.323)	(.336)	(.348)	(.182)	(.177)	(.212)	(.184)
		.896	.876	.874	.834	.886	.912	.796	.662
.15, .15	.5, .45	(.333)	(.332)	(.440)	(.341)	(.189)	(.187)	(.266)	(.211)

Table 2: Coverage probabilities and corresponding average lengths of 95% (shown beneath) CI for AUC obtained by BB and other nonparametric methods based on our simulation and the information contained in Qin and Zhou [19]. Our simulation results are based on 10000 simulated data sets and corresponding 1000 resamples. The grid points on [0,1] is chosen with equal interval length 0.005, $(m, n) = (50, 50)$.

Data/AUC	BB	EL	MW	LT	PB	PTB
Normal:	.9438	.9407	.9379	.9538	.9300	.9690
.8	.1765	.1783	.1808	.1808	.1746	.1971
	.9315	.9352	.9204	.9468	.9150	.9700
.9	.1234	.1281	.1271	.1326	.1228	.1591
	.9066	.8964	.8818	.9289	.8840	.9490
.95	.0823	.0874	.0850	.0930	.0814	.1360
Exponential:	.9465	.9446	.9394	.9551	.9270	.9570
.8	.1708	.1725	.1746	.1748	.1692	.1887
	.9396	.9321	.9200	.9482	.9240	.9740
.9	.1212	.1254	.1247	.1290	.1198	.1501
	.9049	.8977	.8817	NA	.9000	.9460
.95	.0823	.0881	.0859	NA	.0838	.1427

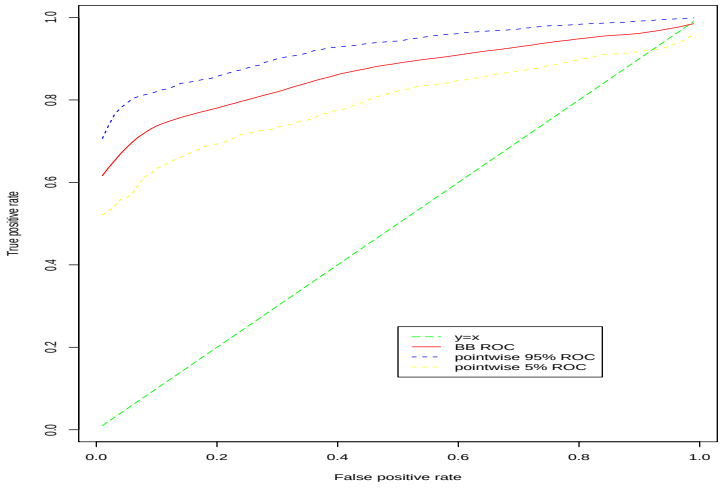

Figure 3: Pointwise 90% credible band of ROC curve using biomarker CA 19-9.