Published in final edited form as:

J Stat Plan Inference. 2010 November 1; 140(11): 3540–3551. doi:10.1016/j.jspi.2010.05.026.

Compare diagnostic tests using transformation-invariant smoothed ROC curves[★]

Liansheng Tang^{a,*}, Pang Du^b, and Chengqing Wu^c

^aDepartment of Statistics, George Mason University, Fairfax, VA 22030, USA

^bDepartment of Statistics, Virginia Tech, Blacksburg, VA 24061, USA

^cDepartment of Epidemiology and Public Health, Yale School of Public Health, New Haven, CT 06510, USA

Abstract

Receiver operating characteristic (ROC) curve, plotting true positive rates against false positive rates as threshold varies, is an important tool for evaluating biomarkers in diagnostic medicine studies. By definition, ROC curve is monotone increasing from 0 to 1 and is invariant to any monotone transformation of test results. And it is often a curve with certain level of smoothness when test results from the diseased and non-diseased subjects follow continuous distributions. Most existing ROC curve estimation methods do not guarantee all of these properties. One of the exceptions is Du and Tang (2009) which applies certain monotone spline regression procedure to empirical ROC estimates. However, their method does not consider the inherent correlations between empirical ROC estimates. This makes the derivation of the asymptotic properties very difficult. In this paper we propose a penalized weighted least square estimation method, which incorporates the covariance between empirical ROC estimates as a weight matrix. The resulting estimator satisfies all the aforementioned properties, and we show that it is also consistent. Then a resampling approach is used to extend our method for comparisons of two or more diagnostic tests. Our simulations show a significantly improved performance over the existing method, especially for steep ROC curves. We then apply the proposed method to a cancer diagnostic study that compares several newly developed diagnostic biomarkers to a traditional one.

Keywords

ROC curve; Smoothing spline; Bootstrap

1. Introduction

In diagnostic biomarker studies, one is concerned about whether a newly developed biomarker is more accurate than traditional biomarkers in correctly discriminating a subject with a certain disease from a subject without the disease. Biomarkers often generate discrete or continuous test results, and the receiver operating characteristic (ROC) curve is a standard statistical tool to describe and compare the accuracy of biomarkers (Zhou et al., 2002). Consider a diagnostic biomarker measured on diseased and non-diseased subjects. Let X_i be the result of the biomarker on the i th diseased subject, where $i = 1, \dots, n_d$, and Y_j be the result on the j th non-diseased subject, where $j = 1, \dots, n_{\bar{d}}$. Suppose X_i follows a distribution with

[★]Funded by: NIH and NSF.

© 2010 Elsevier B.V. All rights reserved.

^{*}Corresponding author.: ltang1@gmu.edu (L. Tang).

continuous survival function F , and Y_j follows a distribution with continuous survival function G . Then the ROC curve of the test is given by

$$R(t) = F(G^{-1}(t)) \quad \text{for } 0 \leq t \leq 1,$$

where t is false positive rate (FPR), and $G^{-1}(t) = \inf\{y : G(y) < t\}$. Clearly, $R(t)$ is monotone increasing with $R(0)=0$ and $R(1)=1$.

Many ROC-based methods exist for estimating and comparing test accuracy. A classical parametric method is the binormal model, which assumes that test results of both the diseased and non-diseased populations are normally distributed after some monotone transformation (Dorfman and Alf, 1969). Metz et al. (1998) extend the binormal model to compare two ROC curves. Binormal methods have been widely used because they give convenient maximum likelihood estimates of ROC curve parameters and their variances when the normal assumption is satisfied. However, real data may not follow normal distributions, even after transformations. If this happens, binormal methods do not provide good estimates; see, e.g., Goddard and Hinberg (1990). The issue with binormal methods has motivated non-parametric ROC methods, which are generally robust to model misspecification. Non-parametric methods for comparing diagnostic biomarkers include empirical ROC curve estimates and smooth ROC curve estimates. The empirical ROC curve is given by $\hat{R}_n(t) = \hat{R}(\hat{G}^{-1}(t))$, where $\hat{R}$ is the empirical survivor function of F and $\hat{G}^{-1}(t) = \inf\{y : \hat{G}(y) < t\}$. The properties of empirical ROC curve have been studied by Hsieh and Turnbull (1996) and Li et al. (1996). Despite its robustness, empirical ROC curve has rugged appearance, which is often inappropriate when test results are continuous.

The drawback of empirical ROC curves leads to the development of alternative non-parametric methods based on kernel smoothing estimates of density functions or distribution functions of diseased and non-diseased populations (Zou et al., 1997; Lloyd, 1998; Lloyd and Yong, 1999; Hall and Hyndman, 2003; Hall et al., 2004). However, the estimated ROC curves from these kernel smoothing methods are not transformation invariant; that is, a monotone transformation to X_i 's and Y_j 's may change the final estimate of ROC curve. To overcome this, several authors have proposed to smooth the empirical ROC curve. Peng and Zhou (2004) apply local linear smoothing methods on the empirical ROC curve by minimizing the integrated kernel weighted square error. Ren et al. (2004) use a P-spline estimator to incorporate correlation among empirical ROC estimates. However, these methods do not guarantee monotonicity and their ROC curve estimates may fall beyond the range $[0,1]$ of an ROC curve. A recent paper by Du and Tang (2009) introduces a spline approach to obtain a monotone ROC estimate which satisfies both the transformation invariant property and the boundary constraints. However, their method treats the empirical ROC values as independent observations, failing to recognize the fact that they are actually inherently correlated. This drawback may have contributed to its difficulty in estimating steep ROC curves, as evident in Du and Tang (2009, Figs. 1(a) and 2(a)). Also, the asymptotic property of their estimate was not investigated.

Most of aforementioned smoothing papers deal with the estimation of a single ROC curve. Although the estimation of a single ROC curve is important, a more common problem in practice is to compare two or more diagnostic biomarkers based on their corresponding ROC curves. Lloyd (1998) compares biomarkers by looking at the difference between kernel smoothed ROC curve estimates. He uses a bootstrap procedure to construct the pointwise confidence interval for the difference. This method inherits the aforementioned drawbacks of kernel smoothed ROC curve estimates. In some diagnostic trials data are collected on

more than two diagnostic tests. Comparing these tests becomes a multiple comparison problem and has been considered by McClish (1987) and Li and Zhou (2008). McClish (1987) uses a studentized range statistic to compare multiple AUCs, but the method is only applicable to independent ROC curves. A recent paper by Li and Zhou (2008) compares biomarkers by applying resampling-based methods to empirical ROC curves. Although their method is robust to model misspecification, the resulting estimate may have an unappealing rugged form.

In this paper, we propose a penalized weighted least square approach to estimate the ROC curve, where the covariance of empirical ROC values is incorporated as a weight matrix. Our estimate inherits all the nice properties of the ROC curve estimate by Du and Tang (2009), namely being smooth, monotone increasing, transformation invariant, and satisfying the boundary conditions. Furthermore, the incorporation of covariance among empirical ROC values allows us to show that our estimate is consistent and achieves the optimal convergence rate of a spline estimate. We then generalize the proposed method to compare multiple ROC curves. Resampling procedures are proposed to make inference on the difference curves and the differences between weighted areas under curves. Our resampling procedures are well-grounded because of the justified consistency of our ROC curve estimate. We then conduct simulation studies to demonstrate the finite-sample performance of our ROC curve estimation and multiple curve comparison procedures. Particularly, our method is significantly better at estimating steep ROC curves than the existing method. Also, the resampling procedures are shown to be efficient at detecting difference between ROC curves. We then apply the proposed techniques to the cancer diagnostic data in Goddard and Hinberg (1990), where six newly developed diagnostic kits are compared to a traditional one.

The rest of the paper is organized as follows. In Section 2 we propose the new method and present the consistency result. In Section 3 we introduce several methods for comparing two or more tests based on our smoothing method. Section 4 presents the simulations and Section 5 is the application to the real example. Some discussion is provided in Section 6.

2. ROC estimation by penalized weighted least squares

2.1. Penalized weighted least squares method

Let $t_p, p=1, \dots, n$, be the selected false positive rates (FPRs) and $\hat{R}_p(t_p)$ be the corresponding empirical ROC values. They are considered as “raw data” $\{(t_p, \hat{R}_p(t_p)) : p=1, \dots, n\}$ with the observed “responses” $\hat{R}_p(t_p), p=1, \dots, n$. To simplify the notation, we use Y_p to denote $\hat{R}_p(t_p)$ from now on. Then we transform the ROC estimation to the following correlated regression problem

$$Y_p = R_0(t_p) + \varepsilon_p, \quad (1)$$

where $R_0 = R_{w0}$ is the true ROC curve, and ε_p is the “random” error defined as the difference between the empirical and the true ROC values at t_p . The asymptotic theory of the empirical ROC in Hsieh and Turnbull (1996) then tells us that for large n_d and $n_{\bar{d}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ is approximately normal with mean zero and the covariance matrix Σ . The (p, q) -th entry of Σ is given by

$$\sum_{pq} = \frac{R_0(t_p \wedge t_q) - R_0(t_p)R_0(t_q)}{n_d} + \frac{(t_p \wedge t_q - t_p t_q)R'_0(G^{-1}(t_p))R'_0(G^{-1}(t_q))}{n_{\bar{d}}}, \quad (2)$$

where $x \wedge y$ means the minimum of x and y .

Next, we incorporate the monotonicity and the boundary conditions into the estimation of R_0 . For an ROC curve function R , consider a square integrable function w satisfying $R'' = wR'$. Then $w = (\log R')'$ becomes free of the monotonicity constraint. Solving the differential equation together with the boundary conditions $R(0)=0$ and $R(1)=1$ yields

$$R(t) = \int_0^t e^{\int_0^u w(v) dv} du \int_0^1 e^{-\int_0^u w(v) dv} dv du. \quad (3)$$

From now on, we will use R_w to denote the ROC function associated with w . To estimate $R_0 = R_{w_0}$, we first compute the estimate $\hat{w}$ of the constraint-free function w_0 and then use (3) to derive $\hat{R} = R_{\hat{w}}$.

In the view of (1), we now have a correlated non-parametric regression problem of estimating a constraint-free square integrable function w_0 . One of the many available techniques serving this purpose is the smoothing spline estimator that minimizes the following penalized weighted least squares

$$\frac{1}{n}(\mathbf{Y} - \mathbf{R}_w)^T \sum^{-1} (\mathbf{Y} - \mathbf{R}_w) + \frac{\lambda}{2} \int_0^1 [w^{(m)}(t)]^2 dt. \quad (4)$$

Here, the first term in (4) is the weighted mean square error with $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ and $\mathbf{R}_w = (R_w(t_1), \dots, R_w(t_n))^T$. It represents the goodness-of-fit of an estimate. The integral term is a roughness penalty on w and the smoothing parameter $\lambda > 0$ controls the tradeoff. A common choice of m is $m=2$ which yields the well-known cubic splines and a $\hat{w}$ that has continuous second derivative. This corresponds to an $\hat{R}$ with continuous third derivative. When such level of smoothness of $\hat{R}$ is not necessary, a lower order spline, e.g., linear spline with $m=1$, can be used instead. In practice, Σ is generally unavailable, and its consistent estimate $\hat{\Sigma}_{n_d n_{\bar{d}}}$ is used as a proxy of Σ in (4), where n_d and $n_{\bar{d}}$ are the numbers of diseased and non-diseased subjects. To obtain $\hat{\Sigma}_{n_d n_{\bar{d}}}$ we replace $R_0(\cdot)$ in (2) by its empirical estimate and the density functions involved in $R'_0(\cdot)$ by the density estimates from standard smoothing techniques.

With a selected set of spline basis functions $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_K)^T$, we can express w as $w = \boldsymbol{\varphi}^T \mathbf{c}$ with $\mathbf{c}$ being the coefficient vector. Then the problem becomes nonlinear regression for $\mathbf{c}$, which can be solved by the Gauss–Jordan procedure. Let $\mathbf{c}^{(v-1)}$ be the estimate from the previous iteration. The Gauss–Jordan procedure updates $\mathbf{c}$ via the linear equation

$$\mathbf{H}^{(v-1)}(\mathbf{c}^{(v)} - \mathbf{c}^{(v-1)}) = -\mathbf{s}^{(v-1)}, \quad (5)$$

where $\mathbf{H} = \mathbf{X}^T \Sigma^{-1} \mathbf{X} / n + \lambda \mathbf{Q}$, $\mathbf{s} = -\mathbf{X}^T \Sigma^{-1} \mathbf{r} / n + \lambda \mathbf{Q} \mathbf{c}$, $\mathbf{X}$ is $n \times K$ with the p th row $\mathbf{x}_p = R(t_p) / \mathbf{c}$, $\mathbf{Q} = \int_0^1 \boldsymbol{\varphi}(u) \boldsymbol{\varphi}^T(u) du$, and $\mathbf{r}$ is the vector of residuals with $r_p = Y_p - \hat{R}(t_p)$.

Two choices in the estimation procedure, namely the spline basis functions $\{\varphi_k, k = 1, \dots, K\}$ and the smoothing parameter λ , need some further comments. Gu (2002, Chapter 2) has details on common choices of cubic spline basis functions. In this paper, we use the scaled Bernoulli polynomials basis functions with $K=n$ knots. When n is large, a more efficient choice of $K=10n^{2/9}$ was suggested in Kim and Gu (2004). For smoothing parameter selection, a common tool is the generalized cross-validation (GCV) score introduced in Craven and Wahba (1979). Let $\Sigma = V^T V$ be the Cholesky decomposition of Σ . Some tedious algebra shows that the GCV score for our problem is

$$V(\lambda) = \frac{1}{n} \sum_{p=1}^n \left(\frac{\tilde{r}_p}{1 - a_p} \right)^2, \quad (6)$$

where $\tilde{\mathbf{r}} = V^{-T} \mathbf{r}$ with $\tilde{r}_p$ as the p -th entry, $a_p = 1/(n-1) \mathbf{x}_p^T \mathbf{H}^{-1} \mathbf{x}_p$, and $\mathbf{x}_p$, $\mathbf{H}$ and $\mathbf{r}$ are as defined in (5). The detailed derivation is similar to that in Du and Tang (2009) and thus omitted here.

2.2. Asymptotic consistency

In this section we establish the consistency and the rate of convergence for our ROC curve estimator. Let $\mathcal{H} = \{w: [0, 1] \rightarrow \mathbb{R}, \int_0^1 (w^{(m)}(t))^2 dt < \infty\}$ be the Sobolev space defining the m th order smoothing splines. The corresponding penalty function is $J(w) = \int_0^1 (w^{(m)}(t))^2 dt$. J also defines a seminorm in $\mathcal{H}$. Let $\mathcal{N}_J$ be its null space. Then $\mathcal{H}$ can be decomposed as $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$. Each $w \in \mathcal{H}$ defines a smooth and monotone increasing function $R_w: [0, 1] \rightarrow [0, 1]$ with

$$R_w(0)=0, \quad R_w(1)=1 \quad \text{and} \quad R_w(t) = \int_0^t e^{\int_0^u w(v) dv} du \int_0^1 e^{\int_0^u w(v) dv} du.$$

Let $\mathbf{t} = (t_1, \dots, t_n)^T$ be the vector of chosen FPR values. We will need the following conditions.

- A1** The true function w_0 belongs to the Sobolev space $\mathcal{H}$. Assume there exists $M > 0$ such that $|w_0| \leq M$.
- A2** Let $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_m)^T$ with $\varphi_i(t) = t^i$ be a basis of the null space $\mathcal{N}_J$ and

$$\sum_n = \frac{1}{n} \sum_{i=1}^m [V^{-T} \varphi_i(\mathbf{t})] [V^{-T} \varphi_i(\mathbf{t})]^T,$$

where V is from the Cholesky decomposition $\Sigma = V^T V$. Assume all the eigenvalues of Σ_n are bounded away from zero.

- A3** $\lambda = O_p(n^{-2m/(2m+1)})$ and $\lambda^{-1} = O_p(n^{2m/(2m+1)})$.

Conditions A1–A3 are generalizations of the conditions for penalized least squares estimation in van der Geer (2000). Condition A1 requires the true function to reside in the working function space $\mathcal{H}$. Condition A2 is assumed to guarantee an entropy bound result. Condition A3 on the smoothing parameter λ matches that in standard spline studies.

We then have the following convergence result.

Theorem 1—Let $\mathbf{t} = (t_1, \dots, t_n)^T$ be the FPR values and $R_{\hat{w}_n}$ be the ROC estimate derived from the minimizer $\hat{w}_n$ of (4). Define the weighted mean square error as

$$\|R_{\hat{w}_n} - R_{w_0}\|_{n, \Sigma}^2 = \frac{1}{n} (R_{\hat{w}_n}(\mathbf{t}) - R_{w_0}(\mathbf{t}))^T \sum_n^{-1} (R_{\hat{w}_n}(\mathbf{t}) - R_{w_0}(\mathbf{t})).$$

Then under Conditions A1–A3, we have $\mathcal{J}(\hat{w}_n) = O_p(1)$ and

$$\|R_{\hat{w}_n} - R_{w_0}\|_{n,\Sigma}^2 = O_p(n^{-2m/(2m+1)}).$$

The above theorem shows that the proposed estimate achieves the optimal rate for a spline function. In addition, $J(\hat{w}_n)$ indicates a “proper” amount of smoothness for $\hat{w}_n$. The proof of the theorem is in the Appendix.

3. Comparing two or more tests

Suppose we have measurements from L diagnostic tests. Let X_{ℓ} denote the measurement from test ℓ ($\ell = 1, \dots, L$) on the i th diseased subject, $i = 1, \dots, n_d$, and let Y_{ℓ} denote corresponding measurement on the j th healthy subject, $j = 1, \dots, n_{\bar{d}}$. Let $F_{\ell}(x)$ be the marginal survivor function for X_{ℓ} and $G_{\ell}(y)$ the marginal survivor function for Y_{ℓ} . The ROC curve for test ℓ is $R_{\ell}(t) = F_{\ell}(G_{\ell}^{-1}(t))$, whose consistent estimator $\hat{R}_{\ell}$ can be obtained by the procedure in Section 2. In this section we consider the comparison of two or more ROC curves based on estimated curves and the areas under these curves.

3.1. Comparing two tests

We first consider comparison of two tests, that is, $L=2$. Let $D(t) = R_1(t) - R_2(t)$ be the true difference between the two ROC curves and $\hat{D}(t) = \hat{R}_1(t) - \hat{R}_2(t)$ be the estimate. The consistency of $\hat{R}_1$ and $\hat{R}_2$ indicates the convergence of $\hat{D}$ to D , which then provides us the basis for using bootstrap to make inference on $D(t)$, $t \in (0, 1)$. The bootstrap procedure is as follows:

- Step 1** Sample the original data with replacement many times to obtain bootstrap resamples, $(X_{1i}^*, X_{2i}^*)_{i=1}^{n_d}$ and $(Y_{1j}^*, Y_{2j}^*)_{j=1}^{n_{\bar{d}}}$. The sampling is performed on the subjects such that the test results on the same subject are always sampled together.
- Step 2** For each resample, estimate $\hat{R}_{\ell}^*(t)$ for test ℓ using the proposed penalized weighted least squares method.
- Step 3** For each resample, compute $\hat{D}^*(t) = \hat{R}_1^*(t) - \hat{R}_2^*(t)$.
- Step 4** For each t , a $100(1-\alpha)\%$ confidence interval of $D(t)$ is then given by

$$[2\hat{D}(t) - \hat{D}^*(t)_{1-\alpha/2}, 2\hat{D}(t) - \hat{D}^*(t)_{\alpha/2}],$$

where $\hat{D}^*(t)_{\alpha/2}$ and $\hat{D}^*(t)_{1-\alpha/2}$ are, respectively, the $(\alpha/2)$ -quantile and the $(1-\alpha/2)$ -quantile of all the $\hat{D}^*(t)$'s from the bootstrap resamples. These interval limits are the basic bootstrap confidence limits defined in Davison and Hinkley (1997). Such limits are suggested to be generally more accurate than the naive one given by $(\hat{D}^*(t)_{\alpha/2}, \hat{D}^*(t)_{1-\alpha/2})$.

We can use a similar procedure to compare the tests by drawing inference on the difference between the weighted areas under curve (wAUC) of the two ROC curve estimates. The wAUC is introduced by Wieand et al. (1989) and defined as

$$\Omega_{\ell} = \int_0^1 R_{\ell}(t) dW(t), \quad (7)$$

where $W(u)$ is a probability measure on the FPR domain $(0, 1)$. Common examples of the wAUC include the area under the curve (AUC) when $W(u) = u$ for $0 < u < 1$, the partial area under the curve (pAUC) between FPRs u_1 and u_2 when $W(u) = (u - u_1)/(u_2 - u_1)$ for $0 < u_1 < u < u_2 < 1$, and the sensitivity at a given level of specificity u_0 when $W(u)$ is a point mass at u_0 . The difference between two wAUCs is then $\Delta = \Omega_2 - \Omega_1$. After the first two bootstrap steps, $\hat{\Omega}_\ell$ can be obtained by substituting $\hat{R}_\ell^*$ into (7), and the estimated difference $\hat{\Delta}^*$ is given by

$$\hat{\Delta}^* = \hat{\Omega}_2^* - \hat{\Omega}_1^*.$$

Let $SD(\hat{\Delta})$ be the sample standard deviation of $\hat{\Delta}^*$'s from all the bootstrap resamples. The test statistic for $H_0: \Omega_1 = \Omega_2$ vs. one- or two-sided alternative is given by $z = \hat{\Delta}/SD(\hat{\Delta})$. And the limits of confidence interval for Δ are $\hat{\Delta} \pm z_{1-\alpha/2}SD(\hat{\Delta})$.

3.2. Comparing more than two tests

Li and Zhou (2008) uses a resampling method to compare areas under empirical ROC curves, but the resulting estimator may have much variation which is typical in empirical curves. Alternatively, we can use a bootstrap procedure with the proposed estimator. Let $\hat{\Omega}_{\ell_1, \ell_2}$ denote the wAUC difference between test ℓ_1 and test ℓ_2 . Our bootstrap procedure for comparing multiple curves is similar to the one under two tests:

- Step 1** Sample the original data with replacement many times to obtain bootstrap resamples, $(X_{1i}^*, \dots, X_{Li}^*)_{i=1}^{n_d}$ and $(Y_{1i}^*, \dots, Y_{Li}^*)_{j=1}^{n_d}$. Again, the unit of resampling is the subject.
- Step 2** For each resample, estimate $\hat{R}_\ell^*(t)$ for test ℓ using the proposed penalized weighted least squares method.
- Step 3** For each resample, compute the pairwise wAUC difference between test ℓ_1 and test ℓ_2 : $\hat{\Delta}_{\ell_1, \ell_2}^* = \hat{\Omega}_{\ell_1}^* - \hat{\Omega}_{\ell_2}^*$.
- Step 4** Compute the standard error of $\hat{\Omega}_{\ell_1, \ell_2}$ as the sample standard deviation of all the $\hat{\Delta}_{\ell_1, \ell_2}^*$'s.
- Step 5** Use a multiple comparison procedure to obtain the best set of tests with the largest wAUCs. If the number of tests is relatively small, common multiple comparison procedures may be used. For example, with an adjusted α level, one can apply the Bonferroni procedure; with sample variance from bootstrapped samples, one can apply Fisher's least significant difference procedure and Tukey's honestly significant difference procedure. If the number of tests is large, say 1000, then the false discovery rate introduced by Benjamini and Hochberg (1995) may be applied to draw a reasonable conclusion. It might be of interest to test whether these diagnostic tests have the same partial area under their ROC curves. More generally, one wants to test hypothesis that $P'\Omega = \mathbf{0}$, where $\Omega = (\Omega_1, \dots, \Omega_L)'$ and P is a p by L full rank matrix ($p < L$). We can apply χ^2 test or F test with the covariance matrix estimated by bootstrap procedure.

4. Simulation studies

4.1. Estimating one ROC curve

We conducted large-scale simulation studies to compare the proposed penalized weighted least squares (PWLS) method with the penalized least squares (PLS) method in Du and Tang (2009) and the local linear smoothing method (LLS) in Peng and Zhou (2004). We simulated binormal, bigamma, and biexponential data with various sample sizes, $n_d = n_{\bar{d}} = 100$ and 200. X and Y were independently generated in all the settings. The binormal observations were generated from $X_j \sim N(3, 3)$ and $Y_j \sim N(0, 1)$. The bi-exponential data were generated from $X_j \sim \text{Exp}(5)$ and $Y_j \sim \text{Exp}(2)$, where $\text{Exp}(r)$ denotes the exponential distribution with density re^{-rx} . The bi-gamma data were $X_j \sim \text{Gamma}(3)$ and $Y_j \sim \text{Gamma}(2)$, where $\text{Gamma}(r)$ denotes the Gamma distribution with density $x^{r-1}e^{-x}/\Gamma(r)$. For each simulation setting we applied both methods to 1000 data replicates and computed the biases and the square roots of MSE (RMSE) of the estimated areas under curve (AUCs) and the estimated ROC values at three TPF values $t = (0.24, 0.49, 0.75)$. Table 1 presents the simulation results. It is obvious that the PWLS method outperforms the PLS and the LLS methods most of the time under various model settings.

We also plotted the average estimated ROC curves in Fig. 1 alongside the true curve. The differences between the estimated ROC curves and the true ROC curves are plotted for a better illustration in Fig. 2. Let us first look at the PWLS versus the LLS methods. In all the six settings, both the PWLS and the LLS estimates appear to follow the true ROC curves quite well. The LLS estimates tend to be slightly more accurate at FPR values very close to 0. But the PWLS estimates always do a much better job in a much wider range, especially in the middle of the curve. Also, we should mention that the individual LLS estimates for some data replicates in our simulation showed non-monotonicity, an undesired property for LLS estimate that was also evident in Peng and Zhou (2004, Fig. 1).

Next, let us compare the PWLS estimates with the PLS estimates in Figs. 1 and 2. These two methods perform equally well in the biexponential and bigamma settings where the true ROC curves are not steep. But the PWLS estimates in the binormal setting, where the true ROC curve is steep, are much better than the PLS estimates. As acknowledged in Du and Tang (2009), the PLS method has difficulty in estimating steep ROC curves and often yields estimates with severe non-concavity as FPR approaches 1. To understand why the PWLS method can rectify such problem, we first note that a steep ROC curve yields a $w = (\log R')'$ function that is relatively flat in the middle of the (0,1) range but quickly shoots towards $-\infty$ or ∞ as t goes to 0 or 1. The data generated from the corresponding distributions are likely to display such abrupt changes on the boundary as well. For the estimation of w close to the boundary, the PLS method tries to balance the flat trend shown by the data points in the middle and the abrupt changing trend shown by the data points close to the ends, and thus produces a $\hat{w}$ that fails to capture the abrupt changes on the boundary. This then leads to the obvious non-concavity in the PLS estimate of ROC curve. On the other hand, note that the diagonals of Σ_{pq} in (2) approaches zero as t_p goes to 0 or 1. Hence in a sense the PWLS method puts much more weights on the boundary data points in the estimation and forces the estimate of w to follow the abrupt changing trends on the boundary. This in turn yields a better estimate of R with significant improvement in non-concavity.

4.2. Comparing ROC curves

We simulated bivariate normal data as outcome measurements for two diagnostic tests. The bivariate normal models had the forms of $(X_1, X_2)^T \sim N\{(\mu, 1)^T, \Sigma_1\}$ and $(Y_1, Y_2)^T \sim N\{(0, 0)^T, \Sigma_1\}$, where

$$\Sigma_1 = \begin{pmatrix} 3 & \sqrt{3}\rho \\ \sqrt{3}\rho & 1 \end{pmatrix} \quad \text{with } \rho=0.5.$$

We used $\mu = 1$ or 3 . Here $\mu = 1$ gives two identical ROC curves such that the difference is $D(t) \equiv 0$ for all t , and $\mu = 3$ gives an ROC curve for test 1 on top of that for test two, with the difference between the curves $D(t) = \Phi\{\sqrt{3} + \Phi^{-1}(t)\} - \Phi\{1 + \Phi^{-1}(t)\}$. Equal numbers of diseased and healthy subjects were considered in the simulation, i.e., $n_d = n_{\bar{d}} = (100, 200, 500)$. We applied the bootstrap technique in Section 3.1 to the simulated data. The 95% confidence bands of estimated ROC curves are provided in Fig. 3. The top panels plot the results for $\mu = 3$, and the bottom panels plot the results for $\mu = 1$. As expected, the confidence bands get narrower when the sample size increases. Also the confidence bands for $\mu = 3$ show significant dominance in diagnostic accuracy of test 1 over test 2, since all the lower bounds are above 0. On the other hand, the confidence bands in the bottom panels show no significant difference between the two ROC curves throughout the range of FPR. We also simulated bi-exponential and bi-gamma data. The results were similar and thus not presented here.

5. An example

A cancer diagnostic data set has been used by Goddard and Hinberg (1990) to illustrate the drawbacks of parametric binormal ROC curves. The data collected were the measurements of seven biomarkers on the blood samples of $n_d=135$ patients with active cancer and $n_{\bar{d}}=218$ patients with inactive cancer. These biomarkers consist of a traditional biomarker, prostatic acid phosphatase level, labeled as A, and six newly developed biomarkers labeled from B to G. Neither the test results nor the log transformed results follow normal distributions. Goddard and Hinberg (1990) list several cautious notes on using binormal methods on this type of data. We used our method to estimate the ROC curves of these biomarkers, with the FPRs t_p 's chosen to be 100 equally spaced points from 0.0001 to 0.9999. The empirical ROC curves and our estimated ROC curves are plotted in Fig. 4. All the empirical ROC curves have rugged appearance, while the ROC curves fitted by our method are much smoother, and follow closely to the trends of the empirical curves. To compare the new biomarkers with the traditional biomarker, we calculated the difference curves based on the ROC estimates, and applied the bootstrap method in Section 3.1 to obtain the 95% pointwise confidence bands. Fig. 5 shows the estimated difference curves and their 95% confidence bands. It clearly indicates that the estimated ROC curve of Biomarker A is significantly different from that of Biomarker E at the upper range of FPRs. Biomarkers C and G have significantly different diagnostic accuracy from A at FPRs from around 0.5 to 0.85. The rest of new biomarkers do not have significant difference from Biomarker A as shown in the plot. To determine whether one or more new biomarkers are different from the traditional biomarker, we used the resampling method in Section 3.2 based on multiple comparison of areas under estimated ROC curves. The estimates $\hat{\Omega}_{\ell 1}$, $\ell=2, \dots, 7$, their standard deviations (SD) and two-sided p -values are listed in Table 2. At a significance level of 0.05, we can see that all the AUC differences are already insignificant even without a Bonferroni correction. Such insignificance matches the earlier result by the empirical method in Goddard and Hinberg (1990).

6. Discussion

In this paper we have proposed a new smooth ROC curve estimation method and several bootstrap procedures based on such estimation to compare two or more diagnostic tests. Our ROC curve estimate is smooth, monotone increasing, transformation invariant, and satisfies

the boundary conditions. It compares favorably to the existing ROC curve estimate by Du and Tang (2009) which also possesses all these properties. The proposed bootstrap comparison procedures also show promising results. For discrete data such as rating data in medical imaging diagnostics, empirical ROC curve estimates tend to be very roughly shaped. Thus it is often desired to have smooth ROC curve fits for underlying continuous data. This makes the proposed method a potentially suitable method for discrete data.

A tantalizing extension is to incorporate the correlation between estimated ROC curves into simultaneous estimation of multiple ROC curves. A helpful result is the asymptotic between-curve covariance presented in Tang and Zhou (2009). However, our limited experience shows that the corresponding empirical covariance estimate incorporating the between-curve covariance often results in a singular matrix and thus cannot yield good ROC curve estimates. This is partly because that the between-curve covariance estimated with the existing standard methods may not be accurate enough.

Acknowledgments

We thank an anonymous referee for the insightful comments that have significantly improved the article. We are grateful to Professor Chris J. Lloyd for kindly providing the cancer diagnostic data set. Liansheng Tang's research was supported by Award Number R15CA150698 from the National Cancer Institute under the American Recovery and Reinvestment Act of 2009. Pang Du's research is supported by NSF DMS-1007126. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

- Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*. 1995; 57(1):289–300.
- Craven P, Wahba G. Smoothing noisy data with spline functions. *Numerische Mathematik*. 1979; 31:377–403.
- Davison, AC.; Hinkley, DV. *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics. 1. Cambridge University Press; 1997 Oct.
- Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals: rating-method data. *Journal of Mathematical Psychology*. 1969; 6(3):487–496.
- Du P, Tang L. Transformation-invariant and nonparametric monotone smooth estimation of ROC curves. *Statistics in Medicine*. 2009; 28:349–359. [PubMed: 18985706]
- Goddard MJ, Hingberg I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine*. 1990; 9:325–337. [PubMed: 2343225]
- Gu, C. *Smoothing Spline ANOVA Models*. Springer-Verlag Inc; 2002.
- Hall P, Hyndman RJ. Improved methods for bandwidth selection when estimating ROC curves. *Statistics & Probability Letters*. 2003; 64:181–189.
- Hall P, Hyndman RJ, Fan Y. Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*. 2004; 91(3):743–750.
- Hsieh F, Turnbull BW. Non- & semi-parametric estimation of the receiver operating characteristics (ROC) curve. *Annals of Statistics*. 1996; 24:25–40.
- Kim YJ, Gu C. Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B*. 2004; 66:337–356.
- Li G, Tiwari R, Wells M. Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *Journal of the American Statistical Association*. 1996; 91:689–698.
- Li G, Zhou K. A unified approach to nonparametric comparison of receiver operating characteristic curves for longitudinal and clustered data. *Journal of the American Statistical Association*. 2008 Jun.103:705–713. <<http://ideas.repec.org/a/bs/jnlasa/v103y2008mjune705-713.html>>. [PubMed: 20209021]

- Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*. 1998; 93:1356–1364.
- Lloyd CJ, Yong Z. Kernel estimators of the ROC curve are better than empirical. *Statistics & Probability Letters*. 1999; 44:221–228.
- McClish DK. Comparing the areas under more than two independent ROC curves. *Medical Decision Making*. 1987; 7(3):149–155. <<http://mdm.sagepub.com/cgi/content/abstract/7/3/149>>. [PubMed: 3613915]
- Metz C, Herman B, Roe C. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. *Medical Decision Making*. Jan-Mar;1998 18(1):110–121. [PubMed: 9456215]
- Peng L, Zhou XH. Local linear smoothing of receiver operator characteristic (ROC) curves. *Journal of Statistical Planning and Inference*. 2004; 118:129–143.
- Ren H, Zhou XH, Liang H. A flexible method for estimating the ROC curve. *Journal of Applied Statistics*. 2004; 31(7):773–784.
- Tang L, Zhou XH. Semiparametric inferential procedures for comparing multivariate ROC curves with interaction terms. *Statistica Sinica*. 2009; 19:1203–1221.
- van der Geer, SA. *Empirical Processes in M-Estimation*. Cambridge University Press; 2000.
- 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.
- Zhou, XH.; McClish, DK.; Obuchowski, N. *Statistical Methods in Diagnostic Medicine*. Wiley; New York: 2002.
- Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*. 1997; 16:2143–2156. [PubMed: 9330425]

Appendix A

Proof of Theorem 1

Consider the Cholesky decomposition $\Sigma = V^T V$. Then with $\tilde{\mathbf{Y}} = V^{-T} \mathbf{Y}$ and $\tilde{\mathbf{R}}_w = V^{-T} \mathbf{R}_w$, we can transform the penalized weighted least square problem in (4) to a standard penalized least square problem. Thus, we only need to verify the asymptotic results for the minimizer $\hat{w}_n$ of the penalized least square

$$\frac{1}{n} \sum_{i=1}^n (Y_i - R_w(t_i))^2 + \frac{\lambda}{2} J(w). \quad (8)$$

Define $P_n = (1/n) \sum_{i=1}^n \delta_{t_i}$ as the empirical measure on the points $t_1, \dots, t_n$ and $\|f\|_n = (\int f^2 dP_n)^{1/2}$ as the corresponding empirical norm of a function f . By the definition of $\hat{w}_n$, we have

$$\|Y - R_{\hat{w}_n}\|_n^2 + \frac{\lambda}{2} J(\hat{w}_n) \leq \|Y - R_{w_0}\|_n^2 + \frac{\lambda}{2} J(w_0),$$

which, after rearranging the terms, yields

$$\|R_{\hat{w}_n} - R_{w_0}\|_n^2 + \frac{\lambda}{2} J(\hat{w}_n) \leq 2(Y - R_{w_0}, R_{\hat{w}_n} - R_{w_0})_n + \frac{\lambda}{2} J(w_0), \quad (9)$$

where $(\cdot, \cdot)_n$ is the inner product corresponding to the empirical norm $\|\cdot\|_n$.

Consider $\mathcal{H}_1 = \{w: [0, 1] \rightarrow \mathbb{R}\}$, $\int_0^1 (w^{(m)}(t))^2 dt < 1$. As shown in van der Geer (2000), the entropy with bracketing for $\mathcal{H}_1$ satisfies $\log N_{[]}(\varepsilon, \mathcal{H}_1, \|\cdot\|_n) \leq M_1 \varepsilon^{-1/m}$ for a fixed constant $M_1 > 0$ and all $\varepsilon > 0$ and $n \geq 1$. Then we have

$$\log N_{[]} \left(\varepsilon, \left\{ \frac{R_w - R_{w_0}}{1 + J^{1/2}(w) + J^{1/2}(w_0)} : w \in \mathcal{H}_1 \right\}, \|\cdot\|_n \right) \leq M_2 \varepsilon^{-1/m}$$

for a fixed constant $M_2 > 0$. Then applying Lemma 8.4 (with $\alpha = 1/m$) in van der Geer (2000), we have

$$|(Y - R_{w_0}, R_{\hat{w}_n} - R_{w_0})_n| = O_p(n^{-1/2}) \|R_{\hat{w}_n} - R_{w_0}\|_n^{1-1/2m} (1 + J^{1/2}(\hat{w}_n) + J^{1/2}(w_0))^{1/2m}.$$

Combining this with (9) yields

$$\|R_{\hat{w}_n} - R_{w_0}\|_n^2 + \frac{\lambda}{2} J(\hat{w}_n) \leq \frac{\lambda}{2} J(w_0) + O_p(n^{-1/2}) \|R_{\hat{w}_n} - R_{w_0}\|_n^{1-1/2m} (1 + J^{1/2}(\hat{w}_n) + J^{1/2}(w_0))^{1/2m}.$$

Solving the above equation gives us $J(\hat{w}) = O_p(1)$ and $\|R_{\hat{w}_n} - R_{w_0}\|_n = O_p(-m/(2m+1))$.

Fig. 1. Estimated ROC curves with and without covariance matrix under binormal, bi-exponential, and bi-gamma distribution: solid lines, true ROC; dashed lines, PWLS ROC curves; dotted lines, PLS ROC curves; and dash dotted lines, LLS ROC curves.

Fig. 2. Differences between the estimated ROC curves and the true ROC curves under three distributions: solid lines, PWLS ROC curve differences; dotted lines, PLS ROC curve differences; and dash dotted lines, LLS ROC curve differences.

Fig. 3. Estimated difference curves and their 95% confidence bands under binormal distribution: solid lines, true ROC difference; dotted lines, average of estimated difference curves; and shaded regions, 95% confidence bands.

Fig. 4. Empirical and smoothed ROC curves for seven biomarkers: solid lines, smooth ROC and dashed lines, empirical curves.

Fig. 5.

Difference curves between Biomarker B-G and the reference biomarker: solid lines, smooth ROC difference curves and shaded regions, 95% pointwise confidence bands.

Table 1

Bias (in %) and RMSE (in %) of ROC estimates.

<i>m</i> (<i>n</i>)	<u>Binorm.</u>			<u>Biexp.</u>			<u>Bigam.</u>					
	<i>R</i> (0.24)	<i>R</i> (0.49)	<i>R</i> (0.75)	AUC	<i>R</i> (0.24)	<i>R</i> (0.49)	<i>R</i> (0.75)	AUC	<i>R</i> (0.24)	<i>R</i> (0.49)	<i>R</i> (0.75)	AUC
<i>Penalized weighted least square method</i>												
100 Bias	0.21	0.33	-0.93	-2.13	0.74	-1.14	-0.31	-1.73	-0.64	-0.61	-0.17	1.60
RMSE	2.05	1.64	1.57	1.68	2.39	2.21	1.93	1.98	2.61	2.36	1.84	1.97
200 Bias	-0.26	0.65	-0.59	-2.09	0.74	-1.22	-0.46	-1.63	0.03	-0.17	-0.08	1.70
RMSE	1.94	1.60	1.50	1.58	2.40	2.24	1.93	1.74	2.67	2.32	1.79	1.75
<i>Penalized least square method</i>												
100 Bias	3.39	-1.07	-3.33	-2.13	1.36	-0.29	-1.19	-1.56	-0.79	-0.83	0.49	1.77
RMSE	1.98	1.39	1.93	1.68	2.46	2.08	1.99	1.95	2.62	2.26	1.87	1.98
200 Bias	3.48	-1.28	-3.69	-2.08	1.41	-0.16	-0.97	-1.63	-0.01	-0.25	0.89	1.70
RMSE	1.92	1.32	1.97	1.58	2.12	1.76	1.68	1.97	2.20	1.89	1.62	1.98
<i>Local linear smoothing method</i>												
100 Bias	-2.16	-0.47	-0.21	-2.44	-2.34	-1.48	-1.05	-1.77	-3.27	-1.49	-1.10	-1.87
RMSE	1.91	1.44	1.15	1.75	2.50	2.31	1.79	1.99	2.47	2.21	1.84	1.99
200 Bias	-2.02	-0.40	-0.14	-2.30	-2.40	-1.22	-0.80	-1.42	-2.99	-1.11	-0.71	-1.74
RMSE	1.70	1.21	0.94	1.62	2.21	1.99	1.50	1.71	2.19	1.86	1.54	1.75

Table 2

Estimated AUC differences.

	$\Delta\hat{A}$	SD	<i>p</i> -Value
B	-0.0191	0.0311	0.5394
C	-0.0430	0.0245	0.0794
D	-0.0070	0.0212	0.7418
E	-0.0348	0.0214	0.1031
F	-0.0300	0.0218	0.1691
G	-0.0347	0.0268	0.1947