Evaluating the diagnostic powers of variables and their linear combinations when the gold standard is continuous

Zhanfeng Wang[a,b], Yuan-chin Ivan Chang[b,1]

[a] *Department of Statistics and Finance, University of Science and Technology of China, Hefei, 230026, China*

[b] *Academia Sinica, Taipei, Taiwan, 11529*

Abstract

The receiver operating characteristic (ROC) curve is a very useful tool for analyzing the diagnostic/classification power of instruments/classification schemes as long as a binary-scale gold standard is available. When the gold standard is continuous and there is no confirmative threshold, ROC curve becomes less useful. Hence, there are several extensions proposed for evaluating the diagnostic potential of variables of interest. However, due to the computational difficulties of these nonparametric based extensions, they are not easy to be used for finding the optimal combination of variables to improve the individual diagnostic power. Therefore, we propose a new measure, which extends the AUC index for identifying variables with good potential to be used in a diagnostic scheme. In addition, we propose a threshold gradient descent based algorithm for finding the best linear combination of variables that maximizes this new measure, which is applicable even when the number of variables is huge. The estimate of the proposed index and its asymptotic property are studied. The performance of the proposed method is illustrated using both synthesized and real data sets.

Keywords: ROC curve, Area under curve, Gold standard, Classification

[1] Corresponding author: ycchang@stat.sinica.edu.tw

1. Introduction

The ROC curve, founded on a binary gold standard, is one of the most important tools to measure the diagnostic power of a variable or classifier, and its importance has been intensively studied by many authors, which can easily be found in the literature and textbooks such as Pepe (2003) and Krzanowski and Hand (2009). Moreover, when the number of variables is huge, many algorithms have been proposed for finding the best combination of variables to increase the individual classification accuracy (Su and Liu (1993), Pepe (2003), Ma and Huang (2005), and Wang et al. (2007a)). However, in many classification or diagnostic problems, the professed binary gold standard is essentially derived from a continuous-valued variable. If there is no such confirmative threshold for the continuous gold standard, then the evaluation of variables/classifiers according to the ROC curve based analysis may vary as the choices of thresholds change and therefore becomes less informative. For example, glycosylated hemoglobin is usually used as a primary diabetic control index, and is originally measured as a continuous-valued variable. Health institutes, such as the World Health Organization and National Institutes of Health (NIH), suggest a cutting point for it based on current findings for diabetic diagnosis and control. Once its cutting point is fixed, then the association between the variables of interests, such as new drugs, and this binary-scale standard can be evaluated using some ROC curve related analysis methods. However, as advances are made in science and medicine about this disease, this criterion will be re-evaluated and revised as necessary. Then, the performance evaluation of variables/classifiers may vary as the binary-recoding scheme is changed. It is clear that an unwarranted performance measure may result in misleading conclusions and may require re-evaluation of all the available diagnostic methods again every time a new standard is proposed. Hence, a measure that directly connects to the continuous gold standard is always preferred, which motivates our study of a new measure when the gold standard is continuous. Our goal in this paper is to find a robust measure, which is not affected by the choice of cutting point of a gold standard or how the binary outcome is derived from a continuous gold standard.

Although there are a lot of reports about the ROC curve, there is still a lack of study when the gold standard is not binary (Krzanowski and Hand, 2009). In Henkelman et al. (1990), they proposed a maximum likelihood method under ordinal scale gold standard. Recently, Zhou et al. (2005),

Choi et al. (2006), and Wang et al. (2007b) considered the ROC curve estimation problems based on some nonparametric and Bayesian approaches, when there is no gold standard. In addition, some ROC-type analysis without a binary gold standard has been considered in Obuchowski (2005) and Obuchowski (2006), where a nonparametric method is used to construct a new measure, and many other applications with continuous gold standard are discussed. However, these approaches, due to computational issue, are not easy to apply to the case that the optimal combination of variables is of interest; especially when the number of variables is large as in modern biological/genetic related studies (Waikar et al. (2009)).

In this paper, an extension of the AUC-type measure is proposed, which is independent of the choice of threshold of the continuous gold standard, and algorithms for finding the best linear combination of variables that maximizes the proposed measure are studied. Under the joint multivariate normality assumption, the algorithm for the linear combination can be founded using the LARS method. When this joint normality assumption is violated, we propose a threshold gradient descent based method (TGDM) to find the optimal linear combination. Thus, our algorithms also inherit the nice properties of LARS and TGDM when dealing with the high dimensional and variable selection problems. Numerical studies are conducted to evaluate the performances of the proposed methods with different ranges of cutting points using both synthesized and real data sets. The estimate of this novel measure and its asymptotic properties are also presented.

In the next section, we first present a novel measure for evaluating the diagnostic potential of individual variables and then an estimate of this measure. The algorithms for finding the best linear combination are discussed in Section 3. Numerical results based on the synthesized data and some real examples follow. A summary and conclusions are given in Section 4. The technical details are presented in Appendix.

2. An AUC-type Measure with a Continuous Gold Standard

Before introducing a novel AUC-type measure based on a continuous gold standard, we first fix the notation and briefly review the definition of the ROC curve and its related measures. Let Z and Y be two continuous real-valued random variables, where Z denotes the gold standard and Y is a variable of interest with diagnostic potential to be measured. Then, for example, Z is a primary index for measuring a disease and Y is some other measure of

subjects that is related to the disease of interest. In some medical diagnostics, the primary index is difficult to measure, and we are usually looking for variables that are strongly associated with Z and easy to measure, to be used as surrogates. That is why we need to evaluate the "level of association" of Y to Z. Likewise, in some bioinformatical studies, in order to develop new treatments, we would like to identify any strong associations between some genomic related factors Y to the continuous gold standard Z. Suppose that there is an unambiguous threshold c of Z that can be used to classify subjects into two subgroups, and assume further that subjects with $Z > c$ are classified as diseased, and otherwise as members of the control group. Then the ROC curve, for such a given c, is defined as $ROC(t) \equiv S_D(S_C^{-1}(t))$, where $S_D(t) = P(Y > t|Z > c)$ and $S_C(t) = P(Y > t|Z \leq c)$, and the AUC of variable Y is defined as

$$AUC(c) = P(Y_c^+ > Y_c^-) \tag{1}$$

where random variables Y_c^+ and Y_c^- respectively denote the Y-value of subjects of the disease and non-disease groups with density functions $f(y|Z > c)$ and $f(y|Z < c)$. That is, Y_c^+ and Y_c^- are random variables for the subpopulations defined by $\{Z > c\}$ and $\{Z \leq c\}$, respectively. It is clear that the $AUC(c)$ defined in (1) is a function of c, which will change as the threshold c of Z varies. Hence, when the threshold is dubious, using $AUC(c)$ as a measure may misjudge the diagnostic power of Y or the level of association between Y and Z.

Let $f_c(t)$ be a probability density function defined on the range of possible values of c, then AUC_I is defined as

$$AUC_I \equiv \int AUC(t) f_c(t) dt. \tag{2}$$

Hence, by its definition, the proposed AUC_I is independent of the choice of cutting point for the continuous gold standard, and any monotonic transformation of Y as well. This kind of threshold independent property is also one of the important properties of the ROC curve and AUC when used as measures of diagnostic performance. Since AUC_I is defined as an integration of $AUC(c)$ over the range of possible cutting points with respect to a weight function $f_c(t)$, the support of $f_c(t)$ should be chosen as a subset of the support of the density of Z. Moreover, we can use $f_c(t)$ to put different weights on all possible cutting points of Z if there is some information about the possible cutting point. If Z is an ordinal discrete variable, then there are

only countable cutting points, and $f_c(t)$ can be chosen as a probability mass function of all possible cutting points, and the integration of (2) becomes

$$AUC_I \quad = \sum_{t_i \in C} AUC(t_i) f_c(t_i), \tag{3}$$

where C is a set of all possible cutting points. In particular, when Z is binary, we can let $f_c(t)$ be a degenerated probability density, then AUC_I is the same as the original AUC.

2.1. Estimate of AUC_I

Let random variables (Y_i, Z_i) denote a pair of measures from subject i, for $i \geq 1$. Suppose that $\{(y_i, z_i),\ i = 1, \ldots, n\}$ are n independent observed values of random variables (Y_i, Z_i), $i = 1, \cdots, n$. For a given cutting point c, a subject i, $i = 1, \ldots, n$, is assigned as a "case" if $z_i > c$ and otherwise labeled as a "control". That is, for a given c, we divide the observed subjects into two groups; let $S_1(c)$ and $S_0(c)$ be the case and control groups with sample sizes n_1 and n_0, respectively. It is obvious that these assignments depend on the choice of c. Then for a fixed c, the empirical estimate of $AUC(c)$ is defined as

$$\hat{A}(c) = \frac{1}{n_0\, n_1} \sum_{i \in S_1(c);\, j \in S_0(c)} \psi(y_i - y_j), \tag{4}$$

where $\psi(u) = 1$, if $u > 0$; $= 0.5$, if $u = 0$ and $= 0$ if $u < 0$. (It is easy to see that $\hat{A}(c)$ does not exist, either $c > \max\{z_i, i = 1, \cdots, n\}$ or $c < \min\{z_i, i = 1, \cdots, n\}$, since for these two cases, we have either $n_1 = 0$ or $n_0 = 0$. Therefore, in this paper, we assume $\hat{A}(c) = 0.5$ when either one of the cases occurs.)

If the whole support of Z is considered as a possible range of cutting points, then a natural estimate of AUC_I can be defined as

$$\hat{A}_I = \int \hat{A}(t) d\hat{F}_c(t), \tag{5}$$

where $\hat{F}_c(t)$ is the empirical estimate of the cumulative distribution function of Z based on $\{z_1, \ldots, z_n\}$. However, in practice, it is rare to choose cutting points at ranges near the two ends of the distribution of Z. Thus, instead of the whole range of Z, we might explicitly define a weight function $f_c(t)$ on a particular critical range. Below, we demonstrate three possible choices: (1)

a uniform distribution over the range of $(-\hat{\sigma}, +\hat{\sigma})$, where $\hat{\sigma}$ is an empirical standard deviation of Z, say $f_1(t)$; (2) a normal density with sample mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ based on the observed values of Z, say $f_2(t)$; or (3) using a kernel density estimate, say $f_3(t)$, to approximate the marginal density of Z. For different weight functions $f_j(t)$, $j = 1, 2, 3$, the estimate of AUC_I is denoted as

$$\hat{A}_{Ij} = \int \hat{A}(t) f_j(t) dt. \tag{6}$$

It is clear that our method can be extended to other reasonable choices of weight functions. The theorem below states the strongly consistent property of $\hat{A}_{Ij}$ for all j.

Theorem 2.1. *Let $(Y \in R^1, Z \in R^1)$ be a pair of random variables with uniformly continuous marginal densities. Assume that $\{(y_1,\ z_1), \ldots, (y_n,\ z_n)\}$ are n observations of the independent and identically distributed random sample (Y_i, Z_i), $i = 1, \ldots, n$. Assume further that Z is the continuous gold standard. Then for a given $f_c(t) = f_j(t)$, $j = 1,\ 2,\ 3$, with probability one, $\hat{A}_{Ij} - AUC_{Ij} \to 0$ as $n \to \infty$, where $\hat{A}_{Ij}$ and AUC_{Ij} are defined as in (6) and(2), respectively, with corresponding $f_c(t) = f_j(t)$.*

Proof of Theorem 2.1 Since bounded function $\hat{A}(c)$ converges almost surely to $AUC(c)$ for all given c and $f_c(t)$ is also bounded density function, the proof of Theorem 2.1 follows from the dominated convergence theorem.

It is difficult to have an explicit form for the variance of $\hat{A}_{Ij}$ due to its integral form. Thus, a bootstrap estimate of the variance of $\hat{A}_{Ij}$ is used and denoted as $V(\hat{A}_{Ij})$. A similar idea is employed in Obuchowski (2006).

Remark 2.2. Note that the method for calculating (6) may depend on the choice of weight function. If the empirical density of the gold standard is used, then the computation of it is straightforward; if a kernel density of the gold standard is used, then a numerical integration method is required. However, in all cases the computation of it are easy since it is an one-dimensional density.

3. Linear combination of variables that maximizes AUC_I

For a classification or diagnostic problem, there are usually many variables measured from each subject, and it is well known that a combination

of variables can usually improve on the classification performance of a single variable. This situation motivates us to study how to find the optimal linear combination of variables that maximizes the proposed measure AUC_I. For classical AUC, Su and Liu (1993) studied the best linear combination under a multivariate normal distribution assumption. Here we extend their idea to AUC_I. In addition, we also aim to address cases with huge number of variables, which usually involve some computational issues and will be discussed later in this section.

3.1. Optimal Linear Combination of Variables Under Joint Normality

For clarity and convenience, we start with a bivariate normal distribution case, since the linear combination of variables, for a given vector of coefficients, can be treated as a single variable.

Let $U = (Y, Z)^T$ be a random vector following a bivariate normal distribution with mean vector $\mu = (\mu_1, \mu_2)^T$ and covariance matrix

$$\Sigma_U = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Suppose that $U_i = (Y_i, Z_i)^T$, $i = 1, 2$, are two independent random vectors generated from the same distribution of U. Define

$$Q_i = \exp\left(-\frac{(U_i - \mu)^T \Sigma_U^{-1} (U_i - \mu)}{2}\right), \ i = 1, 2.$$

Then for a given c,

$$\text{pr}(Y_1 > Y_2, Z_1 > c, Z_2 < c) = \int_{-\infty}^{\infty}\int_{-\infty}^{y_1}\int_{c}^{\infty}\int_{-\infty}^{c} \frac{Q_1 Q_2}{4\pi^2|\Sigma_U|}\, dz_2 dz_1 dy_2 dy_1, \quad (7)$$

where $|\Sigma_U|$ denotes the determinant of matrix Σ_U. The conditional distribution of Y_j given $Z = z_j$ is a normal distribution with mean $\tilde{\mu}_j = \mu_1 + \sigma_1/\sigma_2\rho(z_j - \mu_2)$ with $j = 1,\ 2$ and variance $\tilde{\sigma}_1^2 = (1 - \rho^2)\sigma_1^2$. Let $\eta(z_1, z_2) = 1/(2\pi\sigma_2^2)\exp(-((z_1 - \mu_2)^2 + (z_2 - \mu_2)^2)/(2\sigma_2^2))$. Then, (7) can be rewritten as

$$\begin{aligned}
&\text{pr}(Y_1 > Y_2, Z_1 > c, Z_2 < c) \\
&= \int_{c}^{\infty}\int_{-\infty}^{c} \eta(z_1, z_2) \int_{-\infty}^{\infty}\int_{-\infty}^{y_1} \frac{1}{2\pi\tilde{\sigma}_1^2} \exp\left(-\frac{(y_1 - \tilde{\mu}_1)^2 + (y_2 - \tilde{\mu}_2)^2}{2\tilde{\sigma}_1^2}\right) dy_2
\end{aligned}$$

$$dy_1 dz_2 dz_1$$

$$= \int_c^\infty \int_{-\infty}^c \eta(z_1, z_2) \mathrm{E}(\Phi(\frac{\tilde{\sigma}_1 V + \tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\sigma}_1})) dz_2 dz_1$$

$$= \int_c^\infty \int_{-\infty}^c \eta(z_1, z_2) \mathrm{E}(\Phi(V + \frac{\rho(z_1 - z_2)}{\sigma_2(1-\rho^2)^{1/2}})) dz_2 dz_1, \quad (8)$$

where V is a standard normal random variable and Φ is the standard normal cumulated distribution function. Note that under normality assumption, $\rho = 0$ implies that Y and Z are independent, and it follows from (8) $AUC_I = 0.5$ in this case.

Now, suppose that $\tilde{X} = (X_1, \ldots, X_p)^T$ is a p-dimensional random vector of measures of a subject, and Z is the continuous gold standard as before. Suppose $l \in R^p$ and let $Y = l^T \tilde{X}$ be a linear combination of $\tilde{X}$. Assume further that $\tilde{X}$ follows a multivariate normal distribution with mean vector μ^* and covariance matrix Σ. Then Y follows a normal distribution with mean $\mu_1 = l^T \mu^*$ and variance $\sigma_1^2 = l^T \Sigma l$. The correlation coefficient between Y and Z is $\rho = l^T \mathrm{cov}(\tilde{X}, Z)/((l^T \Sigma l)^{1/2} \sigma_2)$, where $\mathrm{cov}(\tilde{X}, Z) = (\mathrm{cov}(X_1, Z), \ldots, \mathrm{cov}(X_p, Z))^T$. Then, AUC_I for such a linear combination of X_i's, $Y = l^T \tilde{X}$, is a function of l:

$$AUC_I(l) \;=\; \int \mathrm{pr}(l^T \tilde{X}_1 > l^T \tilde{X}_2 | Z_1 > t, Z_2 < t) f_c(t) dt \quad (9)$$

where $(\tilde{X}_i^T, Z_i)^T$, $i = 1, 2$, are independent identically distributed samples of $(\tilde{X}^T, Z)^T$. Our goal is to find the optimal linear combination of $X_1, \ldots, X_p$ such that AUC_I is maximized and it is known that AUC is scale invariant. In order to make the solution identifiable, we search for an l_{opt} such that $AUC_I(l_{opt}) \geq AUC_I(l)$ for all possible $l \in R^p$ with $\|l\| = 1$.

From (8),

$$\frac{\partial}{\partial l} \mathrm{E}\left(\Phi\left(V + \frac{\rho(z_1 - z_2)}{\sigma_2(1-\rho^2)^{1/2}}\right)\right) = \frac{1}{\sqrt{2}} \exp\left(-\frac{\rho^2(z_1 - z_2)^2}{4\sigma_2^2(1-\rho^2)}\right) \frac{z_1 - z_2}{\sigma_2(1-\rho^2)^{3/2}} \frac{\partial \rho}{\partial l}. \quad (10)$$

Therefore,

$$\begin{aligned} \frac{\partial AUC_I(l)}{\partial l} &= \frac{\partial \rho}{\partial l} \int f_c(t) \int_t^\infty \int_{-\infty}^t \frac{1}{2^{3/2}\pi\sigma_2^2} \exp\left(-\frac{(z_1 - \mu_2)^2 + (z_2 - \mu_2)^2}{2\sigma_2^2}\right) \\ &\quad \exp\left(-\frac{\rho^2(z_1 - z_2)^2}{4\sigma_2^2(1-\rho^2)}\right) \frac{z_1 - z_2}{\sigma_2(1-\rho^2)^{3/2}} \frac{1}{\mathrm{pr}(Z_1 > t, Z_2 < t)} dz_2 dz_1 dt \\ &= \frac{\partial \rho}{\partial l} \Delta, \end{aligned} \quad (11)$$

where Δ dentes the integration part of the left hand side of (11). Since $\Delta > 0$, the equation $\partial AUC_I(l)/\partial l = 0$ if and only if $\partial\rho/\partial l = 0$; that is,

$$\frac{\partial}{\partial l}\frac{l^T \mathrm{cov}(\tilde{X}, Z)}{((l^T \Sigma l)^{1/2}\sigma_2)} = 0.$$

It implies that the optimal linear combination coefficient

$$l_{opt} = \Sigma^{-1}\mathrm{cov}(\tilde{X}, Z). \tag{12}$$

Note that, as in Su and Liu (1993), this optimal linear combination coefficient l_{opt} is independent of c, and depends only on the covariance matrix of variables and the covariance between of variables and the gold standard.

3.2. Estimation of the Optimal Linear Combination

Assume that $\{(\tilde{x}_i, z_i), i = 1, \cdots, n\}$ is a set of n independent and identically distributed random samples, where z_i denotes the observed gold standard measures as before, and $\tilde{x}_i$ is its corresponding p-dimensional vector of observed variable values of subject i. Without loss of generality, we assume that all the components of $\tilde{x}$ and z are centralized, since we can always centralize the data by subtracting their sample means, and define $H = (\tilde{x}_1 - \bar{x}, \cdots, \tilde{x}_n - \bar{x})^T$ as an $n \times p$ matrix, and $\tilde{z} = (z_1 - \bar{z}, \cdots, z_n - \bar{z})^T$ as a vector of length p, where $\bar{x} = \sum_{i=1}^n \tilde{x}_i/n$ and $\bar{z} = \sum_{i=1}^n z_i/n$. Hence, the estimate of l_{opt} based on a sample of size n following from (12) is defined as

$$\hat{l} = (H^T H)^{-1} H^T \tilde{z}. \tag{13}$$

Similarly to the linear regression problem, it is clear that $\hat{l}$ is a strongly consistent estimate of l_{opt} under some regularity conditions on $\tilde{X}$ and Z. Define

$$\hat{A}(c, l) = \frac{1}{n_1 n_0} \sum_{i \in S_1(c); j \in S_0(c)} \psi(l^T \tilde{x}_i - l^T \tilde{x}_j). \tag{14}$$

Then

$$\hat{A}_I(l) = \int \hat{A}(t, l) f_c(t) dt \tag{15}$$

is an estimate of $AUC_I(l)$. It is easy to see that for given t, $\hat{A}(t, l)$ convergenes to $\hat{A}(t, l)$ uniformly with respect to l. Hence, using the dominated convergence theorem, it is shown that $\hat{A}_I(\hat{l})$ is a strongly consistent estimate of $AUC_I(l_{opt})$ and the details are omitted here. This result is stated as a theorem below:

Theorem 3.1. *Suppose that the joint distribution of* $\tilde{X} \in R^p, Z \in R^1$ *follows a multivariate normal distribution, where* Z *is the continuous gold standard, and* $\tilde{X}$ *denotes the* p*-dimensional vector of variables. Let* $\{(\tilde{X}_1,\ Z_1), \cdots, (\tilde{X}_n,\ Z_n)\}$ *be independent and identically distributed samples of size* n*. Then for a given density* $f_c(t)$*, with probability one,*

$$\hat{A}_I(\hat{l}) - AUC_I(l_{opt}) \longrightarrow 0,\ \ as\ n \to \infty,$$

where $AUC_I(l_{opt})$ *and* $\hat{A}_I(\hat{l})$ *are defined as in (9) and (15) with* $l = l_{opt}$ *and* $\hat{l}$*, respectively.*

Equation (13) provides a neat solution for the best linear combination of variables under a joint multivariate normality assumption. However, it can be seen from (13) that the calculation of $\hat{l}$ relies on the computation of an inverse matrix. Thus, when the number of variables is large, the direct calculation of $\hat{l}$ using (13) becomes numerically unstable. The situation is worse, when the sample size is relatively small compared to the number of variables. So, we need an alternative numerical approach that can handle problems with large p to overcome this obstacle.

Again, from (13), we find that the estimate $\hat{l}$ can be viewed as a least square estimate of l in the linear regression model below:

$$\tilde{z} = Hl + e, \tag{16}$$

where e is an n-dimensional vector of random error. When p is small, then the solution can be obtained easily as in regression problems. When p is large, then we can apply the least angle regression shrinkage (LARS) method (Efron et al., 2004) to (16) to obtain an estimate of l. Since this is the same as applying LARS in a regression setup, the properties of LARS are therefore inherited. With the assistance of LARS, the proposed measure can be applied to evaluate linear combinations of lengthy variables. The variable selection scheme will follow from LARS as it is used in regression models. However, when the normality assumption is violated or the normal approximation to the joint distribution is not adequate, the empirical results show that the l_{opt} defined in (12) is not a good solution. Thus, an alternative algorithm, which does not rely on the normality assumption, is required and developed below.

Remark 3.2. Since the properties of applying LARS to find the linear combination of variables are the same as those in linear regression. We omit the details of applying LARS under the normality assumption. Instead, we focus on the case without a normality assumption.

3.3. When the Joint Distribution is Unknown

As before, let's start with a one-dimensional case, and the case with a linear combination of variables will follow easily as an extension.

Similarly to the methods used in Ma and Huang (2005), and Wang et al. (2007a), we first use a sigmoid function $S(t) = 1/(1 + \exp(-t))$ to approximate $\psi(\cdot)$ in equation (21). Thus, a smooth estimate of AUC_I is defined as

$$\hat{A}_{Is} = \int \frac{1}{n_1 n_0} \sum_{i \in S_1(t); j \in S_0(t)} S\left(\frac{y_i - y_j}{h}\right) f_c(t) dt. \tag{17}$$

It follows from the results in density estimation literature that for a sufficiently small window width h, $S((y-x)/h) \approx \psi((y-x)$, which implies the following asymptotic properties of $\hat{A}_{Is}$:

Theorem 3.3. *Assume that* $\{(y_1,\ z_1), \cdots, (y_n,\ z_n)\}$ *are* n *independent and identically distributed samples of* $(Y \in R^1, Z \in R^1)$*, where* Z *denotes a continuous gold standard. Denote the marginal densities of* Y *and* Z *by* f_Y *and* f_Z*, respectively. Let* $F(z|y)$ *be conditional cumulative function of* Z *given* $Y = y$*. Suppose that* f_Y *and* f_Z *are larger than 0 and bounded. Assume both* $f_Y(\cdot)$ *and* $F(z|\cdot)$ *are uniformly continuous. Then for a given probability density* $f_c(t)$ *with* $h = O(n^{-\alpha})$*,* $1/5 < \alpha < 1/2$*,*

$$\hat{A}_{Is} - AUC_I \to 0 \quad \textit{almost surely as } n \to \infty,$$

where AUC_I *and* $\hat{A}_{Is}$ *are defined in (2) and (17), respectively.*

(The proof of Theorem 3.3 relies on some classical results of density approximation theory. The details are given in Appendix A.)

As before, we replace y in (17) with $l^T \tilde{x}$, then we have the smooth estimate of $AUC_I(l)$ below:

$$\hat{A}_{Is}(l) = \int \frac{1}{n_1 n_0} \sum_{i \in S_1(t); j \in S_0(t)} S\left(\frac{l^T \tilde{x}_i - l^T \tilde{x}_j}{h}\right) f_c(t) dt. \tag{18}$$

The asymptotic property of $\hat{A}_{Is}(l)$ follows easily from Theorem 3.3, and is summarized as the following theorem without proof.

Theorem 3.4. *Suppose that* $\{(\tilde{x}_1,\ z_1), \cdots, (\tilde{x}_n,\ z_n)\}$ *are* n *independent and identically distributed samples of* $(\tilde{X} \in R^p, Z \in R^1)$*, where* Z *denotes the continuous gold standard, and* $\tilde{X}$ *is a vector of corresponding variables. Let* $f_c(t)$ *be a probability density. Assume that for a given constant vector* $l \in R^p$*, the conditions of Theorem (3.3) holds for* $Y = l^T\tilde{X}$ *and* Z*. Then for* $h = O(n^{-\alpha})$ *with* $1/5 < \alpha < 1/2$*,*

$$\hat{A}_{Is}(l) - AUC_I(l) \to 0 \text{ almost surely as } n \to \infty,$$

where $AUC_I(l)$ *and* $\hat{A}_{Is}(l)$ *are defined in (9) and (18), respectively.*

Remark 3.5. We only need to estimate the density function of the linear combination $l^T\tilde{X} \in R^1$, hence the choice of h does not depend on the length of total variables p. Thus, the density estimation part of the proposed algorithm will not suffer from the curse of dimensionality.

Following Theorem 3.3, we apply the threshold gradient descent method (TGDM) of Friedman and Popescu (2004) to find the best linear combination, $\hat{l}$ which maximizes $\hat{A}_{Is}(l)$. That is, to find a solution

$$\hat{l} = \text{argmax}_l \hat{A}_{Is}(l). \tag{19}$$

From equation (18), we know that AUC_{Is} is also scale invariant as is AUC. That is, $\hat{A}_{Is}(l)$ with window width h will equals to $\hat{A}_{Is}(kl)$ with $h = kh$ for a positive constant k. Hence, an anchor variable is needed such that the solution of (19) is unique.

TGDM Based Algorithm Let $\{(\tilde{x}_1, z_1), \cdots, (\tilde{x}_n, z_n)\}$ be a set of random samples of size n, which satisfies the assumption of Theorem 3.4. Define $s = (s_1, \cdots, s_p)^T$ as a p-dimensional vector with $s_i = 1$, if the corresponding empirical AUC_I of the ith variable is greater than 0.5; otherwise set $s_i = -1$. Let β_i be a p-dimensional vector where only the ith component equals s_i and 0 otherwise. Define $R_i = \hat{A}_{Is}(\beta_i)$, then choose the variable with the maximum R_i value as the anchor variable. In the following algorithm, we assume that $R_1 > R_i$, for $i = 2, \ldots, p$ without loss of generality. Let notation $\hat{l}_i$ denote the ith component of $\hat{l}$, then $\hat{l}_1$ is the coefficient of the anchor variable. In order to make the coefficients identifiable, we set $\|\hat{l}_1\| = 1$. Following the notations defined above, a TGDM-based algorithm for finding the best linear combination of variables that maximizes AUC_{Is} is stated below:

Algorithm:

(0) Initial stage: Let $r = 0$ and choose a threshold parameter τ. Set $l^{(0)} = (s_1, 0, \cdots, 0)^T$.

(1) Given $l = l^{(r)}$, calculate the derivative of the smoothed estimate $\hat{A}_{Is}(l)$ with respect to linear coefficient l, $d(l^{(r)}) = (d_1(l^{(r)}), \cdots, d_p(l^{(r)}))^T = \partial \hat{A}_{Is}(l)/\partial l|_{l=l^{(r)}}$.

(2) Use the threshold gradient descent method to calculate $l = l_0^{(r+1)}$; that is, $l_0^{(r+1)} = l^{(r)} + \delta\ t(\tau, l^{(r)})\ d(l^{(r)})$ for some $\delta > 0$, where $t(\tau, l^{(r)})$ is an indicator vector

$$I\left(d(l^{(r)}) > \tau\ \max\{d_1(l^{(r)}), \cdots, d_p(l^{(r)})\}\right).$$

(3) Find the optimal $\delta^* = \text{argmax}_{\delta>0} \hat{A}_{Is}(l_0^{(r+1)})$ with $l_0^{(r+1)} = l^{(r)} + \delta t(\tau, l^{(r)})\ d(l^{(r)})$, and update $l^{(r+1)} = l^{(r)} + \delta^* t(\tau, l^{(r)})\ d(l^{(r)})$.

(4) Repeat steps (1)-(4) until $\hat{A}_{Is}(l^{(r+1)})$ converges.

Remark 3.6. The initial value of l is chosen as $(s_1, 0, \cdots, 0)^T$, since the first component of l corresponds to the selected anchor variable. In Step (2), we update $l^{(r)}$ along the direction $t(\tau, l^{(r)})\ d(l^{(r)})$, where the number of nonzero components is decided by the threshold parameter τ, and by the definition of $t(\tau, l^{(r)})$, the locations of nonzero components of $t(\tau, l^{(r)})$ are determined by the elements of gradient $d(l^{(r)})$. Step (3) is to find a suitable step size δ^* along the direction of Step (2), then update the linear coefficients of variables. The criterion of convergence of Step (4) has to be predetermined.
(The software used in this paper (GoldAUC) is available at
http://idv.sinica.edu.tw/ycchang/software.html).

4. Numerical studies

In numerical studies, we calculate the proposed measures $\hat{A}_{Ij}$, $j = 1, 2, 3$, corresponding to 3 different $f_c(t)$ as defined before. Since the correlation coefficient is a basic statistic to measure the association between two continuous variables, we therefore include it in our experimental studies. We also compare the performances of our methods with that of Obuchowski's (2006) method (page 485, Equation (9)) described below:

$$\hat{\theta} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi'(y_i, z_i, y_j, z_j), \tag{20}$$

where $i \neq j$,

$$\begin{aligned}\psi'(y_i, z_i, y_j, z_j) &= 1 \quad \text{if } \mathrm{y_i} > \mathrm{y_j} \text{ and } \mathrm{z_i} > \mathrm{z_j}, \text{ or } \mathrm{y_i} < \mathrm{y_j} \text{ and } \mathrm{z_i} < \mathrm{z_j};\\ &= 0.5 \text{ if } \mathrm{y_i} = \mathrm{y_j} \text{ or } \mathrm{z_i} = \mathrm{z_j};\\ &= 0 \quad \text{otherwise.}\end{aligned}$$

The sample sizes used in our numerical studies are $n = 50$ and 100. The window width for the kernel estimate in $\hat{A}_{I3}$ is equal to $n^{1/5}$. The bootstrap sample size for estimating the variance of each case is 200, and there are 100 replicates for each simulation setup. For the first experimental study, the data are generated from bivariate normal distributions with means $\mu_1 = \mu_2 = 1.0$, standard deviations (σ_1, σ_2) equal to $(1.0, 1.0)$, $(1.0, 2.0)$, $(2.0, 1.0)$ and $(2.0, 2.0)$, and correlation coefficients equal to $\rho = 0.0$, 0.25, 0.5, 0.75, and 1.0. Let $\hat{\mu}$ and $\hat{\sigma}^2$ denote the sample mean and variance of z. As in the classical ROC curve analysis, when a variable with no diagnostic power, then its corresponding ROC curve will be the 45 degree diagonal line of the unit square. If this case holds for all possible cutting points, then it implies that $AUC_I = 0.5$. So, we use 0.5 as the value of the null hypothesis in our numerical study. Table 1 shows five statistics for different simulation setups: correlation coefficient of two variables $\hat{\rho}$, $\hat{A}_{Ij}$, $j = 1, 2, 3$ with corresponding $f_c(t)$'s, and $\hat{\theta}$ from Obuchowski (2006). Figure 1 is a plot of statistics $\hat{\rho}^2/V(\hat{\rho})$, $(\hat{A}_{Ij} - 0.5)^2/V(\hat{A}_{Ij})$ for all j's, and $(\hat{\theta} - 0.5)^2/V(\hat{\theta})$ versus ρ, where $V(\hat{\rho})$ and $V(\hat{\theta})$ are the bootstrap estimates of variances of $\hat{\rho}$ and $\hat{\theta}$, respectively.

When the joint distribution of two variables follows a bivariate normal distribution, the correlation coefficient is a natural statistic to describe the association between the two variables. In our study, all five measures increase as the true correlation coefficient ρ increases, which suggests that all measures catch the linear association between variable Y and the gold standard Z as expected. In fact, $\hat{A}_{Ij}$ and $\hat{\theta}$ are very close to their true values 0.5 and 1.0, when ρ are equal to 0.0 and 1.0, respectively. In addition, Figure 1 shows that the values of $\hat{\rho}^2/V(\hat{\rho})$ and $(\hat{A}_{Ij} - 0.5)^2/V(\hat{A}_{Ij})$, $j = 1, 2, 3$, are larger than those of $(\hat{\theta} - 0.5)^2/V(\hat{\theta})$ under current simulation set up.

Table 2 shows the results of five measures when there is no association between variable Y and the gold standard Z. That is, the data set used in this table are generated from the model $y = z^2 + \epsilon$ with standard normal error ϵ, where the gold standard z is generated from three different distributions: (1) normal distribution, (2) t_2 distribution with free degree 2, and (3) a Cauchy

Table 1: Comparison of five measure indexes: $\hat{\rho}$, $\hat{A}_{Ij},\ j = 1,\ 2,\ 3$, and $\hat{\theta}$, where the marker and gold standard, $(y,\ z)$, follow multi-variate normal distribution with means $\mu_1 = \mu_2 = 1.0$, with different standard deviations $\sigma_1,\ \sigma_2$ and distinct correlation coefficients ρ.

n	$(\sigma_1,\ \sigma_2)$	Method	0.0	0.25	0.5	0.75	1.0
50	(1.0, 1.0)	$\hat{\rho}$	0.105(0.076, 0.140)*	0.252(0.118, 0.130)	0.511(0.088, 0.103)	0.747(0.067, 0.064)	1.000(0.000, 0.000)
		$\hat{A}_{I1}$	0.505(0.064, 0.073)	0.621(0.063, 0.069)	0.746(0.053, 0.058)	0.866(0.040, 0.038)	1.000(0.000, 0.000)
		$\hat{A}_{I2}$	0.501(0.065, 0.067)	0.616(0.062, 0.063)	0.743(0.046, 0.052)	0.856(0.035, 0.033)	0.979(0.010, 0.013)
		$\hat{A}_{I3}$	0.498(0.065, 0.066)	0.611(0.062, 0.062)	0.737(0.045, 0.051)	0.846(0.037, 0.034)	0.968(0.011, 0.015)
		$\hat{\theta}$	0.504(0.044, 0.049)	0.583(0.044, 0.048)	0.673(0.038, 0.044)	0.771(0.036, 0.035)	1.000(0.000, 0.004)
	(1.0, 2.0)	$\hat{\rho}$	0.106(0.073, 0.136)	0.263(0.118, 0.131)	0.477(0.099, 0.109)	0.750(0.058, 0.065)	1.000(0.000, 0.000)
		$\hat{A}_{I1}$	0.497(0.067, 0.073)	0.621(0.061, 0.070)	0.730(0.053, 0.061)	0.862(0.034, 0.040)	1.000(0.000, 0.000)
		$\hat{A}_{I2}$	0.495(0.065, 0.066)	0.622(0.061, 0.064)	0.729(0.053, 0.054)	0.859(0.029, 0.033)	0.980(0.008, 0.010)
		$\hat{A}_{I3}$	0.496(0.065, 0.066)	0.622(0.062, 0.064)	0.729(0.051, 0.054)	0.858(0.030, 0.034)	0.983(0.004, 0.009)
		$\hat{\theta}$	0.498(0.044, 0.049)	0.583(0.043, 0.049)	0.660(0.038, 0.044)	0.769(0.032, 0.036)	1.000(0.000, 0.004)
100	(1.0, 1.0)	$\hat{\rho}$	0.085(0.056, 0.098)	0.253(0.083, 0.092)	0.497(0.082, 0.075)	0.747(0.046, 0.044)	1.000(0.000, 0.000)
		$\hat{A}_{I1}$	0.490(0.050, 0.051)	0.620(0.043, 0.048)	0.739(0.046, 0.041)	0.865(0.024, 0.027)	1.000(0.000, 0.000)
		$\hat{A}_{I2}$	0.485(0.053, 0.049)	0.622(0.041, 0.046)	0.741(0.042, 0.038)	0.864(0.023, 0.023)	0.987(0.007, 0.008)
		$\hat{A}_{I3}$	0.483(0.054, 0.049)	0.620(0.042, 0.045)	0.739(0.042, 0.037)	0.861(0.024, 0.023)	0.982(0.006, 0.009)
		$\hat{\theta}$	0.493(0.033, 0.034)	0.581(0.029, 0.033)	0.668(0.033, 0.030)	0.771(0.023, 0.024)	1.000(0.000, 0.001)
	(1.0, 2.0)	$\hat{\rho}$	0.075(0.057, 0.097)	0.266(0.100, 0.091)	0.499(0.081, 0.074)	0.739(0.045, 0.046)	1.000(0.000, 0.000)
		$\hat{A}_{I1}$	0.496(0.049, 0.051)	0.625(0.053, 0.048)	0.739(0.042, 0.041)	0.859(0.025, 0.027)	1.000(0.000, 0.000)
		$\hat{A}_{I2}$	0.493(0.050, 0.049)	0.629(0.051, 0.045)	0.744(0.041, 0.037)	0.862(0.024, 0.023)	0.987(0.006, 0.006)
		$\hat{A}_{I3}$	0.494(0.049, 0.049)	0.630(0.052, 0.045)	0.745(0.041, 0.037)	0.862(0.024, 0.024)	0.99(0.003, 0.005)
		$\hat{\theta}$	0.498(0.032, 0.034)	0.586(0.036, 0.033)	0.667(0.031, 0.030)	0.765(0.022, 0.024)	1.000(0.000, 0.001)

*Empirical standard deviations and mean values of bootstrap standard deviations are in parentheses.

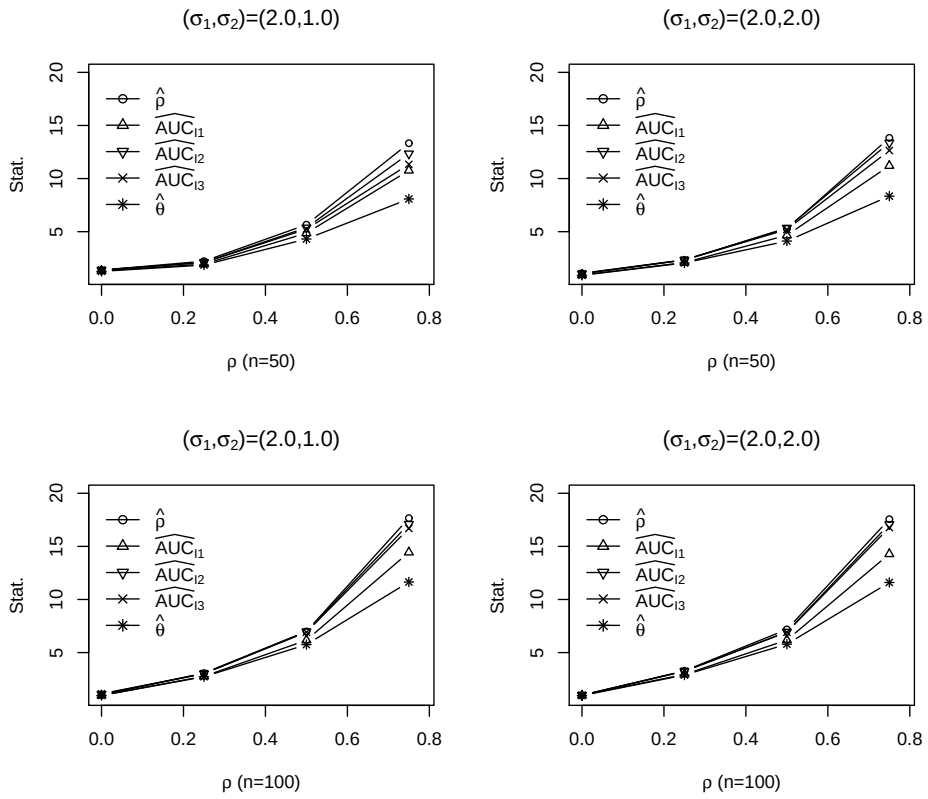

Figure 1: Comparison of five measures: $\hat{\rho}^2/V(\hat{\rho})$, $(\hat{A}_{Ij}-0.5)^2/V(\hat{A}_{Ij})$, $j=1,\ 2,\ 3$, and $(\hat{\theta}-0.5)^2/V(\hat{\theta})$, where (Y,Z) follow bivariate normal distributions with means $\mu_1=\mu_2=1.0$, with different standard deviations $\sigma_1,\ \sigma_2$ and correlation coefficients ρ.

distribution. Since z has symmetrical density functions for all three cases, it is clear that there is no association between Y and Z. That is, the ideal values of the correlation coefficient estimate $|\hat{\rho}|$, ROC-type indexes estimates $|\hat{A}_{Ij} - 0.5|$,$j = 1, 2, 3$ and $|\hat{\theta} - 0.5|$ should be close to 0. We calculate the $25\%, 50\%$ and 75% empirical quantiles based on 100 simulations. The p-values, with a nominal significance level equal to 0.05, for statistics $\hat{\rho}^2/V(\hat{\rho})$, $(\hat{A}_{Ij} - 0.5)^2/V(\hat{A}_{Ij})$, $j = 1, 2, 3$ and $(\hat{\theta} - 0.5)^2/V(\hat{\theta})$ are also reported. It is seen from Table 2 that all three quantiles of $\hat{A}_{I3}$ and $\hat{\theta}$ are very close to 0, while the correlation coefficient seems to over-estimate the association of Y and Z in this experiment. When the tail of the distribution of Z becomes heavier, the quantiles and p-values of $\hat{\rho}$ become further from 0.0 and nominal 0.05, respectively. Especially, when Z is from a Cauchy distribution, the 25% quantiles are larger than 0.5 and the corresponding p-values are greater than 0.3.

The performances of $\hat{A}_{I3}$ and $\hat{\theta}$ are better than those of $\hat{A}_{I1}$ and $\hat{A}_{I2}$ when Z is not from a normal distribution. This is because $\hat{A}_{I3}$ is based on a kernel estimate of $f_c(t)$ and $\hat{\theta}$ is founded on a nonparametric method, they are not affected by the distribution of Z, and therefore very stable even when Z is not normally distributed.

As a summarization and conclusion to the results of Figure 1, and Tables 1 and 2, both $\hat{A}_{I3}$ and $\hat{\theta}$ are recommended for detecting the association between variables and the continuous gold standard. Although $\hat{\theta}$ is considered as a natural extension of the ordinary AUC index, it is worth noting that the performance of $\hat{A}_{Ij}$ (especially $\hat{A}_{I3}$), in these cases, are are very competitive.

4.1. Combination of Variables

Both correlation coefficient (CC) and the TGDM algorithm are used to obtain the optimal linear combinations of variables. We then calculate $\hat{A}_{I3}$ and $\hat{\theta}$ of the corresponding combination of variables based on the coefficient vectors obtained from these two methods. The threshold parameter τ in the TGDM algorithm is equal to 1.0 in our studies. The data set are generated from $Z = l^T\tilde{X} + \epsilon$, where $\tilde{X}$ follows a p dimensional multivariate normal distribution with mean vector $(0, \ldots, 0)^T$ and an identity covariance matrix, and the true $l = (1.0, 1.0, 0.0, \cdots, 0.0)^T$. Error term ϵ is generated from either the standard normal distribution or a Cauchy distribution. In this experimental study, we have tried three different dimensions of X ($p = 4, 10, 20$) for all cases, and only variables x_1 and x_2 have non-zero coefficients. That is, only these two variables are associated with the gold standard. Moreover,

Table 2: Comparison of different methods when there is no association between variable Y and the gold standard Z. The data set (y, z) is generated from model $y = z^2 + \epsilon$ with standard normal error ϵ. Three different distributions of z are used, which are a normal distribution, a t_2 distribution with free degree 2 and a Cauchy distribution.

		Normal		t_2		Cauchy	
n	Model	(25%, 50%, 75%)	p-value*	(25%, 50%, 75%)	p-value	(25%, 50%, 75%)	p-value
50	$\hat{\rho}$	(0.086, 0.175, 0.287)	0.13	(0.277, 0.544, 0.802)	0.30	(0.515, 0.834, 0.948)	0.33
	$\hat{A}_{I1}$	(0.031, 0.060, 0.110)	0.09	(0.048, 0.079, 0.125)	0.12	(0.044, 0.074, 0.125)	0.13
	$\hat{A}_{I2}$	(0.025, 0.054, 0.090)	0.06	(0.041, 0.076, 0.120)	0.13	(0.040, 0.078, 0.150)	0.15
	$\hat{A}_{I3}$	(0.022, 0.050, 0.087)	0.06	(0.036, 0.060, 0.090)	0.09	(0.028, 0.055, 0.088)	0.09
	$\hat{\theta}$	(0.021, 0.043, 0.081)	0.08	(0.034, 0.060, 0.087)	0.09	(0.025, 0.049, 0.090)	0.08
100	$\hat{\rho}$	(0.055, 0.114, 0.202)	0.07	(0.305, 0.527, 0.728)	0.27	(0.603, 0.825, 0.931)	0.37
	$\hat{A}_{I1}$	(0.022, 0.044, 0.072)	0.07	(0.016, 0.035, 0.072)	0.06	(0.029, 0.052, 0.098)	0.15
	$\hat{A}_{I2}$	(0.020, 0.033, 0.059)	0.06	(0.014, 0.033, 0.064)	0.05	(0.026, 0.048, 0.096)	0.14
	$\hat{A}_{I3}$	(0.017, 0.034, 0.060)	0.06	(0.009, 0.032, 0.056)	0.04	(0.018, 0.041, 0.07)	0.08
	$\hat{\theta}$	(0.015, 0.032, 0.049)	0.07	(0.012, 0.031, 0.054)	0.04	(0.014, 0.036, 0.065)	0.06

*Nominal significance level is 0.05.

a software based on the TGDM algorithm to calculate the optimal linear combination of variables is available as an R package. It is also worth noting that there is no algorithm or discussion in Obuchowski (2005) about finding the linear combination of variables based on $\hat{\theta}$.

Table 3 lists the values of $\hat{A}_{I3}$ and $\hat{\theta}$ for individual variables, x_1 and x_2, and the linear combinations based on the CC and TGDM methods. From this table, we find that $\hat{A}_{I3}$ and $\hat{\theta}$ for linear combinations of variables are always larger than for individual variables, which confirms that linear combinations of variables can improve on the the diagnostic power of individual variables. When ϵ follows the standard normal distribution, $\hat{A}_{I3}$ and $\hat{\theta}$ for linear combinations based on both TGDM and CC are very close. However, when ϵ is a Cauchy distribution, the TGDM method has larger $\hat{A}_{I3}$ and $\hat{\theta}$ than combinations based on CC. This is because the CC method relies on the normality assumption, while TGDM does not. In addition, from Table 3, we can see that $\hat{A}_{I3}$ is larger than $\hat{\theta}$. In most of the cases, the standard deviations of TGDM are smaller than those of $\hat{\theta}$, which suggests that the linear combinations based on TGDM have greater diagnostic power, although the difference may not be statistically significant in our simulation.

Table 3: Results of linear combination using correlation coefficient (CC) and TGDM method.

				$Nonzero\ coef.^{+}$			
Distribution	p^{**}	n	Method	x_1	x_2	CC	TGDM
Normal	4	50	$\hat{A}_{I3}$	0.773(0.054)*	0.786(0.052)	0.900(0.024)	0.900(0.028)
			$\hat{\theta}$	0.694(0.043)	0.702(0.042)	0.815(0.028)	0.815(0.031)
		100	$\hat{A}_{I3}$	0.782(0.033)	0.785(0.035)	0.904(0.018)	0.906(0.018)
			$\hat{\theta}$	0.693(0.027)	0.696(0.030)	0.807(0.021)	0.809(0.021)
	10	50	$\hat{A}_{I3}$	0.785(0.048)	0.773(0.046)	0.909(0.021)	0.900(0.031)
			$\hat{\theta}$	0.703(0.037)	0.692(0.040)	0.824(0.027)	0.815(0.033)
		100	$\hat{A}_{I3}$	0.791(0.036)	0.789(0.032)	0.913(0.015)	0.913(0.016)
			$\hat{\theta}$	0.699(0.030)	0.700(0.025)	0.818(0.019)	0.817(0.020)
	20	50	$\hat{A}_{I3}$	0.767(0.051)	0.779(0.053)	0.928(0.018)	0.897(0.034)
			$\hat{\theta}$	0.689(0.042)	0.698(0.042)	0.852(0.026)	0.813(0.039)
		100	$\hat{A}_{I3}$	0.782(0.033)	0.783(0.032)	0.922(0.015)	0.915(0.016)
			$\hat{\theta}$	0.693(0.028)	0.696(0.025)	0.828(0.019)	0.820(0.019)
Cauchy	4	50	$\hat{A}_{I3}$	0.659(0.067)	0.640(0.068)	0.669(0.107)	0.735(0.073)
			$\hat{\theta}$	0.629(0.046)	0.614(0.046)	0.619(0.088)	0.685(0.059)
		100	$\hat{A}_{I3}$	0.660(0.056)	0.657(0.047)	0.659(0.094)	0.724(0.077)
			$\hat{\theta}$	0.629(0.036)	0.625(0.032)	0.615(0.078)	0.679(0.063)
	10	50	$\hat{A}_{I3}$	0.648(0.064)	0.645(0.072)	0.690(0.099)	0.750(0.067)
			$\hat{\theta}$	0.620(0.045)	0.618(0.048)	0.628(0.079)	0.689(0.056)
		100	$\hat{A}_{I3}$	0.648(0.083)	0.638(0.082)	0.664(0.104)	0.733(0.101)
			$\hat{\theta}$	0.625(0.033)	0.618(0.035)	0.614(0.063)	0.683(0.061)
	20	50	$\hat{A}_{I3}$	0.647(0.093)	0.657(0.096)	0.740(0.123)	0.789(0.096)
			$\hat{\theta}$	0.623(0.044)	0.628(0.046)	0.665(0.083)	0.719(0.052)
		100	$\hat{A}_{I3}$	0.634(0.123)	0.638(0.120)	0.649(0.142)	0.739(0.147)
			$\hat{\theta}$	0.624(0.032)	0.627(0.029)	0.604(0.068)	0.689(0.069)

$^{+}Nonzero\ coef.$ represents variables with non-zero coefficients in true model.

*Empirical standard deviations are in parentheses.

**p denotes number of total variables in true model and the number of non-zero variables is $p_1 = 2$.

4.2. Real examples

We apply the proposed measures to three real data sets: tumor, prostate and diabetes data sets, which are used in Obuchowski (2005), Stamey et al. (1989) and Willems et al. (1997), respectively. In the tumor data set, there are 74 patients and only two surgery variables: the computed tomography (CT) and a fictitious test (Fi). The continuous gold standard of this data set is the size of the renal tumor mass. The prostate data has 97 patients with prostate specific antigen as its gold standard together with 6 continuous variables, which are cancer volume, prostate weight, age (Age), benign prostatic hyperplasia amount, capsular penetration, and percentage Gleason scores 4 or 5 (Pgg45). Except variables Age and Pgg45, the others are recoded in log-scale and denoted by Lcavol, Lweight, Lbph, Lcp and Lpsa, accordingly. The original diabetes data consists of 403 subjects, but we follows Willems et al. (1997) to delete 22 subjects with missing variables. Of the remaining 381 subjects from this data set used in our numerical study, 222 are females and 159 are males. The following 8 continuous variables are used in this data set: total cholesterol (Chol), stabilized glucose (Stab.glu), high density lipoprotein (Hdl), cholesterol/HDL ratio (Ratio), age (Age), body mass index (BMI) and waist/hip ratio (WHR). The gold standard for this data set is glycosylated hemoglobin (Glyhb), which is commonly used as a measure of the progress of diabetes. In addition to analyzing the entire diabetes data set, we also investigate female and male subgroups, separately.

We normalize the data before applying the proposed measures to each data set to avoid scale variations. Table 4 presnets $\hat{A}_{I3}$ and $\hat{\theta}$ for individual variables with p-value less than 10^{-7}. From Table 4, we find that $\hat{A}_{I3}$ selects more variables than $\hat{\theta}$ for some cases. Note that $\hat{A}_{I3}$ are much larger than $\hat{\theta}$ with competitive standard deviations in these cases.

Table 5 lists the linear coefficients obtained using the TGDM and CC methods, and their corresponding $\hat{A}_{I3}$ and $\hat{\theta}$ values for all data sets, including the male and female subgroups of the diabetes data set. In the tumor data set, Fi has a larger $\hat{A}_{I3}$ value than CT; that is, Fi has a greater association with the size of the renal tumor mass for tumor data. In the prostate data set, Lcavol has the largest $\hat{A}_{I3}$ value; that is, Lcavol is most highly associated with prostate specific antigen among all variables considered in the prostate data set. For the diabetes data set and its male and female subgroups, the largest $\hat{A}_{I3}$ and the variable with the largest coefficient value is Stab.glu; that is, Stab.glu has the highest potential to diagnose diabetes in terms of glycosylated hemoglobin index. As expected, from Tables 4 and 5, the linear

Table 4: Results of ROC measure indexes: $\hat{A}_{I3}$ and $\hat{\theta}$, of single markers for tumor, prostate, diabetes, diabetes-female and diabetes-male data sets.

Tumor

Data	Method	CT	Fi
Tumor	$\hat{A}_{I3}$	0.943(0.014)*	0.982(0.011)
	$\hat{\theta}$	0.871(0.020)	0.956(0.008)

Prostate

Data	Method	Lcavol	Lweight	Lcp	Pgg45
Prostate	$\hat{A}_{I3}$	0.865(0.022)	0.722(0.034)	0.759(0.035)	0.744(0.035)
	$\hat{\theta}$	0.758(0.027)	0.647(0.027)	0.675(0.031)	0.676(0.028)

Diabetes

Data	Method	Chol	Stab.glu	Ratio	Age
Diabetes	$\hat{A}_{I3}$	-	0.779(0.021)	0.662(0.022)	0.711(0.019)
	$\hat{\theta}$	-	0.687(0.017)	0.600(0.015)	0.644(0.014)
Diabetes-female	$\hat{A}_{I3}$	0.667(0.029)	0.786(0.022)	-	0.732(0.025)
	$\hat{\theta}$	-	0.691(0.021)	-	0.665(0.019)
Diabetes-male	$\hat{A}_{I3}$	-	0.769(0.039)	0.689(0.034)	0.681(0.030)
	$\hat{\theta}$	-	0.682(0.030)	-	-

*Bootstrap standard deviation is in parentheses.

combinations based on TGDM and CC usually have larger $\hat{A}_{I3}$ and $\hat{\theta}$ values than individual variables do, and similarly, $\hat{A}_{I3}$ and $\hat{\theta}$ values for combinations from TGDM are a little bit larger than those obtained using the CC method. In real data sets the relation is seldom linear, which is the reason why the combinations obtained using TGDM perform better than others.

5. Conclusion and Discussion

In this paper, we first propose a new measure for evaluating the potential diagnostic power of individual variables, when there is only a continuous

Table 5: Results of optimal linear coefficients and corresponding ROC measure indexes: $\hat{A}_{I3}$ and $\hat{\theta}$, for tumor, prostate, diabetes, diabetes-female and diabetes-male data sets.

Tumor

Data	Method	Coef.		ROC-type indexes	
		CT	Fi	$\hat{A}_{I3}$	$\hat{\theta}$
Tumor	CC	-0.118	1.076	0.981(0.011)	0.950(0.009)
	TGDM	0.044	1.000	0.983(0.011)	0.957(0.008)

Prostate

Data	Method	Coef.						ROC-type indexes	
		Lcavol	Lweight	Age	Lbph	Lcp	Pgg45	$\hat{A}_{I3}$	$\hat{\theta}$
Prostate	CC	0.642	0.214	-0.118	0.099	0.017	0.147	0.892(0.018)	0.791(0.024)
	TGDM	1.000	0.264	-0.108	0.135	-0.013	0.189	0.891(0.017)	0.789(0.023)

Diabetes

Data	Method	Coef.							ROC-type indexes	
		Chol	Stab.glu	Hdl	Ratio	Age	BMI	WHR	$\hat{A}_{I3}$	θ
Diabetes	CC	0.074	0.668	0.018	0.101	0.101	0.017	0.019	0.816(0.017)	0.717(0.015)
	TGDM	0.061	1.000	-0.027	0.099	0.373	0.140	0.011	0.826(0.018)	0.723(0.016)
Diabetes-female	CC	0.109	0.659	-0.073	0.027	0.106	0.029	0.069	0.834(0.021)	0.737(0.019)
	TGDM	0.253	1.000	-0.164	-0.007	0.389	0.133	0.199	0.842(0.019)	0.741(0.018)
Diabetes-male	CC	-0.005	0.701	0.141	0.243	0.085	-0.049	-0.002	0.786(0.03)	0.691(0.025)
	TGDM	-0.016	1.000	0.009	0.179	0.367	0.100	-0.040	0.811(0.031)	0.706(0.027)

*ROC-type indexes used here are AUC_{I3} and $\hat{\theta}$.

gold standard available and no confirmative threshold for it is known. The proposed measure is an AUC-type index that shares the threshold independent property of the ROC curve and AUC, and can also be used to evaluate the performance of classifiers when the gold standard variable is essentially continuous, and the threshold is controvertible. Numerical results show that the proposed novel index is very competitive to the existence method.

In addition, we propose algorithms, based on the newly defined index, for finding the best linear combination of variables, which is useful from a practical prospect when there are multiple variables considered at a time, and how to evaluate or select a good combination of variables is an important issue. Here we also study numerical methods for finding the linear combination of variables that maximizes the proposed measure. When the normality assumption of variables is valid, the best linear combination solution can be realized as a solution to a linear system. Thus, under an assumption of normality and when the number of variable p is large, the LARS algorithm can be applied to obtain such a linear combination. This also implies that the LARS-type variable selection scheme can be conducted even when no binary-scale gold standard is available. When the joint distribution of variables is unknown, the proposed measure is then approximated using a nonparametric kernel density estimation method. In this case, we proposed a TGDM-based algorithm to calculate the best linear combination of variables. Based on numerical results, we found that our method is numerically stable with computational advantage when there are large number of variables considered and combination of variables is of interest. Moreover, our method can be easily extended to an ordinal-scale gold standard with a suitable choice of a weight function for cutting points, which will be reported elsewhere.

Appendix

Let random variables (Y_i, Z_i) denote a pair of measures from subject i, for $i \geq 1$. Suppose that $\{(y_i, z_i),\ i = 1, \ldots, n\}$ are n independent observed values of random variables (Y_i, Z_i), $i = 1, \cdots, n$. For a given cutting point c, a subject i, $i = 1, \ldots, n$, is assigned as a "case" if $z_i > c$ and otherwise labeled as a "control". That is, for a given c, we divide the observed subjects into two groups; let $S_1(c)$ and $S_0(c)$ be the case and control groups with sample sizes n_1 and n_0, respectively.

Then we propose a natural estimate of AUC index, AUC_I, with continu-

ous gold standard,

$$\hat{A}_I = \int \hat{A}(t) d\hat{F}_c(t), \tag{21}$$

where $\hat{A}(c)$ is defined as

$$\hat{A}(c) = \frac{1}{n_0\, n_1} \sum_{i \in S_1(c);\, j \in S_0(c)} \psi(y_i - y_j),$$

$\psi(u) = 1$, if $u > 0$; $= 0.5$, if $u = 0$ and $= 0$ if $u < 0$ and $\hat{F}_c(t)$ is the empirical estimate of the cumulative distribution function of Z based on $\{z_1, \ldots, z_n\}$. However, in practice, it is rare to choose cutting points at ranges near the two ends of the distribution of Z. Thus, instead of the whole range of Z, we might explicitly define a weight function $f_c(t)$ on a particular critical range.

Since the step function $\psi(\cdot)$ in (21) is not continuously differentiable, a smooth estimate of AUC_I is defined as

$$\hat{A}_{Is} = \int \frac{1}{n_1 n_0} \sum_{i \in S_1(t); j \in S_0(t)} S\left(\frac{y_i - y_j}{h}\right) f_c(t) dt, \tag{22}$$

where $S(t)$ is a sigmoid function $1/(1 + \exp(-t))$ and h is window width.

Appendix A: Proof of Strong Consistency of $\hat{A}_{Is}(l)$

The proof of the strong consistency of smoothed $AUC_I(l)$ estimator $\hat{A}_{Is}(l)$ follows from the following three lemmas.

Lemma 5.1. *Suppose that $X_1, \cdots, X_n$ is a sequence of independent and identically distributed random variables with values in R^1, and a uniformly continuous density $f(\cdot)$. Let $k(x)$ be a bounded probability density and the Dirichlet series $\sum_{n=1}^{\infty} n \exp(-\gamma \eta_n)$, $\eta_n = n\, h^2$ converges for any $\gamma > 0$. Then*

$$\int_{-\infty}^{\infty} |f_n(x) - f(x)| dx \to 0, \ \textit{almost surely as } n \to \infty,$$

where $f_n(x) = \frac{1}{n\, h} \sum_{i=1}^{n} k((x - X_i)/h)$ is a kernel density estimator of $f(x)$.

(The proof of Lemma 5.1 can be found in Nadaraya (1989), Theorem 3.1, page 55. So, it is omitted here.)

Lemma 5.2. *Suppose that $X_1, \cdots, X_n$ is a sequence of independent and identically distributed random variables with values in R^1, and a uniformly continuous density. Then with probability one, as $n \to \infty$*

$$\sup_{x \in R^1} |F_n(x) - F(x)| \to 0,$$

where $F_n(\cdot)$ and $F(\cdot)$ are the empirical distribution and distribution functions of X, respectively.

Proof of Lemma 5.2:
From Nadaraya (1989) (Equation (1.4), page 43), we have

$$\text{pr}(\sup_{x \in R^1} |F_n(x) - F(x)| > \eta\, n^{-1/2}) \leq c \exp(-2\eta^2), \tag{23}$$

which completes the proof of Lemma 5.2.

Lemma 5.3. *Assume that $\{(y_1,\ z_1), \cdots, (y_n,\ z_n)\}$ are n independent and identically distributed samples of $(Y \in R^1, Z \in R^1)$, where Z denotes a continuous gold standard. For a given c, let $\tilde{f}(y|Z > c)$ be a conditional density function of Y given $Z > c$. Suppose that conditions of Theorem 3 holds. Then $\tilde{f}(\cdot|Z > c)$ is uniformly continuous.*

Proof of Lemma 5.3:
By the Bayesian theorem, we have

$$\tilde{f}(y|Z > c) = \frac{\int_c^\infty f(y, z)dz}{\text{pr}(Z > c)}. \tag{24}$$

For any $y_i \in R^1$, $i = 1, 2$,

$$\begin{aligned} &\int_c^\infty f(y_1, z)dz - \int_c^\infty f(y_2, z)dz \\ &= \int_c^\infty [f(z|y_1)f_Y(y_1) - f(z|y_1)f_Y(y_2)]dz + \int_c^\infty [f(z|y_1)f_Y(y_2) - f(z|y_2)f_Y(y_2)]dz \\ &= [f_Y(y_1) - f_Y(y_2)][1 - F(c|y_1)] + [F(z|y_2) - F(z|y_1)]f_Y(y_2), \end{aligned} \tag{25}$$

where $f(z|y)$ is a conditional density function of Z given $Y = y$ and $f_Y(y)$ is a density function of marker Y. From the conditions of Theorem 3, we have $b \equiv \text{pr}(Z > c) > 0$, $f_Y(\cdot) < M$ and both $f_Y(\cdot)$ and $F(z|\cdot) - F(z|\cdot)$ are uniformly continuous. Hence, for any $\epsilon > 0$, there exists a $\delta > 0$, for any y_1 and y_2 satisfying $|y_1 - y_2| < \delta$, we have

$$\begin{aligned} |f_Y(y_1) - f_Y(y_2)| &< b\epsilon/2 \\ |F(z|y_2) - F(z|y_1)| &< b\epsilon/(2M). \end{aligned} \tag{26}$$

Consequently, by (24), (25) and (26) we get that for a given c,

$$\begin{aligned} & |\tilde{f}(y_1|Z>c) - \tilde{f}(y_2|Z>c)| \\ < \; & \frac{1}{b}\{|f_Y(y_1) - f_Y(y_2)|(1-F(c|y_1)) + |F(z|y_2) - F(z|y_1)|f_Y(y_2)\} \\ < \; & \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned} \tag{27}$$

It follows that $\tilde{f}(\cdot|Z>c)$ is uniformly continuous.

Proof of Theorem 3:
By the triangle inequality, we have, for fixed l,

$$\begin{aligned} \left|\hat{A}_{Is} - AUC_I\right| \; & \leq \left|\hat{A}_{Is} - \hat{A}_I\right| + \left|\hat{A}_I - AUC_I\right| \\ & = (I) + (II) \text{ (say)}. \end{aligned} \tag{28}$$

From Theorem 1, (II) converges to 0 almost surely as n goes to ∞; that is

$$\hat{A}_I - AUC_I \to 0 \quad \text{ almost surely as } n \to \infty. \tag{29}$$

From (21) and (17),

$$\begin{aligned} (I) &= \left| \int \tfrac{1}{n_1 n_0} \textstyle\sum_{i\in S_1(t); j\in S_0(t)} S\left(\tfrac{y_i - y_j}{h}\right) f_c(t)dt - \int \tfrac{1}{n_1 n_0} \textstyle\sum_{i\in S_1(t); j\in S_0(t)} \psi(y_i - y_j) f_c(t)dt \right| \\ &\leq \int \left| \tfrac{1}{n_1 n_0} \textstyle\sum_{i\in S_1(t); j\in S_0(t)} S\left(\tfrac{y_i - y_j}{h}\right) - \tfrac{1}{n_1 n_0} \textstyle\sum_{i\in S_1(t); j\in S_0(t)} \psi(y_i - y_j) \right| f_c(t)dt. \end{aligned}$$

Due to $n_1 + n_0 = n$, then at least one of $n_1 \to \infty$ and $n_0 \to \infty$ holds as n tends to ∞. Without loss of generality, assume that n_1 tends to ∞. Then

$$\begin{aligned} (I) \leq & \int \tfrac{1}{n_0} \textstyle\sum_{j\in S_0(t)} \left| \tfrac{1}{n_1} \textstyle\sum_{i\in S_1(t)} S\left(\tfrac{y_i - y_j}{h}\right) - \tilde{F}(y_j|Z>t) \right| f_c(t)dt \\ & + \int \tfrac{1}{n_0} \textstyle\sum_{j\in S_0(t)} \left| \tfrac{1}{n_1} \textstyle\sum_{i\in S_1(t)} \psi(y_i - y_j) - \tilde{F}(y_j|Z>t) \right| f_c(t)dt, \end{aligned} \tag{30}$$

where $\tilde{F}(\cdot|Z>t)$ is the conditional cumulative distribution function of Y given $\{Z>t\}$. Let $\tilde{f}(\cdot|Z>t)$ be its conditional density function. By Lemma 5.3, $\tilde{f}(\cdot|Z>t)$ is uniformly continuous.

Let $h = n^{-\alpha}$, $1/5 < \alpha < 1/2$. Set $\eta_n = nh^2 = n^{1-2\alpha}$, and the Dirichlet series $\sum_{n=1}^{\infty} n\exp(-\gamma\eta_n)$ converges for any $\gamma > 0$. Thus, the conditions of Lemma 5.1 are satisfied. Let $k(t)$ denote the derivative of $S(t)$, then $k(t)$ is

a bounded probability density. Thus, by Lemma 5.1,

$$\begin{aligned} &\sup_{y\in R^1}\left|\tfrac{1}{n_1}\textstyle\sum_{i\in S_1(t)} S\left(\tfrac{y_i-y}{h}\right)-\tilde{F}(y|Z>t)\right| \\ &=\sup_{y\in R^1}\left|\textstyle\int_{-\infty}^{y}\left(\tfrac{1}{n_1 h}\sum_{i\in S_1(t)} k\left(\tfrac{y_i-t}{h}\right)-\tilde{f}(t|Z>t)\right)dt\right| \\ &\leq \textstyle\int_{-\infty}^{\infty}\left|\left(\tfrac{1}{n_1 h}\sum_{i\in S_1(t)} k\left(\tfrac{y_i-t}{h}\right)-\tilde{f}(t|Z>t)\right)\right|dt \longrightarrow 0, \quad \text{almost surely as } n\to\infty. \end{aligned} \tag{31}$$

From Lemma 5.2, we have

$$\sup_{y\in R^1}\left|\frac{1}{n_1}\sum_{i\in S_1(t)}\psi(y_i-y)-\tilde{F}(y|Z>t)\right|\longrightarrow 0, \quad \text{almost surely as } n\to\infty. \tag{32}$$

From (30), (31) and (32), we prove that

$$\hat{A}_{Is}-\hat{A}_I\to 0, \text{ almost surely as } n\to\infty. \tag{33}$$

Put (29) and (33) together to complete the proof of Theorem 3.

Acknowledgements

This work is partially supported via NSC97-2118-M-001-004-MY2 funded by the National Science Council, Taipei, Taiwan, ROC.

References

Choi, Y., Johnson, W., Collins, M., Gardner, I. (2006). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *Journal of Agricuture, Biological and Enviromental Statistics* **11**, 210 – 229.

Efron, B., Johnstone, I., Hastie, T., Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–499.

Friedman, J. H., Popescu, B. E. (2004). Gradient directed regularization for linear regression and classification. Tech. rep., Department of Statistics, Stanford University.

Henkelman, R., Kay, I., Bronskill, M. (1990). Receiver operating characteristic analysis without truth. *Medical Decision Making* **10**.

Krzanowski, W., Hand, D. (2009). *ROC curves for Continuous Data*. CRC Press, London.

Ma, S., Huang, J. (2005). Regularized roc method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21**, 4356–4362.

Waikar, S., Betensky, R., , Bonventre, J. (2009). Creatinine as the gold standard for kidney injury biomarker studies? *Nephrol Dial Transplant* **24**, 3263–3265.

Nadaraya, E. A. (1989). *Nonparametric Estimation of Probability Densities and Regression Curves*. Kluwer Academic.

Obuchowski, N. (2005). Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary. *Statistcs in Medicine* **20**, 3261–3278.

Obuchowski, N. (2006). An roc-type measure of diagnostic accuracy when the gold standard is continuous-scale. *Statistcs in Medicine* **25**, 481–493.

Pepe, M. (2003). *The Statistical Rvaluation of Medical Tests for Classification and Prediction*. University Press, Oxford.

Pepe, M, Thompson, M. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123–140.

Pfeiffer, R., Castle, P. (2005). With or without a goldstandard. *Epidemiology* **16**, .

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: Ii. radical prostatectomy treated patients. *Journal of Urology* **141**, 1076–1083.

Su, J., Liu, J. (1993). Linear combinations of multiple diagnostic markers. *J. Am. Statist. Ass.* **88**, 1350–1355.

Wang, Z., Chang, Y., Ying, Z., Zhu, L., Yang, Y. (2007a). A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve. *Bioinformatics* **23**, 2788–2794.

Wang, C., Turnbull, B., Gröhn, Y., Nielsen, S. (2007b). Nonparametric estimation of roc curves based on bayesian models when the true disease state is unknown. *Journal of Agriculture, Biological and Enviromental Statistics* **12**.

Willems, J., Saunders, J., Hunt, D., Schorling, J. (1997). Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Medical Journal* **90**, 814–820.

Zhou, X.-H., Castelluccio, P., Zhou, C. (2005). Nonparametric estimation of roc curves in the absence of a gold standard. *Biometrics* **61**, 600–609.