UW Biostatistics Working Paper Series

4-25-2008

Semi-Parametric Maximum Likelihood Estimates for ROC Curves of Continuous-Scale Tests

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

Huazhen Lin

Semi-parametric maximum likelihood estimates for ROC curves of continuous-scale tests

Xiao-Hua Zhou [†] [*] **and Huazhen Lin** [*‡]

Summary

In this paper, we propose a new semi-parametric maximum likelihood (ML) estimate of an ROC curve that satisfies the property of invariance of the ROC curve and is easy to compute. We show that our new estimator is $\sqrt{n}$-consistent and has an asymptotically normal distribution. Our extensive simulation studies show the proposed method is efficient, robust, and simple to compute. Finally, we illustrate the application of the proposed estimator in a real data set.

Key words: ROC curves; Sensitivity and specificity; Semi-parametric maximum likelihood estimators.

1. Introduction

When the response of a diagnostic test is continuous, its diagnostic accuracy is best represented by the receiver operating characteristic (ROC) curve [1]. Let F_1 and F_0 denote distribution functions of the test result Y_1 of a diseased subject and Y_0 of a non-diseased subject, respectively. Then, the ROC curve of the test can be written as

$$ROC(u) = 1 - F_1(F_0^{-1}(1-u)), \tag{1}$$

where F_0^{-1} is the inverse function of F_0, and u is the false positive rate (FPR) corresponding to a cut-off point for positivity. It is well-known that the ROC curve of a test must be invariant to any monotone increasing transformation of test results, a fundamental property of an ROC curve. Hence, any sensible

[†]HSR&D Center of Excellence, VA Puget Sound Health Care System, Seattle, WA 98108

[*]Department of Biostatistics, University of Washington, Seattle, WA 98195

[‡]School of Mathematics, Sichuan University, Chengdu, Sichuan 610064, P. R. China

estimation methods should have this property. In the statistical literature, many parametric, semi-parametric, and non-parametric methods have been proposed for estimating an ROC curve. In general, pure parametric methods do not possess the invariance property; the empirical non-parametric and smoothing non-parametric methods have the property of invariance [2,3]. However, the jagged form of the empirical ROC curve estimator can result in underestimating the true ROC curve as the true ROC curve is a smooth function, and the intensive computation and challenging bandwidth selection of the smoothing non-parametric estimators may affect their application in practice.

An intermediate strategy between pure parametric and non-parametric methods is a semi-parametric approach. The most commonly used semi-parametric procedure to estimate the ROC curve is to assume a parametric form for the ROC curve, but avoid making any additional parametric assumptions about the distributions of test results. This type of semi-parametric method has the property of invariance. In this paper, we focus on this type of semi-parametric method.

We assume that the ROC curve has the parametric form,

$$ROC(u) = G(\alpha_0 + \alpha_1 H^{-1}(u)), \tag{2}$$

where G and H are some known cumulative distribution functions. The most common choice for G and H is the binormal form, $G = H = \Phi$, where Φ is the cumulative distribution function of the standard normal random variable.

Under the binormal model, several methods have been proposed by Metz et al. [4], Alonzo and Pepe [5], Pepe and Cai [6], Zou and Hall [7], and Cai and Moskowitz [8]. The first approach, proposed by Metz et al. [4] and denoted by MHS, is to first categorize continuous test data into ordinal-scale categorical data and then to apply the maximum likelihood method to estimating the parameters in the binormal model by assuming that the continuous variable can be approximated by a discrete variable with finite support points; that is, the continuous distribution function can be specified by finite parameters, which is difficult in practice. Furthermore, they use an ad hoc approach to reduce the number of the parameters, hence, their resulting estimate is no longer a ML estimate even though their likelihood is correct. The second approach, proposed by Alonzo and Pepe [5] and denoted by AP, is to estimate the ROC curve by using procedures for fitting generalized linear models to binary data. The third approach, proposed by Pepe and Cai [6] and denoted by PC, is to first write the ROC curve as the distribution

of placement values and then to estimate the ROC curve by maximizing the pseudo likelihood function of the estimated placement values. None of the above three methods is a truly maximum likelihood (ML) estimator, and hence they do not possess the optimal property associated with ML estimators, for example, fully efficient. The fourth method, proposed by Zou and Hall [7] and denoted by ZH, is to use rank data to estimate the ROC curve by assuming semi-parametric distributions for test results of diseased and non-diseased subjects and using maximum likelihood algorithms. However, the likelihood function in the ZH method is a high-dimensional integral, and hence numerical methods include Monte Carlo are required to evaluate the likelihood. Recently, Cai and Moskowitz [8] have proposed a profile maximum likelihood approach from the raw data to estimate the ROC curve, denoted by CM; however, their computation algorithm requires choosing reasonable initial values for a large number of nuisance parameters, which may be difficult in practice when the sample size is large.

In this paper, we propose a new profile maximum likelihood approach to estimate the ROC curve. Compared to the Cai and Moskowitz' method, our method has a smaller number of nuisance parameters to estimate. Furthermore, our estimator can be computed by using an algorithm that is based on a recursive relationship among the nuisance parameters, without specifying initial values for a large number of nuisance parameters. Our estimator is $\sqrt{n}$-consistent and has an asymptotically normal distribution, and our extensive simulation studies show the proposed method is efficient, robust, and simple to compute.

Since the binormal model is the most commonly used form for an ROC curve, from now on we assume that the true ROC curve is defined by

$$ROC(u) = \Phi(\alpha_0 + \alpha_1 \Phi^{-1}(u)), \tag{3}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution of the standard normal distribution. Equivalently, we can derive model (3) by assuming that there exists an unknown monotone increasing function $g(.)$ such that $g(Y_0)$ has the standard normal distribution and $g(Y_1)$ has a normal distribution with mean μ and standard deviation σ. Then, the resulting ROC curve satisfies model (3) with $\alpha_0 = \mu/\sigma$ and $\alpha_1 = 1/\sigma$.

The paper is organized as follows. In Section 2, we propose a semi-parametric maximum likelihood estimator of an ROC curve and an algorithm for computing it, and develop asymptotic properties for

the resulting estimator. In Section 3, we perform simulation studies to asses efficiency and robustness of our estimator relative to the existing methods and to verify the validity of the asymptotic inferences in finite samples. In Section 4, we illustrate the application of our method in an example.

2. Semi-parametric maximum likelihood estimate

Data available for making inferences consist of a random sample of size n_1 from the diseased population with the unknown cumulative distribution function F_1, $\{Y_{1j}, j = 1, \cdots, n_1\}$, and a random sample of size n_0 from the non-diseased population, $\{Y_{0i}, i = 1, \cdots, n_0\}$, with the unknown cumulative distribution function F_0. Denote $n = n_0 + n_1$.

Let f_0 and f_1 be the density functions of F_0 and F_1, respectively. Then, the likelihood function of observations $Y_{0i}, i = 1, \cdots, n_0$, and $Y_{1j}, j = 1, \cdots, n_1$, is given by

$$L = \prod_{i=1}^{n_0} f_0(Y_{0i}) \prod_{j=1}^{n_1} f_1(Y_{1j}). \tag{4}$$

Under model (3), we know that there exists a unknown increasing and differentiable function $g(.)$ such that $g(Y_0)$ has the standard normal distribution and $g(Y_1)$ has a normal distribution with mean μ and standard deviation σ. Denote $\phi(x)$ to be the standard normal density function, therefore, we have that $f_0(y) = \phi(g(y))g'(y)$ and $f_1(y) = \phi(-\alpha_0 + \alpha_1 g(y))\alpha_1 g'(y)$. Hence we can write the likelihood function (4) as

$$L = \prod_{i=1}^{n_0} \phi(g(Y_{0i}))g'(Y_{0i}) \prod_{j=1}^{n_1} \phi(-\alpha_0 + \alpha_1 g(Y_{1j}))\alpha_1 g'(Y_{1j}). \tag{5}$$

Consequently, the ML estimation of the *ROC* curve parameters α_0 and α_1 requires simultaneous estimation of the unknown function g. Denote $dF(x) = F(x) - F(x-)$, where " $-$ " represents the left limit. Our approach to estimating α_0, α_1 and g is based on the nonparametric maximum likelihood approach [9, 10, 11, 12], which seek to maximize the function $\tilde{L}$ given by

$$\tilde{L} = \prod_{i=1}^{n_0} \phi(g(Y_{0i}))dg(Y_{0i}) \prod_{j=1}^{n_1} \phi(-\alpha_0 + \alpha_1 g(Y_{1j}))\alpha_1 dg(Y_{1j});$$

that is,

$$\tilde{L} = \prod_{i=1}^{n_0} \{\Phi(g(Y_{0i})) - \Phi(g(Y_{0i}-))\} \prod_{j=1}^{n_1} \{\Phi(-\alpha_0 + \alpha_1 g(Y_{1j})) - \Phi(-\alpha_0 + \alpha_1 g(Y_{1j}-))\}. \tag{6}$$

Denote the distinct ordered test results from the combined sample, Y_{0i}'s and Y_{1j}'s, by $Y^*_{(1)} < \cdots < Y^*_{(I^*_n-1)}$, where $I^*_n - 1$ is the number of distinct values among Y_{0i}'s and Y_{1j}'s. Therefore we can write the likelihood function (6) as follows:

$$\tilde{L} = \prod_{r=1}^{I^*_n-1} \left(\Phi(g(Y^*_{(r)})) - \Phi(g(Y^*_{(r)}-))\right)^{k^*_r} \left(\Phi(-\alpha_0 + \alpha_1 g(Y^*_{(r)})) - \Phi(-\alpha_0 + \alpha_1 g(Y^*_{(r)}-))\right)^{\ell^*_r}, \tag{7}$$

where frequency counts $k^*_r = \#\{Y_{0i} = Y^*_{(r)}, i = 1, \cdots, n_0\}$ and $\ell^*_r = \#\{Y_{1j} = Y^*_{(r)}, j = 1, \cdots, n_1\}$, corresponding to non-diseased and diseased subjects at distinct ordered test results.

Let $\Omega = \{\text{all monotone increasing functions on}(-\infty, \infty)\}$. Denote $Y^*_{(0)} = -\infty$ and $Y^*_{(I^*_n)} = +\infty$. Since Φ is monotonic and $\alpha_1 > 0$, to maximize $\tilde{L}$, we need to make the $g(Y^*_{(r)})$ as large, and the $g(Y^*_{(r)}-)$ as small. In addition, because g is monotonic and $\int_{-\infty}^{+\infty} d\Phi(g(y)) = 1$, hence, any $g \in \Omega$ that maximizes $\tilde{L}$ must satisfy

$$g(y) = g(Y^*_{(r-1)}), \quad \text{if} \quad Y^*_{(r-1)} \le y < Y^*_{(r)}$$

for $r = 1, \cdots, I^*_n$. Hence the maximum likelihood estimate of g, denoted by $\widehat{g}$, has to be a discrete function that only jumps at observations $Y^*_{(1)} < \cdots < Y^*_{(I^*_n-1)}$.

With $C^*_r = g(Y^*_{(r)})$, $C^*_0 = -\infty$, and $C^*_{I^*_n} = +\infty$, we can write (7) as

$$L_n(\theta, g) = \prod_{r=1}^{I^*_n} \left(\Phi(C^*_r) - \Phi(C^*_{r-1})\right)^{k^*_r} \left(\Phi(-\alpha_0 + \alpha_1 C^*_r) - \Phi(-\alpha_0 + \alpha_1 C^*_{r-1})\right)^{\ell^*_r} \tag{8}$$

when $g = \widehat{g}$. Therefore the estimation of ROC curve parameters α_0 and α_1, which are of primary interest, requires simultaneous estimation of the $I^*_n - 1$ number of nuisance parameters, $C^*_1, \cdots, C^*_{I^*_n-1}$.

Using the same idea as in Metz et al. [4], we note that some of the jump points of $\widehat{g}$, $Y^*_{(r)}$'s, can be ignored for estimating α_0 and α_1, which means we can obtain the estimates of α_0 and α_1 with fewer nuisance parameters. We state the results in Conclusion 1 below.

Denote

$$D(Y^*_{(r)}) = \begin{cases} 2 & \text{if} \quad k^*_r > 0 \quad \text{and} \quad \ell^*_r > 0 \\ 1 & \text{if} \quad k^*_r > 0 \quad \text{and} \quad \ell^*_r = 0 \\ 0 & \text{if} \quad k^*_r = 0 \quad \text{and} \quad \ell^*_r > 0 \end{cases}$$

and

$$\Re = \left\{ Y^*_{(r)} : D(Y^*_{(r)}) = D(Y^*_{(r+1)}) \leq 1, 1 \leq r \leq I^*_n - 2 \right\}.$$

Each jump point in $\Re$ has the same diseased status as its next contiguous jump point. Here, $\Re$ includes all jump points of a contiguous sequence with the same diseased status except the last point in the sequence.

Conclusion 1 . The maximum likelihood estimates of α_0 and α_1 can be determined by some estimating equations that don't depend on those nuisance parameters $C^*_r = g(Y^*_{(r)})$ for which $Y^*_{(r)}$ belongs to $\Re$.

See Appendix A.1 for a proof of Conclusion 1. A practical consequence of the conclusion is that we can ignore the jump points in $\Re$ for estimating α_0 and α_1.

After deleting the points in $\Re$, we denote the remaining jump points of $\widehat{g}$ by $Y_{(1)} < \cdots < Y_{(I_n-1)}$ and let $C_r = g(Y_{(r)})$ for $1 \leq r \leq I_n - 1$, $C_0 = -\infty$ and $C_{I_n} = +\infty$. The MLE of $\theta = (\alpha_0, \alpha_1)^T$ and $\mathbf{C} = (C_1, \cdots, C_{I_n-1})^T$ can be obtained by maximizing

$$\mathcal{L}_n(\theta, g) = \prod_{r=1}^{I_n} \left(\Phi(C_r) - \Phi(C_{r-1})\right)^{k_r} \left(\Phi(-\alpha_0 + \alpha_1 C_r) - \Phi(-\alpha_0 + \alpha_1 C_{r-1})\right)^{\ell_r}, \tag{9}$$

which is essentially (8) with I^*_n replaced by I_n. Here, for $2 \leq r \leq I_n - 1$, $k_r = \#\{Y_{(r-1)} < Y_{0i} \leq Y_{(r)}, i = 1, \cdots, n_0\}$ and $\ell_r = \#\{Y_{(r-1)} < Y_{1j} \leq Y_{(r)}, j = 1, \cdots, n_1\}$; $k_1 = \#\{Y_{0i} \leq Y_{(1)}, i = 1, \cdots, n_0\}$, $\ell_1 = \#\{Y_{1j} \leq Y_{(1)}, j = 1, \cdots, n_1\}$, $k_{I_n} = \#\{Y_{0i} < Y_{(I_n-1)}, i = 1, \cdots, n_0\}$, and $\ell_{I_n} = \#\{Y_{1j} < Y_{(I_n-1)}, j = 1, \cdots, n_1\}$.

To obtain the estimator of θ, in the paper, we propose a two-stage iterative procedure, alternating the parametric and nonparametric estimation steps. Our idea for the nonparametric estimate is from Kvam and Samaniego [11].

Given θ, we can find the estimator of $\mathbf{C}$ by maximizing the likelihood function (9) with respect to

$\mathbf{C}$. The estimator of $\mathbf{C}$ must satisfy the following $(I_n - 1)$ score equations:

$$\begin{aligned}
\frac{\partial \log\{\mathcal{L}_n(\theta, g)\}}{\partial C_1} &= k_1 \frac{\phi(C_1)}{\Phi(C_1)} - k_2 \frac{\phi(C_1)}{\Phi(C_2) - \Phi(C_1)} \\
&\quad + \alpha_1 \ell_1 \frac{\phi(-\alpha_0 + \alpha_1 C_1)}{\Phi(-\alpha_0 + \alpha_1 C_1)} - \alpha_1 \ell_2 \frac{\phi(-\alpha_0 + \alpha_1 C_1)}{\Phi(-\alpha_0 + \alpha_1 C_2) - \Phi(-\alpha_0 + \alpha_1 C_1)} = 0, \\
\frac{\partial \log\{\mathcal{L}_n(\theta, g)\}}{\partial C_r} &= k_r \frac{\phi(C_r)}{\Phi(C_r) - \Phi(C_{r-1})} - k_{r+1} \frac{\phi(C_r)}{\Phi(C_{r+1}) - \Phi(C_r)} \\
&\quad + \alpha_1 \ell_r \frac{\phi(-\alpha_0 + \alpha_1 C_r)}{\Phi(-\alpha_0 + \alpha_1 C_r) - \Phi(-\alpha_0 + \alpha_1 C_{r-1})} \\
&\quad - \alpha_1 \ell_{r+1} \frac{\phi(-\alpha_0 + \alpha_1 C_r)}{\Phi(-\alpha_0 + \alpha_1 C_{r+1}) - \Phi(-\alpha_0 + \alpha_1 C_r)} = 0, \quad 2 \le r \le I_n - 2, \\
\frac{\partial \log\{\mathcal{L}_n(\theta, g)\}}{\partial C_{I_n - 1}} &= k_{I_n - 1} \frac{\phi(C_{I_n - 1})}{\Phi(C_{In-1}) - \Phi(C_{I_n - 2})} - k_{I_n} \frac{\phi(C_{I_n - 1})}{1 - \Phi(C_{I_n - 1})} \\
&\quad + \alpha_1 \ell_{I_n - 1} \frac{\phi(-\alpha_0 + \alpha_1 C_{I_n - 1})}{\Phi(-\alpha_0 + \alpha_1 C_{I_n - 1}) - \Phi(-\alpha_0 + \alpha_1 C_{I_n - 2})} \\
&\quad - \alpha_1 \ell_{I_n} \frac{\phi(-\alpha_0 + \alpha_1 C_{I_n - 1})}{1 - \Phi(-\alpha_0 + \alpha_1 C_{I_n - 1})} = 0. \qquad (10)
\end{aligned}$$

Inspection of (10) shows that finding the estimator of $\mathbf{C}$ in a closed form is a challenge. Hence, an iterative algorithm is required. However, the standard Newton-Raphson iteration requires inversion of an $(I_n - 1) \times (I_n - 1)$ matrix, and this computation can become a problem if I_n is large. But the solution to the likelihood equations can be easily obtained if an initial estimate of C_1 is given. Note that $\ell_r k_r = 0$, for $1 \le r \le I_n$; $\ell_r \ell_{r+1} = 0$ and $k_r k_{r+1} = 0$ for $1 \le r \le I_n - 1$. Suppose that we have selected an initial value of C_1, $\check{C}_1$. Then from the first equation of (10), we obtain an estimate, $\check{C}_2$, of C_2,

$$\check{C}_2 = \begin{cases} \frac{\alpha_0}{\alpha_1} + \frac{1}{\alpha_1} \Phi^{-1} \left(\Phi(-\alpha_0 + \alpha_1 \check{C}_1) + \frac{\alpha_1 \ell_2 \phi(-\alpha_0 + \alpha_1 \check{C}_1) \Phi(\check{C}_1)}{k_1 \phi(\check{C}_1)} \right) & \text{if} \quad k_2 = 0 \\ \Phi^{-1} \left(\Phi(\check{C}_1) + \frac{k_2 \Phi(-\alpha_0 + \alpha_1 \check{C}_1) \phi(\check{C}_1)}{\alpha_1 \ell_1 \phi(-\alpha_0 + \alpha_1 \check{C}_1)} \right) & \text{if} \quad \ell_2 = 0 \end{cases}.$$

For $r = 2, \cdots, I_n - 2$, using the latest estimates, $\check{C}_{r-1}$ and $\check{C}_r$, of C_{r-1} and C_r, we solve the rth equation of (10) to obtain the following estimate of C_{r+1}:

$$\check{C}_{r+1} = \begin{cases} \frac{\alpha_0}{\alpha_1} + \frac{1}{\alpha_1} \Phi^{-1} \left(\Phi(-\alpha_0 + \alpha_1 \check{C}_r) + \frac{\alpha_1 \ell_{r+1} \phi(-\alpha_0 + \alpha_1 \check{C}_r) \left(\Phi(\check{C}_r) - \Phi(\check{C}_{r-1}) \right)}{k_r \phi(\check{C}_r)} \right) & \text{if} \quad k_{r+1} = 0 \\ \Phi^{-1} \left(\Phi(\check{C}_r) + \frac{k_{r+1} \phi(\check{C}_r) \left(\Phi(-\alpha_0 + \alpha_1 \check{C}_r) - \Phi(-\alpha_0 + \alpha_1 \check{C}_{r-1}) \right)}{\alpha_1 \ell_r \phi(-\alpha_0 + \alpha_1 \check{C}_r)} \right) & \text{if} \quad \ell_{r+1} = 0 \end{cases}.$$

Hence, given the initially chosen value of C_1, $\check{C}_1$, we can obtain the estimates, $\check{C}_2, \ldots, \check{C}_{I_n - 1}$, of $C_2, \ldots, C_{I_n - 1}$ by solving the first $I_n - 2$ equations in (10). Now we are left to check whether those

estimates also satisfy the last equation in (10),

$$\Lambda(\check{C}_{I_n-2}, \check{C}_{I_n-1}) = 0, \tag{11}$$

where $\Lambda(C_{I_n-2}, C_{I_n-1}) = \frac{\partial}{\partial C_{I_n-1}} \log\{\mathcal{L}_n(\theta, g)\}$. If $\Lambda(\check{C}_{I_n-2}, \check{C}_{I_n-1}) = 0$, the estimates, $\check{C}_r$, $r = 1, \ldots, I_n - 1$, are the unique solution to Equation (10). If $\Lambda(\check{C}_{I_n-2}, \check{C}_{I_n-1}) \neq 0$, we need to update the initially chosen value estimate, $\check{C}_1$, and repeat the whole estimation process until the last equation in (10) is satisfied.

Let θ_0 be the true value of θ and g_0 be the true function of g. Denote $C_{r0} = g_0(Y_{(r)})$. In the following Conclusion 2, we establish the relationship between C_1 and $\Lambda(C_{I_n-2}, C_{I_n-1})$ to help in updating the initially chosen value, $\check{C}_1$. We provide a proof for Conclusion 2 in Appendix A.2.

Conclusion 2. Let $\theta_n = \theta_0 + o_p(1)$. For any initial chosen value $\check{C}_1$ of C_1, we let $\check{C}_2, \ldots, \check{C}_{I_n-1}$ be the corresponding solution to the first $(I_n - 2)$ equations in (10) and $\check{g}$ be the corresponding function for g. Then, when n is large enough,

1. if $\check{C}_1 < C_{10}$, then $\check{C}_r < C_{r0}$ for $r = 2, \cdots, I_n - 1$, and

$$\Lambda(\check{C}_{I_n-2}, \check{C}_{I_n-1}) = \frac{\partial}{\partial C_{I_n-1}} \log\{\mathcal{L}_n(\theta, g)\}|_{g=\check{g}, \theta=\theta_n} < 0;$$

2. if $\check{C}_1 > C_{10}$, then $\check{C}_r > C_{r0}$ for $r = 2, \cdots, I_n - 1$, and

$$\Lambda(\check{C}_{I_n-2}, \check{C}_{I_n-1}) = \frac{\partial}{\partial C_{I_n-1}} \log\{\mathcal{L}_n(\theta, g)\}|_{g=\check{g}, \theta=\theta_n} > 0.$$

The results of Conclusion 2 provide a mechanism for updating the initially chosen value $\check{C}_1$. If $\Lambda(\check{C}_{I_n-2}, \check{C}_{I_n-1}) < 0$, we should increase our initially chosen value, $\check{C}_1$. On the other hand, if $\Lambda(\check{C}_{I_n-2}, \check{C}_{I_n-1}) > 0$, we should decrease our initially chosen value, $\check{C}_1$.

Given $\mathbf{C}$, we can estimate θ by maximizing (9). We next outline the two-stage iterative procedure for estimating θ and $\mathbf{C}$.

- Step 1. We combine data from the diseased and non-diseased samples and order test results in the combined sample, replace each test result by its true disease status. As a result, we create a sequence of disease statuses for the combined sample. Denote the number of different sequences with the same consecutive disease status by I_n. Then, count the number of el-

ements in each sequence, denoted by $\mathbf{k} = \{k_1, \cdots, k_{I_n} | \sum_{r=1}^{I_n} k_r = n_0\}$ for non-diseased subjects and $\ell = \{\ell_1, \cdots, \ell_{I_n} | \sum_{s=1}^{I_n} \ell_s = n_1\}$ for diseased subjects. For example, if we have data $\{5.38, 2.1, 4.5\}$ for non-diseased subjects and $\{12.5, 10.4, 16.8, 5.1, 13.5\}$ for diseased subjects, the ordered test results in the combined sample are $\{2.1, 4.5, 5.1, 5.38, 10.4, 12.5, 13.5, 16.8\}$, and their corresponding disease statuses are $\{no, no, di, no, di, di, di, di\}$, where *no* and *di* indicate a non-diseased and diseased subject, respectively. Thus, in the above notation, we have $I_n = 4$ and $k_1 = 2, k_2 = 0, k_3 = 1, k_4 = 0$ and $\ell_1 = 0, \ell_2 = 1, \ell_3 = 0, \ell_4 = 4$.

- Step 2. Given values of α_0, α_1, we estimate $C_1, \cdots, C_{I_n-1}$ by solving (10).

- Step 3. Given estimates of $C_1, \cdots, C_{I_n-1}$, we estimate α_0 and α_1 by maximizing (9) respect to α_0 and α_1.

- Step 4. Repeat Steps 2 and 3 until two successive values for $(\alpha_0, \alpha_1, C_1, \cdots, C_{I_n-1})$ converge. Denote the convergent values of the iterate by $\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{C}_1, \cdots, \widehat{C}_{I_n-1}$.

- Step 5. The estimator $\widehat{g}$, which is a discrete function that only jumps at observations $Y^*_{(1)} < \cdots < Y^*_{(I^*_n-1)}$, can be obtained by noting that for a sequence of M contiguous jump points which only involve non-diseased subjects, we have

$$\frac{k^*_r}{\Phi(C^*_r) - \Phi(C^*_{r-1})} = \cdots = \frac{k^*_{r+M-1}}{\Phi(C^*_{r+M-1}) - \Phi(C^*_{r+M-2})} = \frac{\sum_{j=r}^{r+M-1} k^*_j}{\Phi(C^*_{r+M-1}) - \Phi(C^*_{r-1})}. \tag{12}$$

  Since estimation of C^*_{r-1} and C^*_{r+M-1} has been obtained from Step 4, $\widehat{C}^*_r$ can be obtained by the equality of the first and the last terms in the (12), and then, $\widehat{C}^*_{r+1}$ can be obtained by the equality of the second and the last terms in the (12) and so on. Similarly we can estimate values of g at the jump points only involving diseased subjects by noting that for a sequence of M contiguous jump points that only involves diseased subjects, we have

$$\frac{\ell^*_r}{\Phi(\theta' D^*_r) - \Phi(\theta' D^*_{r-1})} = \cdots = \frac{\ell^*_{r+M-1}}{\Phi(\theta' D^*_{r+M-1}) - \Phi(\theta' D^*_{r+M-2})} = \frac{\sum_{j=r}^{r+M-1} \ell^*_j}{\Phi(\theta' D^*_{r+M-1}) - \Phi(\theta' D^*_{r-1})},$$

  where $D^*_r = (-1, C^*_r)'$.

Our final estimate of θ is actually a profile likelihood estimate, which maximizes the profile likelihood for θ given by

$$PL(\theta) = \mathcal{L}_n(\theta, \widehat{g}(\theta)),$$

where $\widehat{g}(\theta)$ maximizes the likelihood $\mathcal{L}_n(\theta, g)$ for a fixed value of θ. This estimator is a function of the test values only through their ranks. Using the results on the properties of maximum profile likelihood estimates derived by Murphy and Van der Varrt[13], we can show that $\widehat{\theta}$ is fully efficient and has the following asymptotic distribution result:

$$n^{1/2}(\widehat{\theta} - \theta_0) \to N(0, \Sigma_0),$$

where

$$\begin{aligned}\Sigma_0 &= \lim_{n\to\infty} \left\{ -\frac{\partial^2 \log \mathcal{L}_n(\theta, g)}{n\partial\theta\partial\theta'} + \left(\frac{\partial^2 \log \mathcal{L}_n(\theta, g)}{n\partial\theta\partial\mathbf{C}'}\right) \right. \\ &\quad \left. \times \left(\frac{\partial^2 \log \mathcal{L}_n(\theta, g)}{n\partial\mathbf{C}\partial\mathbf{C}'}\right)^{-1} \left(\frac{\partial^2 \log \mathcal{L}_n(\theta, g)}{n\partial\mathbf{C}\partial\theta'}\right) \right\}^{-1} |_{\theta=\theta_0, g=g_0}.\end{aligned}$$

Based on the estimates of α_0 and α_1, we can estimate the ROC curve by $\widehat{ROC}(u) = \Phi(\widehat{\alpha}_0 + \widehat{\alpha}_1 \Phi^{-1}(u))$. Using the Taylor series expansion and the asymptotically normal result of $\widehat{\theta}$, we have the following asymptotic distribution result for the estimated ROC curve:

$$n^{1/2}\left(\widehat{ROC}(u) - ROC(u)\right) \to N(0, \sigma_R^2),$$

where

$$\sigma_R^2 = \phi^2(\alpha_{00} + \alpha_{10}\Phi^{-1}(u)) \begin{pmatrix} 1 \\ \\ \Phi^{-1}(u) \end{pmatrix}' \Sigma_0 \begin{pmatrix} 1 \\ \\ \Phi^{-1}(u) \end{pmatrix},$$

α_{00} and α_{10} are the true values of α_0 and α_1, respectively.

3. Simulation studies

In this section we conduct several simulation studies to (1) investigate the efficiency of the proposed estimator by comparing with the existing estimators, (2) assess the robustness of the proposed estimator against the departure from the binormal model, and (3) evaluate the accuracy of the asymptotic variance estimator of the proposed estimator in finite sample sizes.

3.1 *Efficiency*

Since a valid program for the ZH method is no longer available and their Monte Carlo computation is complicated, we will not include the ZH estimator in our simulation study but will investigate the

performance of the ZH estimator by using the estimates from Zou and Hall [16] and from Zou's dissertation in our examples. In this section we investigate the statistical efficiency of the five methods: the proposed method, CM, MHS, AP and PC methods for estimating α_0 and α_1 in the binormal model and for estimating the corresponding ROC curve by numerical studies. We use the root of mean squared error (RMSE) to measure the performance of the various estimators for α_0 and α_1 and the sum of RMSEs (SRMSE) for α_0 and α_1 as an overall performance measure. We evaluate the performance of an estimator $\widehat{ROC}(\cdot)$ for the ROC curve using the average square errors (ASE), defined by

$$\text{ASE} = \frac{1}{n_{grid}} \sum_{k=1}^{n_{grid}} \left\{ \widehat{ROC}(u_k) - ROC(u_k) \right\}^2, \tag{13}$$

where $\{u_k, k = 1, \cdots, n_{grid}\}$ are the grid points at which the functions $ROC(\cdot)$ are estimated. In the simulation studies, we choose $n_{grid} = 100$ and u_k's to be uniformly distributed over $(0, 1)$. We choose 500 simulations for each scenario. Data for non-diseased subjects are generated from the standard normal distribution, and data for diseased subjects are generated from $N(2, 1.44)$; hence the area under the ROC curve (AUC) is 0.9. We choose sizes of the diseased and non-diseased samples to be both equal and unequal numbers, $(n_0, n_1) = \{(100, 100), (200, 100), (200, 200)\}$, to investigate the effect of the sample sizes on the performance of the estimates. The results are displayed in Table 1 and Figure 1.

Table 1 gives bias, SD, RMSE and SRMSE of the resulting estimators for α_0 and α_1 by the five methods. From Table 1 we see that in all cases, our estimator has the smallest SRMSE. Specifically, our estimator consistently has smaller bias, standard error, and RMSE than the CM estimator due to a smaller number of nuisance parameters to estimate: the number of the nuisance parameters are $50.82, 68.33$ and 101.704 on average with $SD = 7.17, 8.01$ and 9.97, respectively in the proposed approach, and $200, 300$ and 400 in the CM method for the case with $(n_0, n_1) = (100, 100), (200, 100)$ and $(200, 200)$, respectively; Although the PC can have the smallest variance, its bias is also large and can even be larger than its variance, which means that the bias is significant and could not be ignored; the AP estimator is less biased than the PC estimator but has a larger variance than the PC estimator. Table 1 also show that all the estimators are improved when n_0 or n_1 increases.

Table 1 goes here

Figure 1 depicts the distribution of the estimated ASEs for the ROC curve over the 500 replications for each method. The proposed estimator and the CM estimator have comparable ASE, which is smaller

than the other methods. The performance of the estimators MHS and AP is close to that of the proposed one in this setting. The PC estimator has larger ASE than the other estimators. Further simulation study (not reported here) shows that when the accuracy of a diagnostic test is not too high or the sample size is large so that I_n can be large, the computation algorithm in the MHS method, which collapses too many jump points, will lose some information.

Figure 1 goes here

Since the number of the nuisance parameters increase with the AUC decreasing for the proposed method, the AUC can affect the comparison of the performance of the proposed approach and the CM method. Therefore, we also conduct additional simulation studies with smaller AUCs. We generate data for non-diseased subjects still from the standard nrmal distribution, but data for diseased subjects from N(1,1) and N(0.5,1), resulting in the areas under the ROC curve (AUC) of 0.76 and 0.64, respectively. Table 2 gives bias, SD, RMSE and SRMSE of the resulting estimators for α_0 and α_1 by the proposed method and the CM method.

Table 2 goes here

From the result in Table 2, we see that SRMSEs of our method are less than or equal to those from the CM; hence, our method may be a little bit better than the CM, or at a minimum, comparable to the CM for the case with low AUC.

In summary, the CM, MHS, and AP estimators have similar efficiency as the proposed estimator with the proposed estimator being slightly better in terms of SRMSEs in Tables 1 and 2. The AUC may affect the performance of the proposed approach and the CM method, because the number of different sequences with the same consecutive disease status, I_n, decreases with the AUC increasing, hence the number of the parameters in our method decreases with the AUC increasing, while the number of the parameters in the CM is still $n + 2$ regardless the AUC increasing or decreasing.

3.2 *Robustness*

In the subsection, we compare robustness of the five methods to the departure from the binormal assumption. It may be reasonable to expect a transformation to result in approximately normal data for

non-diseased subjects, but since the population of diseased subjects is often a mixture of subpopulations of subjects in different stages of the disease/infection, it seems much more reasonable to expect that transformation would result in a mixture of normals rather than a single normal for diseased subjects. So, to investigate the robustness of the binormal model, we simulate test responses of non-diseased subjects from $N(0,1)$, but test responses of diseased subjects from the mixture of two normal distributions $N(1.2, 1.2^2)$, and $N(2.2, 1.5^2)$, with the corresponding mixing proportions of 1/2 and 1/2. We set $(n_0, n_1) = \{(100, 100), (200, 200)\}$ to investigate the effect of the sample sizes on the performance of the estimates.

Figures 2(A) and 3(A) plot the average of the estimated ROC curves over the 500 replications for each method when the sample sizes are $(100, 100)$ and $(200, 200)$, respectively. The proposed estimator has the smallest ASE and hence is the most robust estimate among the five ones considered here. The CM, MHS and AP also have good robustness properties. The PC estimator has larger bias. Figures 2(B) and 3(B) depict the distribution of the ASE for the estimated ROC curves over the 500 replications for each method when the sample sizes are $(100, 100)$ and $(200, 200)$, respectively. The AP estimator has better ASE than the PC estimator, which means the AP estimator is more robust than the PC estimator.

Figures 2 and 3 go here

We also conduct numerical studies with a larger number of the components in a normal mixture. That is, we generate test results of non-diseased subjects from the standard normal distribution but test results of diseased subjects from a mixture of three normal distributions $N(1.2, 1.2^2)$, $N(2.2, 1.5^2)$ and $N(2.2, 1)$ with the corresponding mixing proportions of 1/3, 1/3, and 1/3. The results (not reported here) are similar to those in Figures 2 and 3 except that the PC estimator seems to have the largest bias and ASE, suggesting that the robustness of the PC estimator may decrease as the number of components in normal mixtures increases.

In summary, the CM, MHS, and AP estimators have similar robustness as the proposed estimators with the proposed estimator to be slightly better.

3.3 *Asymptotic inference in finite sample*

Finally, we assess the accuracy of the variance estimate formula given in Section 2 in finite sample

sizes. We investigate the performance of the variance estimate formula using the simulated data in Sections 3.1 and 3.2. Based on 500 simulated data sets, we obtain 500 estimates of $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ and their corresponding standard deviation estimates using the proposed method. From the estimates of α_0 and α_1, we form the empirical standard deviations, denoted by SD, which can be regarded as an approximation to the true standard deviations. We denote the average and the standard deviation of 500 estimated standard errors for the estimated $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ by SE_{ave} and SE_{std}, which summarize the overall performance of the standard error formula. We report our results in Tables 3 and 4 for the simulated binormal and mixture normal data in Sections 3.1 and 3.2, respectively. Our standard error estimators are very close to the "true" sample standard errors. The empirical CI coverage probabilities are close to their nominal levels 0.95.

Tables 3 and 4 go here

4. An example

We illustrate the application of our newly proposed method in an example on the accuracy of biomarkers for detecting pancreatic cancer [14]. This study examined two biomarkers, the antigenic determinant, designated as CA125, and carbohydrate antigen designated as CA19-9. The data consist of 51 measurements on subjects free of disease and 90 measurements on diseased subjects using the two biomarkers. Here, we used the two biomarkers to illustrate the application of our methodology.

Although the binormal ROC model (3) may be robust, as shown in the simulation study, it is also useful to assess whether model (3) is appropriate for the data before we make inferences on the ROC curves of the CA125 and the CA19-9 using the binormal model. Here, we present a graphical method, used by Cai and Moskowitz [8], to test model (3). Figure 4(a) and 5(a) plot the empirical ROC curve, the maximum likelihood estimate of the ROC curve and its 95% pointwise confidence intervals (denoted CI in Figure 4(a) and 5(a)) for CA19-9 and CA125, respectively, showing no obvious difference between the empirical ROC curve and the estimated ROC curve based on the binormal model. So, the binormal model is reasonable for the two biomarkers.

Figures 4 and 5 go here

Table 5 lists the estimates for the coefficients α_0 and α_1, and Figure 4(b) and Figure 5(b) plot the estimated ROC curves using the six methods for CA19-9 and CA125, respectively. Here, ZH's parameter estimates were taken from Zou and Hall [7] and Zou's Ph.D dissertation. Note that the parameters (α and β) in Zou and Hall [7] are related to our parameters, α_0 and α_1 in the following way: $\alpha_0 = \alpha/\beta, \alpha_1 = 1/\beta$.

Table 5 go here

From Table 5, we can see that the ZL, CM, AP, MHS and ZH estimates are quite similar. The PC estimator is a little different from the others. The result is consistent with the simulation, which shows that the PC estimator may be biased.

Discussion

In this paper we have proposed a semi-parametric MLE for the ROC curve under the bi-normal ROC curve model (3). The estimator is asymptotically normal. The asymptotic results also hold for the more general specification of parametric ROC curve model given by (2), for example, when G and H are symmetric distributions and when H belongs to a location-scale family. Our simulation results have indicated that the proposed estimators also have good finite-sample properties and have similar efficiency and robustness as the CM, MHS, AP and ZH estimators with the proposed estimator is slightly better than all the other estimators considered here in numerical implementation or efficiency. Same with the CM and ZH estimators, our method is rank-based, the observations can be replaced by ranks.

Hanley[15] has shown that the bi-normal ROC curve model for ordinal-scale tests enjoys a certain degree of robustness against departure from the bi-normality assumption. Our own simulation studies have also demonstrated this result. However, given limitations of any simulation study, we want to emphasize that it is important to check the assumption of the bi-normality in any application; for example one may use the graphical method suggested by Cai and Moskowitz[8] and illustrated in Section 4.

If the ROC curve of a diagnostic test depends on a subject's covariates, we need to model covariate effects on the ROC curves using a regression model. The most commonly used methods are direct

regression models of covariates on ROC curves. For example, given the vector of covariates x, we can model the effect of x on the ROC curves by the following model,

$$ROC(u, x) = G(\alpha_0 + \alpha_1' x + H(u)). \tag{14}$$

For estimating such models, both the method of estimating equations and the quasi-likelihood method have been proposed by Alonzo and Pepe [5] and Pepe and Cai [6], respectively. As shown in this paper, without any covariates, our estimator slightly outperforms these two methods in the setting we considered. Hence, it may be worthwhile to extend our methods to the setting where the ROC curve depends on subject covariates.

In the paper, the two biomarkers CA19-9 and CA125 were analyzed 'marginally', without recognition that each sample is really bivariate (two diagnostic tests on the same patients). Zou and Hall [16] and Metz, Herman and Roe [17] proposed methods to estimate ROC curve from paired samples. We will extend our method to the data with potential dependence in another paper.

ACKNOWLEDGMENTS

We like to thanks the editor, an associate editor and anonymous referees for many helpful suggestions and comments that result in an improved version of this manuscript. Zhou's research is supported by NIH R01EB005829 and NSF grant DMS 0603913. Lin's research is supported by the Fund of National Natural Science (Grant 10771148) of China. Dr. Zhou is the Director of Biostatistics Unit in Northwest HSR&D Center of Excellence, Department of Veterans Affairs Medical Center, Seattle, Washington. This paper presents the findings and conclusions of the authors. It does not necessarily represent those of VA HSR&D Service.

Appendix

We first define some additional notation that is needed to prove Conclusions 1, 2 and Theorem 1. Define $D_r = (-1, C_r)'$, $D_r^* = (-1, C_r^*)'$, $D_{r0} = (-1, C_{r0})'$, $\check{D}_r = (-1, \check{C}_r)'$ and $b_0 = \lim n_1/(n_0 + n_1)$.

A.1 *Proof of Conclusion 1*

Let $\lambda_n(\theta, g) = \frac{1}{n}\log L_n(\theta, g)$, where $L_n(\theta, g)$ is defined by (8). Denote $\mathbf{C}^* = (C_1^*, \cdots, C_{I_n^*-1}^*)'$, $C_r^* = g(Y_{(r)}^*)$, and $Y_{(1)}^* < \cdots < Y_{(I_n^*-1)}^*$ to be distinct ordered test results of $Y_{0i}, i = 1, \cdots, n_0$ and $Y_{1j}, j = 1, \cdots, n_1$. It can be shown that the MLE of θ and $\mathbf{C}^*$ must satisfy the following equations:

$$\frac{\partial \lambda_n(\theta, g)}{\partial C_r^*} = \frac{1}{n}\left(\frac{k_r^*}{\Phi(C_r^*) - \Phi(C_{r-1}^*)} - \frac{k_{r+1}^*}{\Phi(C_{r+1}^*) - \Phi(C_r^*)}\right)\phi(C_r^*)$$
$$+\frac{1}{n}\left(\frac{\ell_r^*}{\Phi(\theta' D_r^*) - \Phi(\theta' D_{r-1}^*)} - \frac{\ell_{r+1}^*}{\Phi(\theta' D_{r+1}^*) - \Phi(\theta' D_r^*)}\right)\alpha_1\phi(\theta' D_r^*) = 0, \tag{15}$$

for $1 \le r \le I_n^* - 1$, and

$$\frac{\partial \lambda_n(\theta, g)}{\partial \theta} = \frac{1}{n}\sum_{r=2}^{I_n^*-1} \ell_r^* \frac{\phi(\theta' D_r^*)D_r^* - \phi(\theta' D_{r-1}^*)D_{r-1}^*}{\Phi(\theta' D_r^*) - \Phi(\theta' D_{r-1}^*)}$$
$$+\frac{1}{n}\ell_1^* \frac{\phi(\theta' D_1^*)D_1^*}{\Phi(\theta' D_1^*)} - \frac{1}{n}\ell_{I_n^*}^* \frac{\phi(\theta' D_{I_n^*-1}^*)D_{I_n^*-1}^*}{1 - \Phi(\theta' D_{I_n^*-1}^*)} = 0, \tag{16}$$

where $k_r^* = \#\{Y_{0i} = Y_{(r)}^*, i = 1, \cdots, n_0\}$ and $\ell_r^* = \#\{Y_{1j} = Y_{(r)}^*, j = 1, \cdots, n_1\}$. If both $Y_{(r)}^*$ and $Y_{(r+1)}^*$ correspond to non-diseased subjects, then $\ell_r^* = \ell_{r+1}^* = 0$, $k_r^* < 0$ and $k_{r+1}^* < 0$. Hence, rom (15), we have $\frac{k_r^*}{\Phi(C_r^*) - \Phi(C_{r-1}^*)} = \frac{k_{r+1}^*}{\Phi(C_{r+1}^*) - \Phi(C_r^*)}$. Extending this argument to a sequence of M contiguous jump points which only involve non-diseased subjects, we have

$$\frac{k_r^*}{\Phi(C_r^*) - \Phi(C_{r-1}^*)} = \cdots = \frac{k_{r+M-1}^*}{\Phi(C_{r+M-1}^*) - \Phi(C_{r+M-2}^*)},$$

which is equal to

$$\frac{\sum_{j=r}^{r+M-1} k_j^*}{\Phi(C_{r+M-1}^*) - \Phi(C_{r-1}^*)}.$$

Similar arguments indicate that for a sequence of M contiguous jump points which only involves diseased subjects, we have

$$\frac{\ell_r^*}{\Phi(\theta' D_r^*) - \Phi(\theta' D_{r-1}^*)} = \cdots = \frac{\ell_{r+M-1}^*}{\Phi(\theta' D_{r+M-1}^*) - \Phi(\theta' D_{r+M-2}^*)},$$

which is equal to

$$\frac{\sum_{j=r}^{r+M-1} \ell_j^*}{\Phi(\theta' D_{r+M-1}^*) - \Phi(\theta' D_{r-1}^*)}.$$

Therefore, if we denote $Y_{(1)} < \cdots < Y_{(I_n-1)}$ to be the last point of contiguous jump points with same disease status, and $C_r = g(Y_{(r)}), r = 1, \cdots, I_n - 1$, for $1 \le r \le I_n - 1$, we can write (15) and (16) as

$$\frac{\partial \lambda_n(\theta, g)}{\partial C_r} = \frac{1}{n}\left(\frac{k_r}{\Phi(C_r) - \Phi(C_{r-1})} - \frac{k_{r+1}}{\Phi(C_{r+1}) - \Phi(C_r)}\right)\phi(C_r)$$
$$+\frac{1}{n}\left(\frac{\ell_r}{\Phi(\theta' D_r) - \Phi(\theta' D_{r-1})} - \frac{\ell_{r+1}}{\Phi(\theta' D_{r+1}) - \Phi(\theta' D_r)}\right)\alpha_1 \phi(\theta' D_r) = 0 \tag{17}$$

and

$$\frac{\partial \lambda_n(\theta, g)}{\partial \theta} = \frac{1}{n}\sum_{r=2}^{I_n-1} \ell_r \frac{\phi(\theta' D_r) D_r - \phi(\theta' D_{r-1}) D_{r-1}}{\Phi(\theta' D_r) - \Phi(\theta' D_{r-1})}$$
$$+\frac{1}{n}\ell_1 \frac{\phi(\theta' D_1) D_1}{\Phi(\theta' D_1)} - \frac{1}{n}\ell_{I_n} \frac{\phi(\theta' D_{I_n-1}) D_{I_n-1}}{1 - \Phi(\theta' D_{I_n-1})} = 0, \tag{18}$$

respectively, where k_r and ℓ_r are defined in Section 2. Note that (17) and (18) do not depend on the nuisance parameters $C_r^* = g(Y_{(r)}^*)$'s with $Y_{(r)}^* \in \Re$. Hence Conclusion 1 follows.

A.2 *Proof of Conclusion 2*

Let $\check{C}_1 = C_{10} + \varepsilon$ for any $\varepsilon < 0$ and $\check{C}_2, \cdots, \check{C}_{I_n-1}$ be the solution to the first $I_n - 2$ score equations in (10) given $C_1 = \check{C}_1$ and $\check{g}$ is the corresponding function for g. Define $\check{\Phi}_r = \Phi(\check{C}_r)$ for $r = 1, \cdots, I_n - 1$. Let $\check{\Phi}_{I_n}$ be the solution to the following equation:

$$\begin{aligned} G_n(x) \equiv\ & k_{I_n-1}\frac{\phi(\check{C}_{I_n-1})}{\Phi(\check{C}_{I_n-1}) - \Phi(\check{C}_{I_n-2})} - k_{I_n}\frac{\phi(\check{C}_{I_n-1})}{x - \Phi(\check{C}_{I_n-1})} \\ & +\alpha_1 \ell_{I_n-1}\frac{\phi(\theta' \check{D}_{I_n-1})}{\Phi(\theta' \check{D}_{I_n-1}) - \Phi(\theta' \check{D}_{I_n-2})} - \alpha_1 \ell_{I_n}\frac{\phi(\theta' \check{D}_{I_n-1})}{x - \Phi(\theta' \check{D}_{I_n-1})} = 0. \end{aligned}$$

Since

$$\frac{k_r}{n} - (1 - b_0)\left[\Phi(C_{r0}) - \Phi(C_{r-1,0})\right] = o_p(1) \tag{19}$$

and

$$\frac{\ell_r}{n} - b_0\left[\Phi(\theta_0' D_{r0}) - \Phi(\theta_0' D_{r-1,0})\right] = o_p(1), \tag{20}$$

for $1 \le r \le I_n$, we have

$$G_n(x) - g_n(x) = o_p(1), \tag{21}$$

where

$$\begin{aligned} g_n(x) &= (1-b_0)\phi(\check{C}_{I_n-1})\left[\frac{\Phi(C_{I_n-1,0})-\Phi(C_{I_n-2,0})}{\Phi(\check{C}_{I_n-1})-\Phi(\check{C}_{I_n-2})}-\frac{1-\Phi(C_{I_n-1,0})}{x-\Phi(\check{C}_{I_n-1})}\right] \\ &\quad +b_0\alpha_1\phi(\theta'\check{D}_{I_n-1})\left[\frac{\Phi(\theta' D_{I_n-1,0})-\Phi(\theta' D_{I_n-2,0})}{\Phi(\theta'\check{D}_{I_n-1})-\Phi(\theta'\check{D}_{I_n-2})}-\frac{1-\Phi(\theta' D_{I_n-1,0})}{x-\Phi(\theta'\check{D}_{I_n-1})}\right]. \end{aligned} \tag{22}$$

Since $G_n(\check{\Phi}_{I_n})=0$, we have

$$g_n(\check{\Phi}_{I_n}) = o_p(1). \tag{23}$$

Furthermore, we have,

$$\begin{aligned} &\frac{1}{n}\frac{\partial}{\partial C_{I_n-1}}\log\{L_n(\theta,g)\}|_{g=\check{g}} \\ =\ & (1-b_0)\phi(\check{C}_{I_n-1})\left[\frac{\Phi(C_{I_n-1,0})-\Phi(C_{I_n-2,0})}{\Phi(\check{C}_{I_n-1})-\Phi(\check{C}_{I_n-2})}-\frac{1-\Phi(C_{I_n-1,0})}{1-\Phi(\check{C}_{I_n-1})}\right] \\ & +b_0\alpha_1\phi(\theta'\check{D}_{I_n-1})\left[\frac{\Phi(\theta' D_{I_n-1,0})-\Phi(\theta' D_{I_n-2,0})}{\Phi(\theta'\check{D}_{I_n-1})-\Phi(\theta'\check{D}_{I_n-2})}-\frac{1-\Phi(\theta' D_{I_n-1,0})}{1-\Phi(\theta'\check{D}_{I_n-1})}\right]+o_p(1) \\ =\ & g_n(1)+o_p(1) = g_n(\Phi(C_{I_n,0}))+o_p(1), \end{aligned}$$

Hence, if the assumption that

$$\check{\Phi}_r \ > \ \Phi(C_{r0}) \tag{24}$$

holds for $r=I_n$, by (23) and the fact that $g_n(x)$ is a strict increasing function of x, we obtain

$$\frac{1}{n}\frac{\partial}{\partial C_{I_n-1}}\log\{L_n(\theta,g)\}|_{g=\check{g}} < 0$$

for sufficient large ε, and hence the second part of Conclusion 2 follows.

Now we prove the assumption (24) holds for $r=2,\cdots,I_n$. We use the inductive method to prove (24). The inductive method relies on I_n-1 steps. The first step consists of an conclusion for $r=2$. From (10), we see that $\check{C}_2$ satisfies

$$\begin{aligned} G_1(x) &= \frac{1}{n}\frac{\partial}{\partial C_1}\log\{\mathcal{L}_n(\theta,g)\}|_{C_1=\check{C}_1,C_2=x} \\ &= \frac{k_1}{n}\frac{\phi(\check{C}_1)}{\Phi(\check{C}_1)}-\frac{k_2}{n}\frac{\phi(\check{C}_1)}{\Phi(x)-\Phi(\check{C}_1)} \\ &\quad +\alpha_1\frac{\ell_1}{n}\frac{\phi(\theta'\check{D}_1)}{\Phi(\theta'\check{D}_1)}-\alpha_1\frac{\ell_2}{n}\frac{\phi(\theta'\check{D}_1)}{\Phi(\theta'\tilde{x})-\Phi(\theta'\check{D}_1)}=0. \end{aligned} \tag{25}$$

By (19) and (20), we have

$$G_1(x) - g_1(x) = o_p(1), \tag{26}$$

where

$$\begin{aligned} g_1(x) &= (1-b_0)\phi(\check{C}_1)\left\{\frac{\Phi(C_{10})}{\Phi(\check{C}_1)} - \frac{\Phi(C_{20})-\Phi(C_{10})}{\Phi(x)-\Phi(\check{C}_1)}\right\} \\ &\quad + b_0\alpha_1\phi(\theta'\check{D}_1)\left\{\frac{\Phi(\theta' D_{10})}{\Phi(\theta'\check{D}_1)} - \frac{\Phi(\theta' D_{20})-\Phi(\theta' D_{10})}{\Phi(\theta'\tilde{x})-\Phi(\theta'\check{D}_1)}\right\}, \end{aligned}$$

Thus by (25) and (26), we have $g_1(\check{C}_2) = o_p(1)$. Since $\check{C}_1 < C_{10}$, $g_1(C_{20}) < 0$. Note that $g_1(C_2)$ is an increasing function of C_2. Hence

$$\check{C}_2 < C_{20},$$

and (24) holds for $r = 2$.

The step j consists of showing that the inequality (24) is true for $r = j+1$ if the inequality (24) holds when $r = 1, \ldots, j$. Using the same argument as before with $r = 2$, we can prove that (24) holds for $r = j+1$ given $\check{\Phi}_r < \Phi(C_{r0}), r = 2, \cdots, j$. Hence the inequality (24) holds for $r \leq I_n$.

Using the same argument as before with $\check{C}_1 = C_{10} + \varepsilon$, we can obtain the first part of Conclusion 2.

References

1 Zhou XH., Obuchowski NA., and McClish DK. Statistical Methods in Diagnostic Medicine. *Wiley & Sons* 2002, New York, USA.

2 Hsieh F. and Turnbull BW. Nonparametric estimation of the receiver operating characteristic curve. *Ann. Statis.* 1996, **24**, 25-40.

3 Peng L, Zhou X. H. Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inferences* 2004, **118**, 129-143.

4 Metz CE., Herman BA., and Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine* 1998, **17**, 1033-1053.

5 Alonzo TA. and Pepe MS. Distribution-free ROC analysis using binary regression techniques. *Biostatistics* 2002, **3**, 421–432.

6 Pepe MS. and Cai T. The analysis of placement values for evaluating discriminatory measures. *Biometrics* 2004, **60**: 528-535.

7 Zou KH. and Hall WJ. Two transformation models for estimating an ROC curvederived from continuous data. *Journal of Applied Statistics* 2000, **5**, 621-631.

8 Cai T. and Moskowitz C.S. Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics* 2004, **5**, 573–586.

9 Kaplan E. and Meier, P. Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* 1958, **53**, 457-481.

10 Kiefer J. and Wolfowitz J. Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Annals of mathematical statistics* 1956, **27**, 887-906.

11 Kvam PH. and Samaniego FJ, Nonparametric maximum likelihood estimation based on ranked set samples. *Journal of the American Statistical Association* 1994,**89**, 526-537.

12 Li G. On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statistics & Probability Letters* 1995, **25**, 95-104.

13 Murphy SA. and Van der Vaart AW. On profile likelihood. *Journal of the American Statistical Association* 2000, **95**, 449-465.

14 Wieand S., Gail M.H., James B.R., and James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989, **76**,585-592.

15 Hanley JA. The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical Decision Making* 1998, **8**, 197-203

16 Zou KH. and Hall WJ. Semiparametric and parametric transformation models for comparing diagnostic markers with paired design. *Journal of Applied Statistics* 2002, **29**, 803-816.

17 Metz CE., Herman BA. and Roe CA. Statistical Comparison of Two ROC-curve Estimates Obtained from Partially-paired Datasets. *Medical Decision Making* 1998, **18**, 110-121.

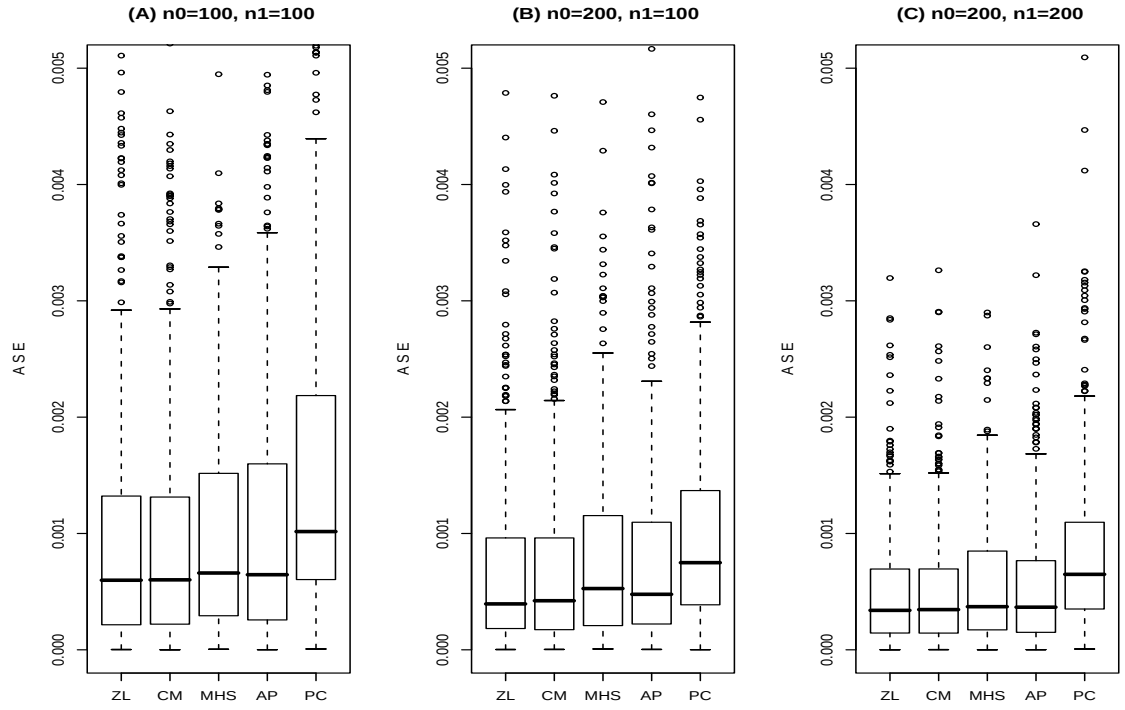

Figure 1: The distribution of ASE for the estimated ROC curve from the binormal simulated model in Section 3.1 over the 500 replications.

Table 1: Estimates of (α_0, α_1) compared with their actual values over the 500 replications for the binormal simulated data with high AUC in section 3.1.

			$\alpha_0 = 2/1.2$			$\alpha_1 = 1/1.2$			
n_0	n_1	Method	Bias	SD	RMSE	Bias	SD	RMSE	SRMSE
100	100	ZL	0.038	0.209	0.212	0.012	0.137	0.137	0.349
		CM	0.074	0.213	0.226	0.043	0.141	0.148	0.373
		MHS	0.002	0.212	0.212	-0.033	0.142	0.146	0.358
		AP	0.025	0.225	0.226	0.041	0.164	0.169	0.395
		PC	-0.105	0.169	0.199	-0.163	0.088	0.185	0.384
200	100	ZL	0.028	0.180	0.182	0.009	0.113	0.113	0.295
		CM	0.039	0.184	0.188	0.022	0.115	0.117	0.305
		MHS	0.022	0.208	0.209	-0.014	0.142	0.143	0.352
		AP	0.027	0.191	0.193	0.034	0.131	0.136	0.328
		PC	-0.105	0.142	0.177	-0.129	0.092	0.158	0.335
200	200	ZL	0.016	0.140	0.141	0.002	0.093	0.093	0.234
		CM	0.024	0.142	0.144	0.012	0.094	0.095	0.239
		MHS	-0.009	0.137	0.137	-0.014	0.102	0.103	0.240
		AP	0.009	0.144	0.145	0.017	0.103	0.104	0.249
		PC	-0.108	0.123	0.164	-0.132	0.087	0.158	0.322

Table 2: Estimates of (α_0, α_1) compared with their actual values over the 500 replications for the binormal simulated data with low AUC in section 3.1.

			$\alpha_0 = 1/1$			$\alpha_1 = 1/1$			
n_0	n_1	Method	Bias	SD	RMSE	Bias	SD	RMSE	SRMSE
100	100	ZL	0.025	0.166	0.168	0.003	0.124	0.124	0.292
		CM	0.031	0.168	0.171	0.021	0.125	0.126	0.298
200	100	ZL	0.009	0.134	0.134	0.006	0.102	0.102	0.236
		CM	0.017	0.137	0.138	0.017	0.103	0.105	0.243
200	200	ZL	0.008	0.111	0.111	-0.001	0.082	0.082	0.193
		CM	0.015	0.112	0.113	0.011	0.083	0.084	0.197
			$\alpha_0 = 0.5/1$			$\alpha_1 = 1/1$			
n_0	n_1	Method	Bias	SD	RMSE	Bias	SD	RMSE	SRMSE
100	100	ZL	0.025	0.151	0.153	0.012	0.115	0.116	0.269
		CM	0.023	0.152	0.153	0.024	0.111	0.114	0.267
200	100	ZL	0.005	0.122	0.122	0.021	0.095	0.097	0.219
		CM	0.010	0.121	0.121	0.020	0.096	0.098	0.219
200	200	ZL	0.004	0.100	0.100	0.004	0.075	0.075	0.175
		CM	0.010	0.099	0.100	0.013	0.075	0.076	0.176

Table 3: Average(SE_{ave}) and standard deviation (SE_{std}) of the standard error estimator over the 500 replications for the binormal simulated data with high AUC in Section 3.1

		$\alpha_0 = 2/1.2$			$\alpha_1 = 1/1.2$		
n_0	n_1	SD	$SE_{ave}(SE_{std})$	Coverage	SD	$SE_{ave}(SE_{std})$	Coverage
50	100	0.230	0.226(0.049)	0.939	0.160	0.161(0.043)	0.917
100	100	0.209	0.202(0.040)	0.920	0.137	0.133(0.031)	0.922
200	100	0.180	0.186(0.035)	0.931	0.113	0.113(0.024)	0.933
200	200	0.140	0.140(0.020)	0.947	0.093	0.091(0.015)	0.929

Table 4: Average(SE_{ave}) and standard deviation (SE_{std}) of the standard error estimator over the 500 replications for the mixture normal data in Section 3.2

			α_0		α_1	
n_0	n_1	k*	SD	$SE_{ave}(SE_{std})$	SD	$SE_{ave}(SE_{std})$
50	50	3	0.239	0.244(0.034)	0.159	0.167(0.036)
100	100	3	0.164	0.168(0.016)	0.111	0.113(0.018)
100	100	2	0.155	0.161(0.014)	0.097	0.105(0.015)
200	200	2	0.106	0.112(0.006)	0.068	0.071(0.007)

*where k is the number of terms in the mixture of normals for the diseased data.

Table 5: Estimates of (α_0, α_1) for CA19-9 and CA125 as diagnostic markers of pancreatic cancer

	CA19-9		CA125	
method	$\widehat{\alpha}_0(SD)$	$\widehat{\alpha}_1(SD)$	$\widehat{\alpha}_0(SD)$	$\widehat{\alpha}_1(SD)$
ZL	1.192(0.158)	0.431(0.081)	0.7277(0.1858)	1.005(0.1309)
MHS	1.177(0.160)	0.399(0.082)	0.7240(0.185)	1.001(0.137)
CM	1.235(0.129)	0.480(0.074)	0.7605(0.234)	1.0648(0.154)
AP	1.142(0.153)	0.468(0.110)	0.7594(0.251)	1.1295(0.255)
PC	1.343(0.192)	0.490(0.040)	0.6180(0.202)	0.8421(0.139)

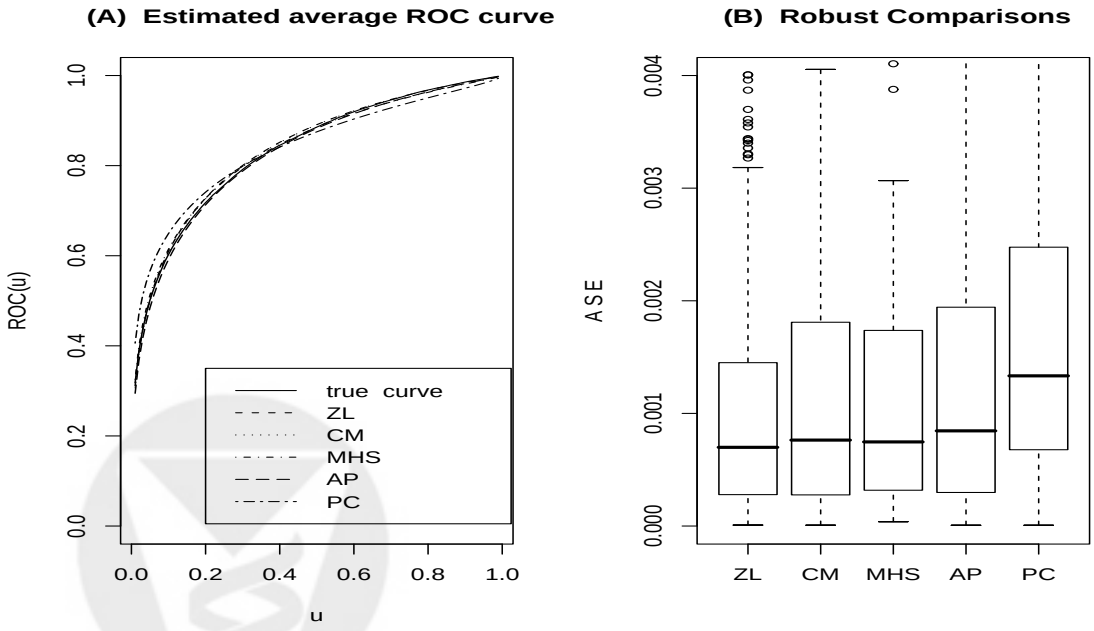

Figure 2: The diseased data are from a mixture of two normal distributions, but modeled with the binormal model when $n_0 = n_1 = 100$. (A) The average of the estimated ROC curves; (B) the distribution of ASE for the estimated ROC curves over the 500 replications.

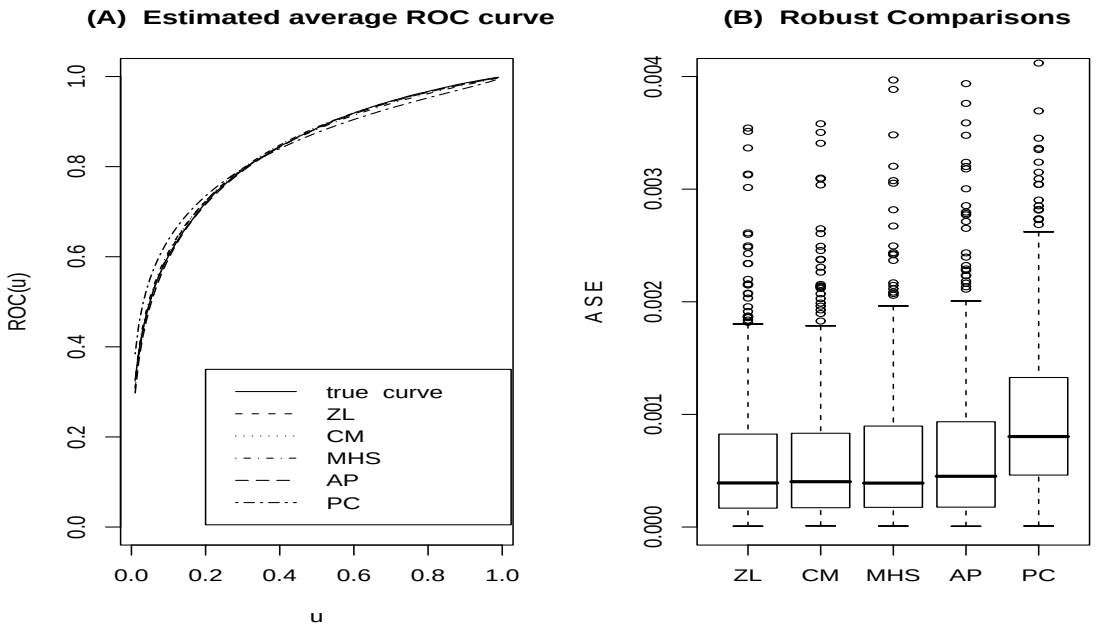

Figure 3: The diseased data are from a mixture of two normal distributions, but modeled with the binormal model when $n_0 = n_1 = 200$. (A) The average of the estimated ROC curves; (B) the distribution of ASE for the estimated ROC curves over the 500 replications.

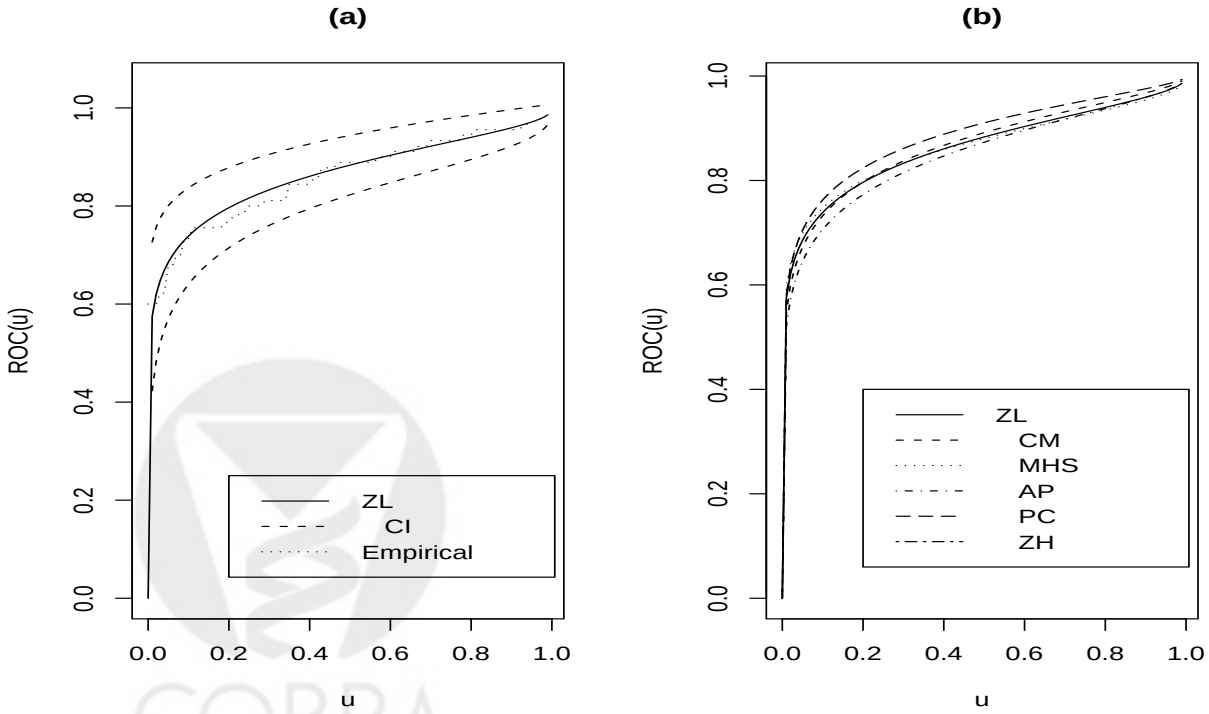

Figure 4: Estimated ROC curves of CA19-9 in the pancreatic cancer data set.

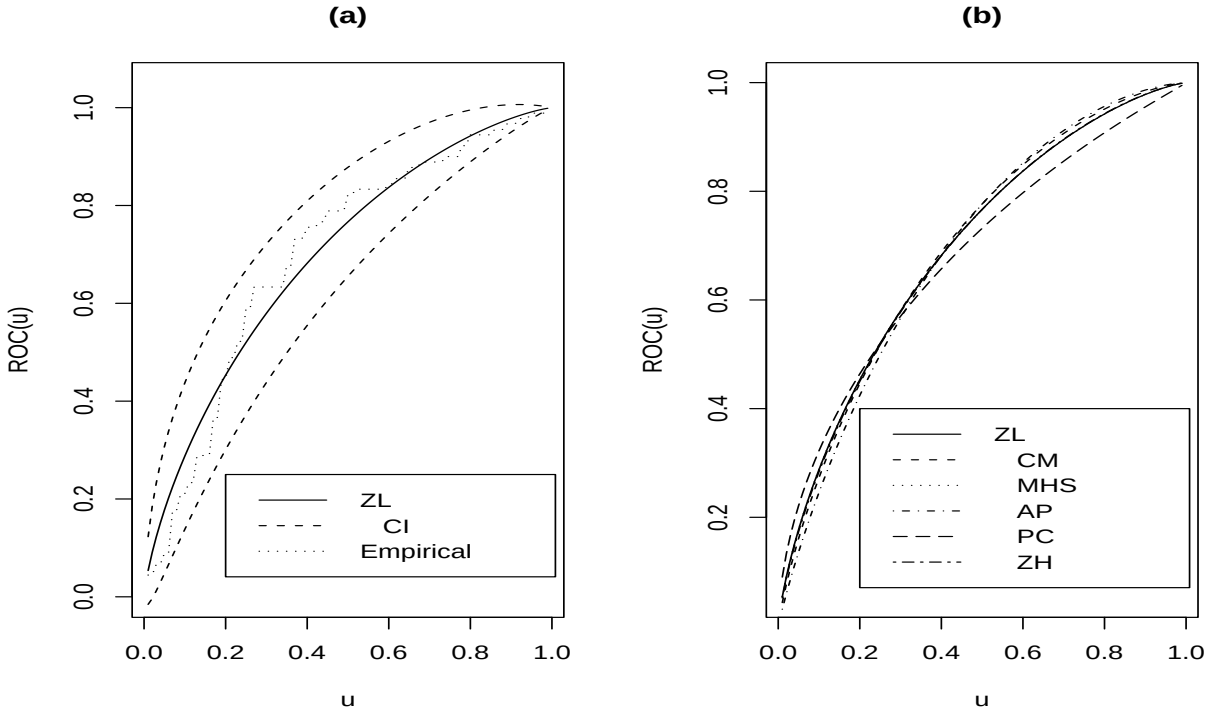

Figure 5: Estimated ROC curves of CA125 in the pancreatic cancer data set.