6-9-2009

Nonparametric and Semiparametric Estimation of the Three Way Receiver Operating Characteristic Surface

Jialiang Li
National University of Singapore

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

Nonparametric and Semiparametric Estimation of the Three Way Receiver Operating Characteristic Surface

Jialiang Li*

Department of Statistics & Applied Probability

National University of Singapore

Singapore 117546

Xiao-Hua Zhou

Department of Biostatistics

University of Washington

Seattle, WA 98195 USA

Abstract

In many situations the diagnostic decision is not limited to a binary choice. Binary statistical tools such as receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) need to be expanded to address three-category classification problem. Previous authors have suggest various ways to model the extension of AUC but not the ROC surface. Only simple parametric approaches are proposed for modeling the ROC measure under the assumption that test results all follow normal distributions. We study the estimation methods of three dimensional ROC surfaces with nonparametric and semiparametric estimators. Asymptotical results are provided as a basis for statistical inference. Simulation studies are performed to assess the validity of our proposed methods in finite samples. We consider an Alzheimer's disease example from a clinical study in the U.S. as an illustration. The nonparametric and semiparametric modelling approaches for the three way ROC analysis can be readily generalized to diagnostic problems with more than three classes.

*Corresponding author: 6 Science Drive 2, Singapore 117546. Email: stalj@nus.edu.sg.

1

1   Introduction

Traditional ROC curve analysis relies on the existence of a binary gold-standard for the true disease status (Zhou et al. (2002), Pepe (2003)). While there are many medical tests where binary gold-standards are appropriate, there are many medical diagnostic situations in which the true disease status has more than two dimensions. For such a diagnostic task one potential solution is to dichotomize the gold-standard so that the existing methods for binary classification can be applied subsequently. Possible methods for creating this artificial binary gold-standard include dichotomization at one or two cut-off points. However, work by Obuchowski (2005) has shown that the unnatural dichotomization of non-binary gold-standards can induce a rather substantial bias in the estimation of the tests diagnostic accuracy. Dichotomizing non-binary gold-standards will bias estimated AUC values towards one, meaning that the diagnostic accuracy appears to be greater than it actually is.

There are numerous examples of medical diagnostic tests that do not easily fit into the binary disease classification structure, and many medical diagnoses appear to have a natural ordinal gold-standard of disease status. These diagnoses have an ordered gradation of illness from not diseased to seriously ill. For example, cognitive function declines from normal function, to mild impairment, to severe impairment and/or dementia. Another example would be the stage of cancer progression at the time of detection, from localized cancer through distant metastases already present. Hence, new statistical methods are needed for the assessment of diagnostic accuracy when the true disease status has an ordinal scale.

In this paper, we consider diagnostic tests which distinguish subjects each from one of three ordered classes. The accuracy of such diagnostic tests can be statistically evaluated by constructing three dimensional ROC surfaces. Scurfield (1996) brought out the mathematical definition of ROC measures for more than two categories. However, rigorous statistical inferences arrived later, motivated by sophisticated multi-category biostatistical research problems. Mossman (1999) introduced the concept of three-way ROC analysis into medical decision making. Nakas and Yiannoutsos (2004) considered the estimation of volume under the ROC surface (VUS) for ordered

2

three-class problem. Xiong et al. (2006) derived similar inference methods by using the delta method. Li and Fine (2008) further proposed the estimation of VUS for unordered classification by following the probabilistic interpretation and applied VUS as a model selection criterion in microarray study. However, none of the previous authors have discussed the estimation of three-dimensional ROC surfaces. The only relevant work we found is Heckerling (2001) with a pure parametric approach.

In this paper we aim at estimating the ROC surfaces with nonparametric and semiparametric methods. We extend the estimation procedures for comparing multivariate ROC curves (Tang and Zhou, 2009) for a two-class diagnostic testing problem to a three-class diagnostic testing problem. Theoretical results are given for the consistency and asymptotical normality of the estimators. Our paper follows a similar structure of the paper by Hsieh and Turnbull (1996) which previously addressed similar estimators for two-class ROC curves. Extending their work which has not been crystal clear for researchers to apply in multivariate tasks, our derivation will potentially benefit future scientific studies involving multi-category classifications.

We introduce the definition of three-way ROC surface in Section 2. In Section 3, we study the nonparametric estimation of ROC surface by using empirical distribution estimators. In Section 4, we consider a semiparametric model for ROC surface and propose two methods to fit the model. In Section 5, a monte carlo study is carried out to assess the performance of the nonparametric and semiparametric estimators. In Section 6, we consider a three-category classification example and apply the procedure in this paper to analyze the real data set. The extension to high dimensional ROC analysis is given in Section 7. We conclude in Section 8 with some final remarks.

2 Definition of ROC Surface

Suppose each subject is from one of three different categories. We employ a continuous diagnostic test for all subjects and record their test results. Let $\mathbf{X} = (X_1, X_2, \cdots, X_{n_x})^T$ denote test results on n_x subjects from Class I; $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_{n_y})^T$ denote test results on n_y subjects from Class

II; and $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_{n_z})^T$ denote test results on n_z subjects from Class III.

The test results X_i ($i = 1, \cdots, n_x$) are i.i.d. with distribution F; the test results Y_j ($j = 1, \cdots, n_y$) are i.i.d. with distributions G; and the test results Z_k ($k = 1, \cdots, n_z$) are i.i.d. with distributions H. F, G and H are continuous probability distributions on $\mathbb{R}$. The three kinds of test results $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ are independent to each other.

Denote the whole set of observations for the test as $\mathbf{T} = (T_1, \cdots, T_{n_x+n_y+n_z})^T = (\mathbf{X}^T, \mathbf{Y}^T, \mathbf{Z}^T)^T$. A diagnostic decision for each subject is based on the following rule by specifying two ordered decision thresholds $c_1 < c_2$:

1. IF $T_i \leq c_1$ THEN decision is "Class I";

2. ELSE IF $c_1 < T_i \leq c_2$ THEN decision is "Class II";

3. ELSE decision is "Class III".

Repeat the procedure for $i = 1, \cdots, n_x + n_y + n_z$ to make diagnostic decisions for all subjects (Nakas and Yiannoutsos, 2004).

For a pair of thresholds (c_1, c_2), we can compute the correct classification probability for the three classes,

$$e_1 = P(X_i \leq c_1) = F(c_1) \qquad \text{correct classification of Class I,}$$

$$e_2 = P(c_1 < Y_i \leq c_2) = G(c_2) - G(c_1) \quad \text{correct classification of Class II,}$$

$$e_3 = P(Z_i > c_2) = 1 - H(c_2) \qquad \text{correct classification of Class III.}$$

We can then plot (e_1, e_2, e_3) in the three-dimensional space for all possible $(c_1, c_2) \in \mathbb{R}^2$ and produce an ROC surface.

By writing the correct classification probability for the intermediate class e_2 as a function of e_1

4

and e_3, the equation for defining an ROC surface for the test may be written as

$$Q(u,v) = \begin{cases} G(H^{-1}(1-u)) - G(F^{-1}(v)) & \text{if } F^{-1}(v) \leq H^{-1}(1-u), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Thus defined, the map $(u,v) \mapsto Q(u,v)$ is monotone non-increasing for both arguments. Just as ROC curve for binary diagnostic represents the trade-off between sensitivity and specificity which are the correct classification probabilities for the two classes (diseased and healthy), ROC surface represents the three-way trade-off among the correct classification probabilities for the three classes.

Clearly, there are another two equivalent forms to define the ROC surface by using e_1 or e_3 as functions of the other two elements (Scurfield, 1996). We choose the current form for the algebraical simplicity to presenting the results. The statistical inferences remain invariant for the three choices.

In order to characterize the overall diagnostic accuracy of the test in a single summary value, we may consider the volume under the ROC surface (VUS). It is defined to be

$$\text{VUS} = \int_0^1 \int_0^1 Q(u,v) \, du \, dv. \tag{2}$$

This is a generalization of the two dimensional area under the ROC curve. Mossman (1999) has noticed that VUS is mathematically equivalent to the probability

$$P(X < Y < Z)$$

for X, Y and Z randomly selected from the three classes, respectively. For a useless test (when F, G and H are identical distributions), VUS is 1/6.
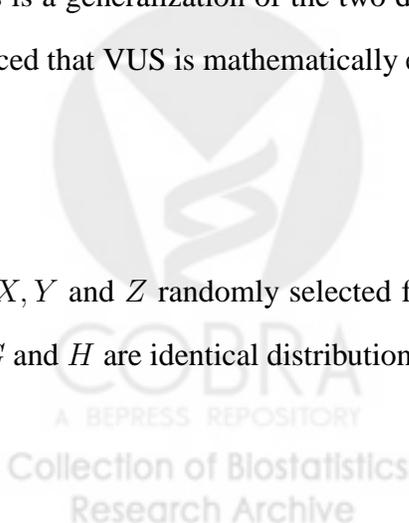
3   Nonparametric Estimation

The nonparametric estimator of ROC surface may be obtained by replacing all the distribution functions in (1) with their empirical counterparts. We construct the estimator as

$$\tilde{Q}(u,v) = \begin{cases} \hat{G}(\hat{H}^{-1}(1-u)) - \hat{G}(\hat{F}^{-1}(v)) & \text{if } \hat{F}^{-1}(v) \leq \hat{H}^{-1}(1-u), \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $\hat{F}$, $\hat{G}$, and $\hat{H}$ are the empirical distribution functions for test results from the three classes, respectively. We realized that both $\hat{G}(\hat{H}^{-1}(.))$ and $\hat{G}(\hat{F}^{-1}(.))$ are simply comparisons of the ordered observations from the two different classes.

We can easily show that the estimator (3) is uniformly consistent to the true ROC surface because we observe

$$\sup_{u,v} |\tilde{Q}(u,v) - Q(u,v)|$$
$$\leq \sup_{u} |\hat{G}(\hat{H}^{-1}(1-u)) - G(\hat{H}^{-1}(1-u))| + \sup_{u} |G(\hat{H}^{-1}(1-u)) - G(H^{-1}(1-u))|$$
$$+ \sup_{v} |\hat{G}(\hat{F}^{-1}(v)) - G(\hat{F}^{-1}(v))| + \sup_{v} |G(\hat{F}^{-1}(v)) - G(F^{-1}(v))|.$$

The first and third terms converge to zero almost surely by applying the Glivenko-Cantelli theorem (see eg. Shao (1999)). The second and fourth term converge to zero almost surely by applying the Bahadur's representation of the sample quantiles (Bahadur (1966)).

Furthermore, we have the following asymptotic distribution theorem for the nonparametric ROC estimator. Denote

$$a_n = \frac{n_x n_y n_z}{n_x n_y + n_x n_z + n_y n_z}.$$

Theorem 3.1. *Let B_i $(i = 1, 2, 3)$ be independent Brownian bridge processes. Assume F, G and H have continuous derivatives f, g and h, respectively. The partial derivative of $Q(u,v)$ is bounded on any subregion $\mathcal{T}$ of $(0,1) \times (0,1)$. As $n_x \to \infty$, $\frac{n_x n_y}{n_x n_y + n_x n_z + n_y n_z} \to \lambda_2 > 0$, $\frac{n_x n_z}{n_x n_y + n_x n_z + n_y n_z} \to \lambda_1 > 0$ and $\frac{n_y n_z}{n_x n_y + n_x n_z + n_y n_z} \to \lambda_3 > 0$. The following approximation holds*

6

asymptotically

$$\sqrt{a_n}(\tilde{Q}(u,v) - Q(u,v)) = \sqrt{\lambda_1}\big[B_1(G(H^{-1}(1-u))) - B_1(G(F^{-1}(v)))\big]$$
$$+\sqrt{\lambda_2}\frac{g(H^{-1}(1-u))}{h(H^{-1}(1-u))}B_2(1-u) + \sqrt{\lambda_3}\frac{g(F^{-1}(v))}{f(F^{-1}(v))}B_3(v).$$

The proof of this theorem is contained in the Appendix. A similar theorem for two-dimensional ROC curve has been given in Hsieh and Turnbull (1996) while the authors omitted the proof due to its lengthy presentation.

We can apply this theorem to obtain the asymptotic distribution of the estimated volume under the ROC surface which is given by

$$\widehat{\text{VUS}} = \int_0^1 \int_0^1 \tilde{Q}(u,v)\,du\,dv. \tag{4}$$

The following corollary summarizes the asymptotic distribution of $\widehat{\text{VUS}}$.

Corollary 3.2. *Assume the same conditions as in Theorem 3.1, we have*

$$\sqrt{a_n}(\widehat{\text{VUS}} - \text{VUS}) \to_d N(0, \sigma^2),$$

where

$$\sigma^2 = \lambda_1 var\Big[\int_0^1 B_1(G(H^{-1}(1-u)))du - \int_0^1 B_1(G(F^{-1}(v)))dv\Big]$$
$$+\lambda_2 var\Big[\int_0^1 \frac{g(H^{-1}(u))}{h(H^{-1}(u))}B_2(u)du\Big] + \lambda_3 var\Big[\int_0^1 \frac{g(F^{-1}(v))}{f(F^{-1}(v))}B_3(v)dv\Big]$$
$$= \lambda_1 var\Big[\int_0^1 B_1(G(H^{-1}(1-u)))du - \int_0^1 B_1(G(F^{-1}(v)))dv\Big]$$
$$+\lambda_2 var\Big[\int_0^1 B_2\big(H[G^{-1}(u)]\big)du\Big] + \lambda_3 var\Big[\int_0^1 B_3\big(F[G^{-1}(v)]\big)dv\Big]$$
$$= \lambda_1\Big[\|G(H^{-1})\|^* + \|G(F^{-1})\|^* - 2\langle G(H^{-1}), G(F^{-1})\rangle^*\Big]$$
$$+\lambda_2\|H(G^{-1})\|^* + \lambda_3\|F(G^{-1})\|^*, \tag{5}$$

7

$$\langle a, b \rangle^* = \int_0^1 \int_0^1 a(t) \wedge b(s) dt ds - \int_0^1 a(t) dt \cdot \int_0^1 b(t) dt \ \textit{and} \ \|a\|^* = \langle a, a \rangle^*.$$

The corollary provides a useful formula to calculate the variance of the estimated VUS. In practice σ^2 can be estimated by replacing all the unknown distribution functions with their empirical estimates.

By expressing (4) as a three-sample U-statistic with a kernel $1\{x < y < z\}$, where $1\{.\}$ is an indicator function, one can use the projection method (Hoeffding, 1948) to derive the asymptotic variance formula (Nakas and Yiannoutsos, 2004). The two methods are asymptotically identical. We compare the finite sample difference of these two methods in the simulation section.

4   Semiparametric Estimation

It is appealing for practitioners to assume parametric functional forms for the surface ROC curve and display a smooth ROC surface. Hsieh and Turnbull (1996), Hsieh (1995) and Li et al. (1996) considered similar estimation methodology for binary ROC curves. We consider the following parametric functional form for ROC surface.

$$Q_\Phi(u, v) \ = \ \begin{cases} \Phi\{\beta_1 + \beta_2 \Phi^{-1}(1-u)\} - \Phi\{\beta_3 + \beta_4 \Phi^{-1}(v)\} & \text{if } \beta_1 + \beta_2 \Phi^{-1}(1-u) > \beta_3 + \beta_4 \Phi^{-1}(v), \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

where Φ is the standard normal distribution and $\boldsymbol{\beta} = (\beta_1, \beta_1, \beta_3, \beta_4)^T$ are parameters which specify the characteristics of the ROC surface. This model form is motivated when F, G and H are all normal distributions. In fact, if normality assumption is correct $\boldsymbol{\beta}$ may be expressed as functions of the means and variances of the three normal distributions.

$$\beta_1 = \frac{\mu_z - \mu_y}{\sigma_y}, \qquad \beta_2 = \frac{\sigma_z}{\sigma_y}, \qquad \beta_3 = \frac{\mu_x - \mu_y}{\sigma_y}, \qquad \beta_4 = \frac{\sigma_x}{\sigma_y}. \tag{7}$$

We propose two estimation procedures which find the parameter estimates $\hat{\boldsymbol{\beta}}$ by solving the

8

following estimating equations, respectively.

$$\text{Method 1:} \quad \sum_{i=1}^{P_u}\sum_{j=1}^{P_v}\nabla Q_\Phi(u_i, v_j)\big[\tilde{Q}(u_i, v_j) - Q_\Phi(u_i, v_j)\big] = 0, \tag{8}$$

$$\text{Method 2:} \quad \sum_{i=1}^{P_u}\sum_{j=1}^{P_v}\sum_{k=1}^{n_y}\frac{\nabla Q_\Phi(u_i, v_j)\big[1\{\hat{F}^{-1}(1 - u_i) \le Y_k \le \hat{H}^{-1}(v_j)\} - Q_\Phi(u_i, v_j)\big]}{Q_\Phi(u_i, v_j)(1 - Q_\Phi(u_i, v_j))} = 0, \tag{9}$$

where $\{u_i : i = 1, \cdots, P_u\}$ and $\{v_i : i = 1, \cdots, P_v\}$ are chosen partitioning points on $(0, 1)$, $\nabla Q_\Phi(u, v)$ is the partial derivative of $Q_\Phi(u, v)$ with respect to $\boldsymbol{\beta}$ given by

$$\nabla Q_\Phi(u, v) = \begin{pmatrix} \phi(\beta_1 + \beta_2\Phi^{-1}(1 - u)) \\ \phi(\beta_1 + \beta_2\Phi^{-1}(1 - u))\Phi^{-1}(1 - u) \\ -\phi(\beta_3 + \beta_4\Phi^{-1}(v)) \\ -\phi(\beta_3 + \beta_4\Phi^{-1}(v))\Phi^{-1}(v) \end{pmatrix},$$

and ϕ is the density of the standard normal distribution. Each of the two proposed methods defines a system of four non-linear equations.

In Method 1, the empirical ROC estimator $\tilde{Q}$ is treated as the observed ROC surface. We obtain the estimating equations by differentiating the sum of squared distances between the observed and fitted ROC surfaces across a range of (u, v) values. In Method 2, we make use of the fact that the expected value of the indicator $1\{\hat{F}^{-1}(1 - u) \le Y_i \le \hat{H}^{-1}(v)\}$ converges to the true ROC surface in large sample. We thus propose the quazi-likelihood-type estimation equation. In fact, Method 1 can be regarded as a simplified version of Method 2 after we remove the weights in the denominator in (9).

The solutions for the two methods can be acquired via a numerical optimization technique such as the Newton-Raphson method.

The consistency of proposed estimators can be shown by arguing that both estimating equations are asymptotically unbiased. The proof is straightforward and is omitted in this paper. For the

9

asymptotic distribution of the parameter estimates, we have the following results.

Theorem 4.1. *In addition to conditions in Theorem 3.1, we suppose that true value $\boldsymbol{\beta}_0$ lives in a compact parameter space, the first and second order partial derivatives of $Q_\Phi(u, v)$ with respect to $\boldsymbol{\beta}$ are continuous and bounded away from zero and infinity. We have*

$$\text{Method 1:} \quad \sqrt{a_n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rightarrow_d N(0, (\Lambda^T\Lambda)^{-1}\Lambda^T\Sigma\Lambda(\Lambda^T\Lambda)^{-1});$$

$$\text{Method 2:} \quad \sqrt{a_n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rightarrow_d N(0, (\Lambda^T\mathbf{W}\Lambda)^{-1}\Lambda^T\mathbf{W}\Sigma\mathbf{W}\Lambda(\Lambda^T\mathbf{W}\Lambda)^{-1}).$$

The detailed expressions for the quantities involved in the sandwich covariance matrices will be provided in the Appendix where we present the proof of this theorem.

There are several possible methods to estimate the asymptotic covariance matrix. A "model-based" approach uses known (normal) distributions to replace F, G and H in the formula given in Appendix and simply substitutes $\boldsymbol{\beta}$ with the estimates. Alternative, we can choose to use the empirical distribution estimators. Such an empirical estimator may protect the variance estimator from model mis-specification, i.e., F, G and H being different from normal distribution. Finally we should not forget bootstrap is always one effective means to estimate the covariance of parameter estimates. We focus only on the first two approaches in the following simulations and computations.

It is noted that Method 1 is equivalent to Xiong et al. (2006) where the authors applied sampling properties of parametric estimators given by Obuchowski and McClish (1997).

5   Simulation

In order to assess the performance of our estimation procedures, we performed numerical studies using 1000 simulations. At each simulation, test scores for three categories were generated from three different normal populations: $X \sim N(1, 2.5^2)$, $Y \sim N(2, 2^2)$ and $Z \sim N(3, 4^2)$. For such

tri-normal data, the true parameters involved in model (6) are

$$\beta_1 = 0.5, \qquad \beta_2 = 2, \qquad \beta_3 = -0.5, \qquad \beta_4 = 1.25.$$

Furthermore, the true volume under this ROC surface is computed to be

$$\int\int_{z>x} \left\{ \Phi(\frac{z-\mu_y}{\sigma_y}) - \Phi(\frac{x-\mu_y}{\sigma_y}) \right\} d\Phi(\frac{z-\mu_z}{\sigma_z}) d\Phi(\frac{x-\mu_x}{\sigma_x}) = 0.325.$$

The sample sizes were designed such that $n_x = n_y = n_z = n$. We chose $n = 50$ and 100. We have also considered varying sample sizes for three classes and got similar results as presented here.

From the generated sample, we first calculated the nonparametric estimators for ROC surface. The VUS estimates, using (4), were computed at each simulation. The standard errors were estimated by using the formula in Corollary 3.2. We summarized the estimates, standard deviation of estimates and estimated standard errors in Table 1. Coverage rates of the 95% asymptotical normal confidence intervals for VUS were also reported. We compared our method with the method based on U-statistic formulation of VUS. The implementation for such a method followed Nakas and Yiannoutsos (2004). The two methods perform equally well in our simulation studies. The computation for U-statistic method was more intensive, comparing to Brownian bridge methods.

We then considered the semiparametric model and applied the two proposed model-fitting procedures to estimate the parameters. We also estimated the standard errors of the parameter estimates by using model-based (MB) and empirical (EP) methods. For the empirical method, nonparametric density estimates for f, g and h were computed by a smoothing method (Zou et al. (1997)) with the Gaussian kernel. We recorded the coverage rates of 95% asymptotical normal confidence intervals using the two kinds of standard errors. The results for fitting semiparametric models were summarized in Table 2.

The parameter estimates were all reasonably close to the true parameter values. Standard errors estimated from Method 2 in general are closer to the standard deviations of the parameter estimates than those from Method 1. Consequently the coverage rates of 95% confidence intervals

constructed from Method 2 are closer to the nominal level than those from Method 1. We also notice that the magnitude of variation of parameter estimates from Method 2 are smaller, demonstrating Method 2 produces more efficient estimates. The performance of Method 1 improves as sample size increases.

In this case, distribution assumption (normality) is satisfied for the data. Model-based and empirical standard errors are close to each other in all simulation studies. In real data analysis we suggest to report both. If discrepancy is observed we might want to trust empirical standard error.

We conducted simulation studies for data generated from Weibull and gamma distributions. The numerical results for semiparametric models exhibit substantial bias and inefficiency, compared to the results when the normality assumption is satisfied. We thus alert the users that the distribution assumption is crucial for the validity of the semiparametric model.

6 Example

In this section, we consider applying our proposed statistical methods to a practical example in human health research. All the subjects in this example were 65 years or older enrolled at the 30 NIH funded Alzheimer's Research Centers before July 2002 throughout the U.S. who died and underwent autopsy. All had taken the Mini-Mental State Examination (MMSE) within 2 years before death. The definitive examination of Alzheimer's disease (AD) was based on the extent of neuritic plaques and neurofibrillary tangles, the hallmarks of AD, at brain autopsy. The NIA/Reagan Institute criteria, based on the frequency of both plaques and tangles in the neocortex, classify patients into one of the three different diseases classes: Class I include subjects with a high likelihood of dementia being due to AD; Class II include subjects with intermediate likelihood of dementia being due to AD; and Class III include subjects with low likelihood of dementia being due to AD. Each patient underwent the test and a continuous test result was recorded thereafter. A total of 3728 results were included in this analysis where the sample size for the three individual classes are 2283, 850 and 595, respectively. See Zhou and Castellucio (2004) and Koepsell et al. (2008)

for more details about this data.

We were interested in analyzing how accurately the continuous test results are able to classify patients into the three categories. Depending on the threshold values, the test accuracy for each class may vary. In order to have an overall examination of the diagnostic performance of the test, we consider using the methods proposed in this paper to conduct the three way ROC analysis.

The nonparametric ROC surface for the test was produced in Figure 1. We can inspect the three dimensional polyhedron from different perspectives. Two panels were shown for this figure. The surface is useful to investigate the diagnostic accuracy of the test for the three classes at different threshold values. The three coordinates correspond to the correct classification rates of the three stages of AD. Researchers can then choose appropriate thresholds to ensure satisfactory classification accuracy in each dimension by using such a graphical device. The estimated volume under the ROC surface is 0.39 (95% confidence interval [0.37, 0.41]), indicating that the probability that the test correctly classifies three random subjects from the population, each from one of the three stages of AD is about 40%. Such an overall accuracy measure can then be compared to other tests with similar diagnostic goals.

It may sometimes be appealing to show a smooth surface. We thus consider using the semiparametric methods to estimate the ROC surface under normality assumption. The estimated coefficients and their standard error are reported in Table 3. Subsequently, we plot the semiparametric estimations of ROC surface in Figure 2 for the two proposed estimation equations.

The parameter estimates are similar from the model-based and empirical methods since normality is appropriate for this data and the sample size of this study is large. The ROC surfaces fitted according to the two sets of estimating equations present only slight differences. Both semiparametric estimations capture the primary characteristics of the nonparametric ROC surface. Being continuous and smooth, these graphical tools will provide helpful information and interpretation about the diagnostic performance for three classes at any thresholds. The clinical utility of the three category diagnostic test can be visually governed and supported by the displayed ROC surface.
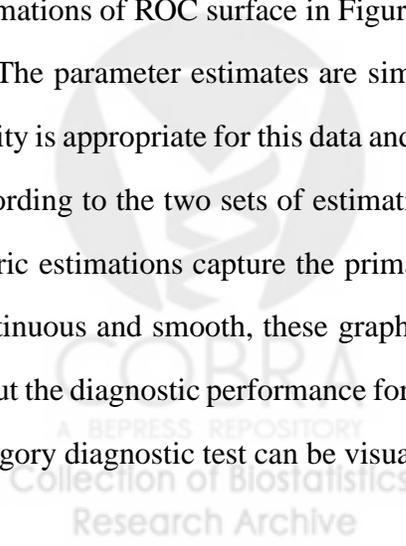
7 High dimensional ROC manifold

The methodology and theory for high dimensional ROC analysis can be similarly derived, following a similar construction. Consider the M-category classification problem where the M classes are ordered. Examples include the Glasgow Outcome Scale (GOS) in head injury, where the five categories are Good Recovery, Moderate Disability, Severe Disability, Vegetative State and Death; the Barthel Index and modified Rankin score in stroke; the Expanded Disability Status scale in multiple sclerosis as well as various measures of quality of life. Diagnosticians must assign each individual into one of the multiple classes.

Let F_m denote the distribution function of the test results for the mth class. The M-class ROC "surface" is thus an $(M-1)$-dimensional manifold with the following inductive expression:

$$ROC(\mathbf{u}) = 1 - F_M(c_{M-1}), \tag{10}$$

where $\mathbf{u} = (u_1, \cdots, u_{M-1})$ and

$$
\begin{aligned}
c_1 &= F_1^{-1}(u_1), \\
c_m &= F_m^{-1}\{F_m(c_{m-1}) + u_m\}, \qquad m = 2, \cdots, M-1.
\end{aligned}
$$

We remark that the manifold (10) is defined for $\mathbf{u} \in [0,1]^{M-1}$ such that the implicit thresholds satisfy

$$c_1 < c_2 < \cdots, c_{M-1},$$

and is equal to zero everywhere else.

We may use the empirical estimator $\hat{F}_m$ for all the distribution functions involved in definition (10) and construct the nonparametric estimator for ROC manifold.

Such an estimator $\widehat{ROC}(\mathbf{u})$ is consistent to the true ROC manifold, following the consistency of the empirical distribution estimators. One can also construct an asymptotic Brownian bridge approximation for the process $\sqrt{a_n^*}(\widehat{ROC}(\mathbf{u}) - ROC(\mathbf{u}))$ with suitably defined a_n^*. These results

14

lead to the statistical inferences for high dimensional ROC analysis.

We have to mention one more important diagnostic measure defined for multi-category classification. As a direct generalization of the area under the 2D ROC curve and the volume under the 3D ROC surface, we may consider the hypervolume under the multi-dimensional ROC manifold (HUM). Its definition is based on the following $M - 1$ integral of ROC manifold,

$$\text{HUM} = \int_0^1 \cdots \int_0^1 ROC(\mathbf{u}) \, d\,u_1 \cdots d\,u_{M-1}. \tag{11}$$

In general, larger value of HUM indicates more accurate classification probability for all classes. It is very easy to see that for a test taking continuous values HUM is equivalent to the probability

$$P(T^{(1)} < T^{(2)} < \cdots T^{(M)}),$$

where $T^{(m)}$ is the test result for a subject randomly selected from the mth class.

The estimation of HUM can be carried out by either integrating the observed ROC manifold or defining an M-sample U-statistic with the kernel function $1\{t^{(1)} < t^{(2)} < \cdots < t^{(M)}\}$. The variance and asymptotic distributions can be derived under these two routes too. In fact, Nakas and Yiannoutsos (2004) provided inference procedures for the variance of HUM by using the U-statistic approach. Their work, however, avoided the painstaking construction of multi-dimensional ROC manifold whose estimation is often more challenging than estimating the volume under ROC manifold and its variance.

We notice that ROC surface would be hard to visualize for classification dimension higher than three. More insightful constructions for high dimensional diagnostic accuracy study are needed.

8 Discussion

Nonparametric and semiparametric estimators for ROC surface are both useful tools to evaluate the accuracy of a test. Nonparametric estimator shows the observed diagnostic performance of

15

the test among three classes. It reflects the sampling characteristics. Semiparametric estimator provides a good mathematical model under assumed population distributions for the test results in three classes. We notice that the choice of normal distribution in semiparametric model may be replaced by other well-known parametric families such as Weibull or gamma, where appropriate.

We remark that in general the analysis based on three-dimensional ROC surfaces and VUSs can not be replaced with a reduced-dimension analysis, i.e., pairwise two-dimensional ROC curves and AUCs. As shown in Scurfield (1996) the pairwise AUCs can be derived from VUSs but not the other way around. Such a conclusion implies that knowing pairwise AUCs may not guide us to understand the VUS. ROC surfaces yield more important diagnostic information for three-category classification than the ROC curves which are only projections of ROC surfaces on the coordinates.

The procedure of a statistical hypothesis test to compare the accuracy of two diagnostic tests for three classes can be constructed naturally, following our proposed estimation methods. Recently Xiong et al. (2007) studied parametric methods for this goal. Nonparametric tests should be available based on the theoretical results obtained in this paper.

All graphical and numerical computations were conducted in MATLAB 7.3.0 (for additional product information see http://www.mathworks.com). The program to fit the semiparametric models is available from the authors.

Acknowledgments

Appendix

Proof of Theorem 3.1. To show that $\sqrt{a_n}(\tilde{Q}(u,v) - Q(u,v))$ converges weakly to the claimed Brownian bridge process, it suffices to show the following results.

(A) The finite-dimensional distribution $\sqrt{a_n}(\tilde{Q}(u_1, v_1) - Q(u_1, v_1), \tilde{Q}(u_2, v_2) - Q(u_2, v_2), \cdots,$ $\tilde{Q}(u_k, v_k) - Q(u_k, v_k))$ converges in distribution to multivariate normal distribution.

(B) The process $\sqrt{a_n}(\tilde{Q}(u,v) - Q(u,v))$ is tight.

For (A), without loss of generality, let us consider $k = 2$. We notice

$$
\left[\begin{array}{c} \tilde{Q}(u_1, v_1) - Q(u_1, v_1) \\ \tilde{Q}(u_2, v_2) - Q(u_2, v_2) \end{array} \right] = \left[\begin{array}{c} M_1(u_1, v_1) + M_2(u_1, v_1) + M_3(u_1, v_1) \\ M_1(u_2, v_2) + M_2(u_2, v_2) + M_3(u_2, v_2) \end{array} \right],
$$

where

$$
\begin{aligned}
M_1(u,v) &= \{\hat{G}(\hat{F}^{-1}(1-u)) - \hat{G}(\hat{H}^{-1}(v))\} - \{G(\hat{F}^{-1}(1-u)) - G(\hat{H}^{-1}(v))\} \\
M_2(u,v) &= G(\hat{F}^{-1}(1-u)) - G(F^{-1}(1-u)) \\
M_3(u,v) &= G(\hat{H}^{-1}(v)) - G(H^{-1}(v))
\end{aligned}
$$

It is straightforward to apply theorems for L-estimators (eg. Shao, 1999) to show that

$$
\sqrt{a_n} \left[\begin{array}{c} M_i(u_1, v_1) \\ M_i(u_2, v_2) \end{array} \right]
$$

converges in distribution to a bivariate normal distribution for $i = 1, 2, 3$.

Furthermore, because $E\{M_1(u,v)|M_2(u,v), M_3(u,v)\} = 0$, we have that the three bivariate normal distributions are uncorrelated (hence independent). The covariance matrix of the finite dimensional normal distribution is easily computed to be the same as the finite dimen-

17

sional version of the claimed Brownian bridge process. The tightness (B) can be verified in the same manner as in Billingsley (1999). A detailed but lengthy proof is available from the author upon request.

Proof of Theorem 4.1. Denote

$$\mathbf{Q} = (Q(u_1, v_1), Q(u_1, v_2), \cdots, Q(u_{P_u}, v_{P_v}))^T$$

and

$$\hat{\mathbf{Q}} = (\hat{Q}(u_1, v_1), \hat{Q}(u_1, v_2), \cdots, \hat{Q}(u_{P_u}, v_{P_v}))^T.$$

Both are column vectors of length $P_u P_v$.

Using the covariance structure of Brownian bridges given in Theorem 3.1, the asymptotical distribution of $\sqrt{a_n}(\hat{\mathbf{Q}} - \mathbf{Q})$ can be derived to be $P_u P_v$-dimensional multivariate normal with mean zero and the following covariance:

$$\Sigma = \lambda_1 \Sigma_1 + \lambda_2 \Sigma_2 + \lambda_3 \Sigma_3, \tag{12}$$

where the $(iP_v + j, i'P_v + j')$th element (corresponding to the pair of indices (u_i, v_j) and $(u_{i'}, v_{j'})$) of Σ_k (k=1,2,3) are given below:

$$\Sigma_1 : \quad Q(u_i, v_j) \wedge Q(u_{i'}, v_{j'}) - Q(u_i, v_j)Q(u_{i'}, v_{j'})$$

$$\Sigma_2 : \quad \frac{g(H^{-1}(1 - u_i))g(H^{-1}(1 - u_{i'}))}{h(H^{-1}(1 - u_i))h(H^{-1}(1 - u_{i'}))}(1 - u_i \vee u_{i'} - (1 - u_i)(1 - u_{i'})),$$

$$\Sigma_3 : \quad \frac{g(F^{-1}(v_j))g(F^{-1}(v_{j'}))}{f(F^{-1}(v_j))f(F^{-1}(v_{j'}))}(v_j \wedge v_{j'} - v_j v_{j'}).$$

When the semiparametric model (6) is assumed, $\mathbf{Q}(u, v) = \mathbf{Q}_\Phi(u, v)$. We then denote the

18

partial derivative of $\mathbf{Q}_\Phi$ to be

$$\Lambda = \begin{bmatrix} \nabla Q_\Phi(u_1, v_1)^T \\ \nabla Q_\Phi(u_1, v_2)^T \\ \ldots \\ \nabla Q_\Phi(u_{P_u}, v_{P_v})^T \end{bmatrix}_{P_u P_v \times 4} .$$

The left hand side of the estimating equation (8) can be written as

$$\mathcal{L}(\boldsymbol{\beta}) = \Lambda^T (\hat{\mathbf{Q}} - \mathbf{Q}).$$

The argument in Section 3 shows that each element of the vector $\mathcal{L}(\boldsymbol{\beta}_0)$ converges to zero uniformly across u_i and v_j at the true parameter $\boldsymbol{\beta}_0$.

A Taylor series expansion around the true parameter $\boldsymbol{\beta}_0$ gives the following approximation:

$$\mathcal{L}(\hat{\boldsymbol{\beta}}) \approx \mathcal{L}(\boldsymbol{\beta}_0) + \mathbf{J}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

where

$$\mathbf{J} = -\Lambda^T \Lambda = -\sum_{i=1}^{P_u} \sum_{j=1}^{P_v} \nabla Q_\Phi(u_i, v_j) \nabla Q_\Phi(u_i, v_j)^T.$$

Under the usual regularity conditions for parameter estimation $\mathbf{J}$ is positive definite with probability converging to 1, evaluated at the true parameter value. This fact thus suffices to conclude that a unique consistent solution to the estimating equation $\mathcal{L}(\boldsymbol{\beta}) = 0$ exists by applying the results in Foutz (1977).

Combining all previous results, we arrive at the claimed results

$$\begin{aligned} \sqrt{a_n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= \sqrt{a_n} \mathbf{J}^{-1} \Lambda (\hat{\mathbf{Q}} - \mathbf{Q}) \\ &\to_d N\big(0, (\Lambda^T \Lambda)^{-1} \Lambda^T \boldsymbol{\Sigma} \Lambda (\Lambda^T \Lambda)^{-1}\big). \end{aligned}$$

19

The proof of asymptotical normality for the estimators obtained from equation (9) follows a similar argument. The asymptotical covariance matrix for $\sqrt{a_n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is given by

$$(\Lambda^T \mathbf{W} \Lambda)^{-1} \Lambda^T \mathbf{W} \boldsymbol{\Sigma} \mathbf{W} \Lambda (\Lambda^T \mathbf{W} \Lambda)^{-1},$$

where $\mathbf{W} = diag\big[\{Q_\Phi(u_1, v_1)(1 - Q_\Phi(u_1, v_1))\}^{-1}, \{Q_\Phi(u_1, v_2)(1 - Q_\Phi(u_1, v_2))\}^{-1}, \cdots, \{Q_\Phi(u_{P_u}, v_{P_v})(1 - Q_\Phi(u_{P_u}, v_{P_v}))\}^{-1}\big]$. $\square$

References

[1] Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics*. **37**, 577-580.

[2] Billingsley, P. (1999). *Convergence of Probability Measures*. 2nd ed. Wiley.

[3] Foutz, R. V. (1977). On the unique consistent solution to the likelihood equation. *J. Am. Statist. Assoc.* **72**, 147-148.

[4] Heckerling, P. S. (2001). Parametric three-way receiver operating characteristic surface analysis using Mathematica. *Medical Decision Making*; **20**: 409-417.

[5] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Journal of Mathematical Statistics*. **19**: 293-325.

[6] Hsieh, F. (1995). The empirical process approach for semiparametric two-sample models with heterogeneous treatment effect. *Journal of Royal Statistical Society, Series B*. **57**: 735-748.

[7] Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*. **24**: 25-40.

[8] Koepsell T. D., Chi Y. Y., Zhou X. H., Lee W. W., Ramos E. M., and Kull W. A. (2007) An alternative method for estimating efficacy of the AN1792 Vaccine for Alzheimer's Disease. *Neurology*. **69**: 1868-1872.

[9] Li, G., Tiwari, R. C., Wells, M. T. (1996). Quantile comparison functions in two-sample problems, with application to comparison of diagnostic markers. *Journal of American Statistical Association*. **91**: 689-698.

[10] Li, J. and Fine, J. P. (2008). ROC analysis for multiple classes and multiple categories and its application in microarray study. *Biostatistics*. **9**: 566-576.

[11] Mossman, D. (1999). Three-way ROCs. *Medical Decision Making* **19**: 78-89.

[12] Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*. **23**: 3437-3449.

[13] Obuchowski, N. (2005). Estimating and comparing diagnostic tests accuracy when the gold standard is not binary. *Academic Radiology*, **12**:1198-1204.

[14] Obuchowski, N. and McClish, D. K. (1997). Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Statistics in Medicine*. **16**: 1529-1542.

[15] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Express.

[16] Scheffe, H. (1959). *The Analysis of Variance*. Jon Wiley & Sons.

[17] Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*. **40**: 253-269.

[18] Shao, J. (1999). *Mathematical Statistics*. Springer.

[19] Spizzichino, F. (2001). *Subjective Probability Models for Lifetimes*. Chapman & Hall/CRC.

[20] Tang, L. and Zhou, X.-H. (2009). Least squared based inferential procedure for comparing multivariate ROC curves. Submitted.

[21] van der Vaart, A. W. (1999). *Asymptotic Statistics*. Cambridge press.

[22] Xiong, C., van Belle, G., Miller, J. P., Morris, J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in Medicine*. **25**, 1251-1273.

[23] Xiong, C., van Belle, G., Miller, J. P., Yan, Y., Gao, F., Yu, K., Morris, J. C. (2007). A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. *Biometrical Journal*. **49**, 682-693.

[24] Zhou, X. H., Obuchowski, N. A., McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine.* Wiley: New York.

[25] Zhou X. H., Castellucio P. (2004). Adjusting for non-ignorable verification bias in clinical studies for Alzheimer's disease. *Statistics in Medicine* **23**: 221-230.

[26] Zou, K. H., Hall, W. J., Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* **16**: 2143-2156.

Table 1: Simulation Results for Estimating Volume Under the ROC Surface (VUS): Brownian bridge method follows the results in Corollary 3.2; U-statistic method follows Nakas and Yiannoutsos (2004); $\widehat{VUS}$ is the average of estimated VUS in 1000 simulations; SD is the standard deviation of the estimated VUS in 1000 simulations; SE is the average of the estimated standard errors of VUS in 1000 simulations; Coverage is the average of observed coverage rates of 95% asymptotical normal confidence intervals in 1000 simulations.

	Brownian bridge method				U-statistic method			
n	$\widehat{VUS}$	SD	SE	Coverage	$\widehat{VUS}$	SD	SE	Coverage
50	0.337	0.052	0.047	89.7%	0.320	0.051	0.043	85.4%
100	0.330	0.035	0.034	94.8%	0.324	0.031	0.032	93.0%

Table 2: Simulation Results for Estimating Semiparametric ROC Surface: ESTs are the average of 1000 parameter estimates; SDs are the standard deviations of the 1000 parameter estimates; MB SEs are the average of 1000 standard error estimates using model-based approach (percentages in the parenthesis are the observed coverage rates for 95% confidence interval); EP SEs are the average of 1000 standard error estimates using empirical approach (percentages in the parenthesis are the observed coverage rates for 95% confidence interval).

n	Parameter	Method 1				Method 2			
		EST	SD	MB SE	EP SE	EST	SD	MB SE	EP SE
50	β_1	0.47	0.39	0.28 (83%)	0.26 (80%)	0.44	0.34	0.34 (95%)	0.31 (91%)
	β_2	2.16	0.63	0.40 (89%)	0.38 (85%)	2.06	0.39	0.41 (94%)	0.40 (92%)
	β_3	-0.52	0.26	0.16 (78%)	0.16 (79%)	-0.52	0.25	0.23 (92%)	0.22 (92%)
	β_4	1.31	0.27	0.18 (85%)	0.18 (86%)	1.31	0.22	0.23 (95%)	0.23 (95%)
100	β_1	0.50	0.23	0.19 (91%)	0.18 (85%)	0.49	0.22	0.23 (97%)	0.22 (91%)
	β_2	2.03	0.36	0.26 (90%)	0.25 (89%)	1.98	0.27	0.27 (95%)	0.27 (95%)
	β_3	-0.55	0.17	0.11 (83%)	0.11(82%)	-0.52	0.17	0.16 (96%)	0.16 (94%)
	β_4	1.24	0.18	0.12 (88%)	0.13 (89%)	1.26	0.16	0.16 (95%)	0.16 (92%)

Table 3: Estimated Coefficients for Semiparametric ROC Surface: ESTs are the estimated parameters; two kinds of standard errors (SE) are calculated and listed as Model-based SE and Empirical SE.

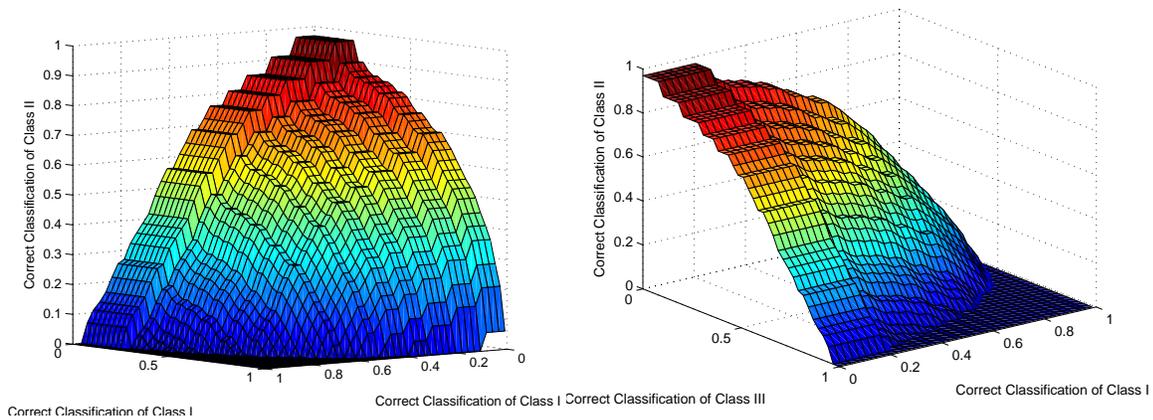| Parameter	Method 1			Method 2		
	EST	Model-based SE	Empirical SE	EST	Model-based SE	Empirical SE
β_1	0.47	0.06	0.05	0.37	0.06	0.05
β_2	1.25	0.05	0.05	1.17	0.05	0.04
β_3	-0.89	0.03	0.03	-0.90	0.03	0.03
β_4	0.91	0.03	0.03	0.88	0.03	0.03

23

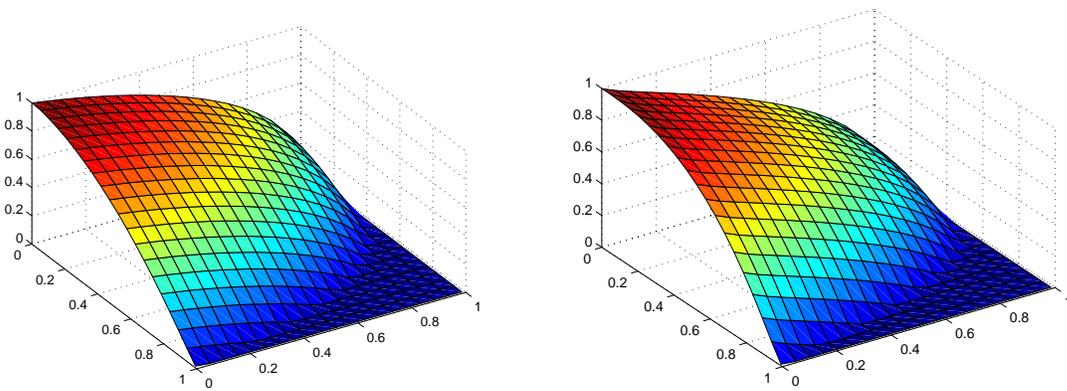Figure 1: Nonparametric estimation of ROC surface viewed from two different perspectives.

Figure 2: Semiparametric estimation of ROC surface: Estimating equation 1 (left) and Estimating equation 2(right).