



Linear Combinations of Multiple Diagnostic Markers

John Q. Su, Jun S. Liu

Journal of the American Statistical Association, Volume 88, Issue 424 (Dec., 1993),
1350-1355.

Your use of the JSTOR database indicates your acceptance of JSTOR's Terms and Conditions of Use. A copy of JSTOR's Terms and Conditions of Use is available at <http://www.jstor.org/about/terms.html>, by contacting JSTOR at jstor-info@umich.edu, or by calling JSTOR at (888)388-3574, (734)998-9101 or (FAX) (734)998-9113. No part of a JSTOR transmission may be copied, downloaded, stored, further transmitted, transferred, distributed, altered, or otherwise used, in any form or by any means, except: (1) one stored electronic and one paper copy of any article solely for your personal, non-commercial use, or (2) with prior written permission of JSTOR and the publisher of the article or other text.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Journal of the American Statistical Association is published by American Statistical Association. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/asa.html>.

Journal of the American Statistical Association
©1993 American Statistical Association

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2000 JSTOR

Linear Combinations of Multiple Diagnostic Markers

John Q. Su and Jun S. Liu*

The receiver operating characteristic (ROC) curve is a simple and meaningful measure to assess the usefulness of diagnostic markers. To use the information carried by multiple markers, we note that Fisher's linear discriminant function provides a linear combination of markers to maximize the sensitivity over the entire specificity range uniformly under the multivariate normal distribution model with proportional covariance matrices. With no restriction on covariance matrices, we also provide a solution of the best linear combination of markers in the sense that the area under the ROC curve of this combination is maximized among all possible linear combinations. We illustrate both situations discussed in the article with a cancer clinical trial data.

KEY WORDS: Linear discriminant function; Receiver operating characteristic curve; Sensitivity; Specificity.

1. INTRODUCTION

In medical applications, many different diagnostic markers are used to detect physiologic abnormalities or irregularities. A marker's usefulness is generally assessed based on its "sensitivity" and "specificity," defined as follows. Suppose that a diagnostic marker is used on a control group (disease condition negative) of m people and a disease group (disease condition positive) of n people; then obtain m values, X_i , $i = 1, \dots, m$, from each person in the control group and n values, Y_j , $j = 1, \dots, n$, from each person in the disease group. To use the outcomes obtained from these $m + n$ subjects, we imagine that X 's and Y 's are iid samples from two different distributions with cumulative distribution functions $F(\cdot)$ and $G(\cdot)$. Then for a cutoff value c , the marker's specificity is defined as $F(c)$, and the corresponding sensitivity is defined as $1 - G(c)$. The diagnostic marker is called "positive" when its value exceeds some given cutoff value c . As c takes on all possible values, the diagnostic marker generates a locus of $\{F(c), 1 - G(c)\}$; the curve is termed a receiver operating characteristic curve (ROC) by some investigators (Swets and Pickett 1982). Bamber (1975) noted that the area under this curve is equal to $P(Y > X)$.

In cases where two or more diagnostic markers are involved, various methods for evaluating and comparing the performance of diagnostic markers have been proposed. Hanley and McNeil (1982, 1983), McClish (1987), and DeLong, DeLong, and Clarke-Pearson (1988) have presented comparisons of markers based on the difference of areas under ROC curves. Greenhouse and Mantel (1950) and Linnit (1987) considered the comparison of two sensitivities at a single fixed level of specificity. Wieand, Gail, James, and James (1989) proposed a class of statistics for such comparisons based on a weighed average of sensitivities, so that one can compare the sensitivities of these diagnostic markers over restricted ranges of specificity, or over the entire area under ROC curves, or at a fixed common specificity.

But different markers are usually sensitive to different aspects of disease in the real situation. It is important to use two or more good diagnostic markers simultaneously so that one may obtain a new diagnostic marker with higher sen-

sitivity. Of particular interest to us are the linear combinations of two or more markers, where the problem becomes one of linear discriminant analysis or classification. The theory of linear discriminant analysis has been studied extensively. Fisher (1936) considered the classical linear discriminant problem by choosing the coefficients so that the ratio of the difference of means of the linear combination in the two groups to its variance is maximized. Welch (1939) suggested a procedure to minimize the total probability of misclassification, and Von Mises (1945) suggested minimizing the maximum probability of misclassification in the two groups. Various authors have suggested minimizing the total cost of misclassification based on the fact that misclassifications have different cost in the different groups. When the covariance matrices for two multivariate normal distributions are proportional, it is a classical result that the best linear discriminant uniquely exists. In this article we note that such best linear discriminant also results in the best ROC curve that dominates all the other linear combinations.

Anderson and Bahadur (1962) studied the procedures for classifying two multivariate normal distributions with unequal covariance matrices using linear combinations and provided a complete class of admissible rules. Among all these admissible rules, they considered a minimax-type discriminant procedure, a procedure for minimizing one probability of misclassification for a specified probability of the other, and a procedure with prior probabilities. In this article we provide the optimal linear combination among all admissible rules of Anderson and Bahadur under the criterion that the area under the corresponding ROC curve of the combination is maximized; hence the ROC criterion. The reason for using ROC criterion is that one may wish to use the best diagnostic marker or combination of markers that behaves well under at least a few particular specificities, if not under the entire range or restricted range of specificity.

Section 2 discusses the best linear combination when the two covariance matrices of normal and disease groups are proportional to each other and draws connections with the classical linear discriminant. Section 3 is devoted to the non-proportional cases, in which an optimal linear combination procedure is proposed under ROC criterion. Section 4 considers the uncertainty of the estimated combination coeffi-

* John Q. Su is Biostatistician, Research Division of Syntex INC., Palo Alto, CA 94303. Jun S. Liu is Assistant Professor of Statistics, Department of Statistics, Harvard University, Cambridge, MA 02138. The authors thank the associate editor, the referees, and Sam Wieand for their helpful comments on the article.

cients from observations. Section 5 presents an analysis of a real example, and Section 6 provides further remarks.

2. DOMINATING ROC CURVE FOR THE CASE OF PROPORTIONAL COVARIANCE MATRICES

In this section it is shown that Fisher’s best linear discriminant maximizes sensitivity uniformly over the entire range of specificity when the two distributions are assumed normal with proportional covariance matrices. A simple argument is given for the two-marker case, which can be generalized to the multiple-marker situation.

Let $X_i, i = 1, \dots, m$ be the marker values of m patients in the control group, where $X_i = (X_{i1}, X_{i2})^T$ and $X_{ik}, k = 1, 2$ is the k th marker value of the i th patient in this group; let $Y_j, j = 1, \dots, n$ be the marker values of n patients in the disease group. Suppose that X and Y are both normally distributed bivariate random variables. We consider the situation where the two covariance matrices are equal or, more generally, proportional to each other. That is, $X \sim N(\mu_x, \Sigma)$ and $Y \sim N(\mu_y, \sigma^2 \Sigma)$, so that the two covariance matrices differ by an unknown scaling factor σ^2 . Because it is generally difficult to deal with multidimensional data, we are interested in reducing dimensionality by constructing an effective linear combination of different markers, which implies that we look for certain “good” linear coefficients (α, β) so that for any marker values X or Y , one-dimensional random variables, $U = (\alpha, \beta)X$ and $V = (\alpha, \beta)Y$, are constructed.

In medical practice, proportionality among covariance matrices seems to be a sensible assumption when the populations of non-disease and disease groups have similar performances in different aspects, but with shifted means and different scales of variance. In the first part of this section, we show that Fisher’s (1936) linear discriminant coefficient dominates all the other possible linear combinations in the sense that it provides the highest sensitivity uniformly at any given specificity. Similar nice properties for the best linear discriminant have also been given by Anderson and Bahadur (1962) and Lachenbruch (1975), among others.

Lemma 2.1. If $\Sigma = I$ and $\mu_x = (0, 0)^T, \mu_y = (\mu, 0)^T$, where μ is a positive constant, then the coefficients for the best linear combination that provides the largest area under the ROC curve is $(\alpha_0, \beta_0) \propto (1, 0)$; that is, proportional to the mean vector $\mu_y - \mu_x$.

The following theorem can be obtained from Lemma 2.1 with linear transformations and the properties of the normal distribution.

Theorem 2.1. The best linear combination coefficient is just Fisher’s discriminant coefficient: $(\alpha_0, \beta_0) \propto (\mu_y - \mu_x)^T \Sigma^{-1}$. Furthermore, such a combination results in a ROC curve dominating all the others.

Based on these coefficients, we can obtain the specificity and the sensitivity of the best linear combination and then generate the ROC curve for the best linear combination. Let the coefficients for the best linear combination be $a = (\alpha_0, \beta_0)^T$; we note that $U = a^T X \sim N(a^T \mu_x, a^T \Sigma_x a)$ and $V = a^T Y \sim N(a^T \mu_y, a^T \Sigma_y a)$. Let the specificity be $F_a(c)$,

where $F_a(\cdot)$ is the cumulative distribution function of U , and let the sensitivity be $H_a(c) = 1 - G_a(c)$, where $G_a(\cdot)$ is the cumulative distribution function of V . Then for any given p_0 , there is a constant c such that

$$F_a(c) = \Phi\left\{\frac{c - a^T \mu_x}{\sqrt{a^T \Sigma_x a}}\right\} = p_0, \tag{1}$$

and the cutoff point c can be expressed as

$$c = a^T \mu_x + \Phi^{-1}(p_0) \sqrt{a^T \Sigma_x a}. \tag{2}$$

The corresponding sensitivity can be obtained as

$$H_a(c) = 1 - \Phi\left(\frac{a^T(\mu_x - \mu_y) + \Phi^{-1}(p_0) \sqrt{a^T \Sigma_x a}}{\sqrt{a^T \Sigma_y a}}\right). \tag{3}$$

This result for the situation of proportional covariance matrices can be generalized to higher dimensions with no difficulty. In general, $X \sim N(\mu_x, \Sigma_x)$ and $Y \sim N(\mu_y, \Sigma_y)$; both can be p dimensional. Then Theorem 2.1 is still true; the best projection vector is $a \propto (\mu_y - \mu_x)^T \Sigma^{-1}$, and the ROC curve is generated exactly the same way by using (1), (2), and (3) as we have just shown.

For the case of nonproportional covariance matrices, Anderson and Bahadur (1962) provided a complete class of admissible rules for classifying two multivariate normal distributions using linear combinations. In other words, there generally does not exist a dominating combination that is uniformly better than any others. Therefore, we cannot hope for a dominating ROC curve that would produce the highest sensitivity at every specificity uniformly, unless some special assumptions on the covariance matrices are made. Among all these admissible rules, Anderson and Bahadur (1962) also considered various other criteria to give “best” linear discriminant procedure. For example, they presented the Bayesian solution for a given prior distribution of population proportions and considered the admissible procedure that minimizes one probability of misclassification for a specified probability of the other. The latter procedure can be used in our case, minimizing the sensitivity for a specified specificity. But this procedure depends on the specified probability of misclassification. In the next section we derive the best admissible linear combination under the ROC criterion despite nonproportionality.

3. ROC CRITERION AND THE BEST LINEAR COMBINATION WITH NO RESTRICTION ON COVARIANCE MATRICES

Numerous indices have been used to summarize the information contained in the ROC curve. The most popular single quantitative index of diagnostic accuracy, the ROC criterion, is defined as the area under the ROC curve. In one marker situation, the larger the area, the more information the marker provides. Generalizing this criterion, we call a linear combination coefficient vector a the best under ROC criterion if the area under the ROC curve generated by $a^T X$ and $a^T Y$ is the largest among all linear combinations. In this section we generalize the result of Section 2 to the case of p markers with nonproportional covariance matrices to pro-

vide the unique solution to the best linear combination under the ROC criterion. Before going on to the higher-dimensional case, the following lemma for one-dimensional cumulative distribution functions is useful.

Lemma 3.1. Let $F(\cdot)$ and $G(\cdot)$ be two one-dimensional cumulative distribution functions representing the marker distributions for control group and disease group. The area under the ROC curve can be expressed as $A = E[F(Y)]$, where $Y \sim G(\cdot)$.

The lemma follows easily from Bamber's (1975) result that $A = P(Y > X)$ and from a conditional expectation argument.

Generally, let \mathbf{X} , a p -dimensional random variable, be a vector of values of p different markers obtained from a randomly picked patient in the non-disease group, and let \mathbf{Y} , another p -dimensional random variable, be a similar vector of a randomly picked patient in the disease group. Then $\mathbf{X} \sim N(\mu_x, \Sigma_x)$ and $\mathbf{Y} \sim N(\mu_y, \Sigma_y)$, where μ_x and μ_y are $p \times 1$ vectors and Σ_x and Σ_y are $p \times p$ positive definite matrices. Consider the linear combination, $U = \mathbf{a}^T \mathbf{X}$ and $V = \mathbf{a}^T \mathbf{Y}$, where $\mathbf{a} = (a_1, \dots, a_p)^T$.

Lemma 3.2. If $\Sigma_x = I$ and $\mu_x = \mathbf{0}$, $\mu_y = \mu$, where $\mathbf{0}$ and μ are $p \times 1$ vectors, and $\Sigma_y = \Sigma$, which is a positive definite matrix, then the coefficients for the best linear combination are $\mathbf{a}_0 \propto (I + \Sigma)^{-1} \mu$.

Using Lemma 3.2 and a linear transformation of the random variables, we obtain the following theorem.

Theorem 3.1. The coefficients for the best linear combination are

$$\mathbf{a}_0 \propto \Sigma_x^{-1/2} (I + \Sigma_x^{-1/2} \Sigma_y \Sigma_x^{-1/2})^{-1} \Sigma_x^{-1/2} \mu = (\Sigma_x + \Sigma_y)^{-1} \mu,$$

where $\mu = \mu_y - \mu_x$.

Using the coefficient vector \mathbf{a}_0 obtained in the theorem, one can treat $\mathbf{a}_0^T \mathbf{X}$ and $\mathbf{a}_0^T \mathbf{Y}$ as one-dimensional random variables and construct the ROC curve of the linear combination in the same way by using (1), (2), and (3) in Section 2. The area under the corresponding ROC curve based on such construction is provided in the following corollary.

Corollary 3.1. Under the normality assumption on distributions of both control and disease groups, the area under the ROC curve of the optimal linear combination is

$$A = \Phi(\sqrt{\mu^T (\Sigma_x + \Sigma_y)^{-1} \mu}),$$

where $\mu = \mu_y - \mu_x$.

4. ESTIMATED MEANS AND COVARIANCE MATRICES

In most cases we do not know the true means and covariance matrix of any practical distribution; we can only estimate them from the sample at hand. In clinical practice we can assume that $X_1, \dots, X_m \sim N(\mu_x, \Sigma_x)$ are iid samples from the control population and $Y_1, \dots, Y_n \sim N(\mu_y, \Sigma_y)$ are iid samples from the disease population, where μ_x and μ_y are p -dimensional vectors and Σ_x and Σ_y are $p \times p$ ma-

trices. Let

$$S = \sum_{i=1}^m (X_i - \bar{X})(X_i - \bar{X})^T + \sum_{j=1}^n (Y_j - \bar{Y})(Y_j - \bar{Y})^T$$

be the regular pooled sum of squares. The following result holds.

Theorem 4.1. If it is known that $\Sigma_x = \Sigma_y = \Sigma$, then $\hat{T} = (m + n - p - 3)S^{-1}$ is an unbiased estimate of Σ^{-1} and $\hat{\mathbf{a}}_0 = \hat{T}(\bar{Y} - \bar{X})$ is an unbiased estimate of the best linear combination vector $\mathbf{a}_0 = \Sigma^{-1} \mu$, where $\mu = \mu_y - \mu_x$. Furthermore,

$$E\|\hat{\mathbf{a}}_0 - \mathbf{a}_0\|_2^2 \stackrel{\text{def}}{=} E[(\hat{\mathbf{a}}_0 - \mathbf{a}_0)^T \Sigma (\hat{\mathbf{a}}_0 - \mathbf{a}_0)]$$

$$\approx \frac{p}{m + n - p - 3} \times \left(\mu^T \Sigma^{-1} \mu + \frac{(m + n)(m + n - 3)}{mn} \right),$$

where the order of error of the approximation is $O((m + n)^{-1})$. The latter formula can be used to assess the uncertainty of the estimation.

The general case of unequal covariance matrices is similar to a multidimensional Behrens-Fisher problem. It is well known that the problem has no close-form distributional solution in general. In this case the two covariance matrices are estimated via the sample sums of squares, $S_x = \sum (X_i - \bar{X})(X_i - \bar{X})^T$ and $S_y = \sum (Y_j - \bar{Y})(Y_j - \bar{Y})^T$. From these derivations we have

$$S_x \sim \text{Wishart}(m - 1, p, \Sigma_x),$$

$$S_y \sim \text{Wishart}(n - 1, p, \Sigma_y).$$

$S_x/(m - 1)$ and $S_y/(n - 1)$ are maximum likelihood estimators (MLE's) of Σ_x and Σ_y with mean squared errors (MSE's) of order $1/m$ and $1/n$. Using the δ method, we see immediately that

$$\left(\frac{S_x}{m - 1} + \frac{S_y}{n - 1} \right)^{-1} (\bar{Y} - \bar{X})$$

is a consistent estimate of $(\Sigma_x + \Sigma_y)^{-1} \mu$ with MSE of order $1/\min(n, m)$.

5. EXAMPLE

For illustration, we apply the preceding method to the data obtained from a clinical trial. The North Central Cancer Treatment Group (NCCTG) and Mayo Clinic recently completed a trial designed to determine whether or not four monoclonal antibodies—carcinoembryonic antigen (CEA), CA 19-9, CA 125, and CA 72-4—were prognostic for recurrence of colorectal cancer. All the patients entered into this study had nonmetastatic colorectal cancer and had undergone surgery that removed the primary lesion, so that the patients were assumed to be disease-free following the surgery. All patients in the study were followed for recurrence. This follow-up included regular physical examinations. During the physical examination blood was drawn and analyzed, so that a numerical score was obtained for each of

the monoclonal antibodies. A detailed description of the medical aspects of the study were presented by Ritts and Wieand (1992). For simplicity, suppose that one considers only the two monoclonal antibodies CA 19-9 and CA 72-4. Suppose that one is only interested in those numerical values of these two antibodies measured right before recurrence for patients with recurrence and those values measured at the last follow-up for patients without recurrence.

In this example the distributions of both markers are skewed to the right, mainly because most of the marker value range in an interval from 0 to a positive number. But some patients with tumor recurrence have marker values that extend far beyond the normal range. To improve the normality, we take a logarithm transformation of the marker values. Let $X_i, i = 1, \dots, 91$ be the logarithm of antibody values of 91 patients without recurrence, where $X_i = (X_{i1}, X_{i2})^T$ and X_{i1} and X_{i2} are values of CA 19-9 and CA 72-4. Let $Y_j = (Y_{j1}, Y_{j2})^T, j = 1, \dots, 110$ be the logarithm of antibody values of 110 patients with recurrence. From the Q-Q plots, we find that the distributions of transformed marker values for the patients without recurrence are very close to normal, but there are still outliers in the recurrence group. One explanation of this is that the marker values for the patients could jump far beyond normal range at the time close to tumor recurrence.

Based on the behavior of the majority data, we assume that $X \sim N(\mu_x, \Sigma_x)$ and $Y \sim N(\mu_y, \Sigma_y)$. The standard normal estimation gives $\hat{\mu}_x = (2.49, .81)^T, \hat{\mu}_y = (3.77, 1.51)^T$, and

$$\hat{\Sigma}_x = \begin{pmatrix} .68 & .03 \\ .03 & .18 \end{pmatrix}, \quad \hat{\Sigma}_y = \begin{pmatrix} 3.97 & .69 \\ .69 & 1.42 \end{pmatrix}.$$

On the other hand, if the covariance matrices are assumed to be proportional to each other—that is, $X \sim N(\mu_x, \Sigma)$ and $Y \sim N(\mu_y, \sigma^2 \Sigma)$ —then maximum likelihood estimates of the unknowns are

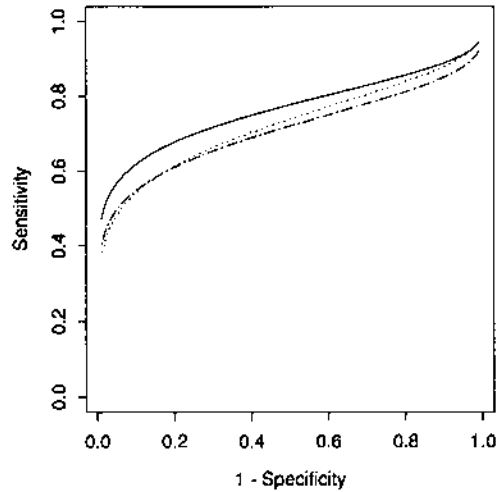


Figure 2. ROC Curves With Normality: Best Linear Combination Under the ROC criterion. —, combination; ·····, CA 19-9; - · - ·, CA 72-4.

$$\hat{\Sigma} = \begin{pmatrix} .61 & .07 \\ .07 & .21 \end{pmatrix}, \quad \hat{\sigma}^2 = 6.89.$$

Although the estimates differ fairly significantly, the preceding two procedures give very similar results on the combination coefficients. For the linear combination, $U = (\alpha, \beta)X$ and $V = (\alpha, \beta)Y$, the method in Section 2 gives the best coefficients as follows under the proportionality assumption, $(\alpha, \beta) \propto (1.00, 1.55)$, and the general method in Section 3 results $(\alpha, \beta) \propto (1.00, 1.50)$. Both pairs of coefficients give an ROC curve dominating the corresponding two single-variable ROC curves of CA 19-9 and CA 72-4, as demonstrated in Figures 1 and 2. The sensitivity of the combined marker improves significantly on the high specificity end. The area under the ROC curve corresponding to the combined marker, with both coefficients, is .76, compared to the area .72 when CA 19-9 is used alone, and the area .71 when CA 72-2 is used alone.

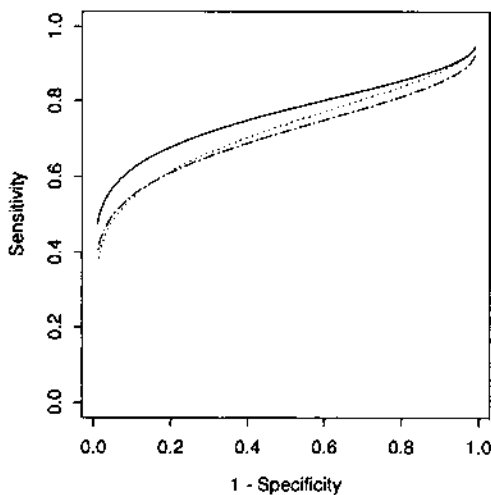


Figure 1. ROC Curves With Normality: Best Linear Combination Under Proportional Assumption. —, combination; ·····, CA 19-9; - · - ·, CA 72-4.

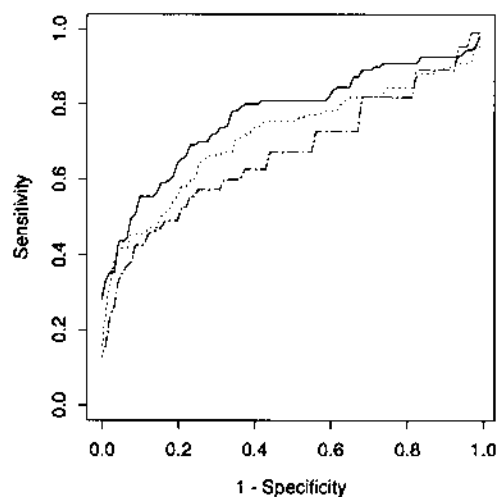


Figure 3. Empirical ROC Curves: Best Linear Combination Under Proportional Assumption. —, combination; ·····, CA 19-9, - · - ·, CA 72-4.

Figures 1 and 2 show that the two curves using two different estimated combination coefficients are very close to each other, clearly indicating that the best linear combination works very well and would be the marker of choice in our case.

6. REMARKS

1. The means and covariances for most populations in practice are unknown. With observations and assumed models, a maximum likelihood estimate procedure is adequate. In our example with assumed proportionality, a simple Fortran program of maximizing the likelihood is needed numerically to evaluate the common matrix $\hat{\Sigma}$ and the ratio $\hat{\sigma}^2$. The uncertainty involved in such estimation procedures, which we do not intend to address here, should not be neglected and may be estimated by standard δ methods.
2. The best combination coefficients, estimated under normality, also work quite reasonably with actual nonnormal distributions, where the ROC curves are estimated empirically. Figures 3 and 4 illustrate the resulting ROC curves of two combined markers using those coefficients, as well as the individual ROC curves of CA 19.9 and CA 72.4. Indeed, with normality assumption, we choose the best combination of the two distributions based on their first two moments. Such a procedure will be beneficial if the underlying true distributions are not too skewed, and the Box-Cox transformation can be used to improve normality. When the normality assumption of X and Y fails, there will be in general some degradations in the performance of our method, similar to that of a linear discriminant analysis. However, when the distributions are still proportional ellipsoidal (e.g., multidimensional t distributions with proportional structural matrices), the conclusions of Section 2 still hold, but the solution in Section 3 is less clear. Some of these features, together with a nonparametric method, are currently under investigation.
3. If the two covariance matrices are not proportional to each other, then the ROC curves of individual markers may cross over each other, and the ROC curve of the new marker

resulting from the best linear combination of all markers under ROC criterion in Section 3 may cross with them as well. This phenomenon has been predicted by Anderson and Bahadur (1962), who showed that there is no linear combination superior to the others over the entire range. If one would like to find a linear combination which is superior at a single given specificity, the procedure of Anderson and Bahadur (1962) can be easily applied.

4. As one referee points out, using logistic regression to identify tests that best predict presence or absence of disease is also common. On one hand the logistic regression is generally less efficient than the normal discriminant analysis when the normal assumption is met (Efron 1975; Ruiz-Velasco 1991) and thus is less efficient than the method we proposed. As was exposed by Cox and Snell (1989), however, the logistic regression presupposes a stable statistical relation such that once a vector of explanatory variables, x , is given, then the probability that this individual belongs to one of the two groups is determined. The distribution of x is, therefore, irrelevant. This feature renders the logistic regression more robust than normal discrimination and also explains that its lower efficiency is due to its less use of information, or assumption (distributional information of x is not used). Because our method is a variation of normal discrimination, the same comparisons hold between logistic regression and our method.

APPENDIX: PROOFS

Proof of Lemma 2.1. It is obvious that $U \sim N(0, \alpha^2 + \beta^2)$ and $V \sim N(\alpha\mu, \sigma^2(\alpha^2 + \beta^2))$. Let the specificity be $F(c)$, where $F(\cdot)$ is the cumulative distribution function of U , and let the sensitivity be $H(c) = 1 - G(c)$, where $G(\cdot)$ is the cumulative distribution function of V . Then for any given p_0 , there is a constant c such that $F(c) = \Phi(c/\sqrt{\alpha^2 + \beta^2}) = p_0$. Thus

$$\begin{aligned}
 H(c) &= 1 - \Phi\left(\frac{c - \alpha\mu}{\sigma\sqrt{\alpha^2 + \beta^2}}\right) \\
 &= 1 - \Phi\left(\frac{\Phi^{-1}(p_0)}{\sigma} - \frac{\alpha\mu}{\sigma\sqrt{\alpha^2 + \beta^2}}\right).
 \end{aligned}$$

It is clear that $H(c)$ is maximized when $\alpha > 0$ and $\beta = 0$.

Proof of Theorem 2.1. It is noticed that shifting the two distributions by a common mean does not affect the solution of our best coefficient of linear combination. Because Σ is a positive definite matrix, we can find a positive definite matrix $\Sigma^{-1/2}$ such that $\Sigma^{-1} = (\Sigma^{-1/2})^2$. Hence $X_0 = \Sigma^{-1/2}(X - \mu_x) \sim N(0, I)$, and $Y_0 = \Sigma^{-1/2}(Y - \mu_x) \sim N(\Sigma^{-1/2}(\mu_y - \mu_x), \sigma^2 I)$. Now we understand from Lemma 2.1 that the vector of best projection is proportional to the mean vector $\Sigma^{-1/2}(\mu_y - \mu_x)$ for X_0 and Y_0 , which is

$$\{\Sigma^{-1/2}(\mu_y - \mu_x)\}^T \Sigma^{-1/2} = (\mu_y - \mu_x)^T \Sigma^{-1}$$

for the original X and Y .

Proof of Lemma 3.2. For any coefficient vector a of a linear combination, the new pair of random variables are normally distributed as $U \sim N(0, a^T a)$ and $V \sim N(a^T \mu, a^T \Sigma a)$. Assuming that the distribution function of U is $F(\cdot)$, by Lemma (3.1) the area under the ROC curve for such linear combination is $A = E[F(V)] = E[\Phi(\alpha Z + \beta)]$, where

$$Z \sim N(0, 1), \quad \alpha(a) = \sqrt{\frac{a^T \Sigma a}{a^T a}} \quad \text{and} \quad \beta(a) = \frac{a^T \mu}{\sqrt{a^T a}}.$$

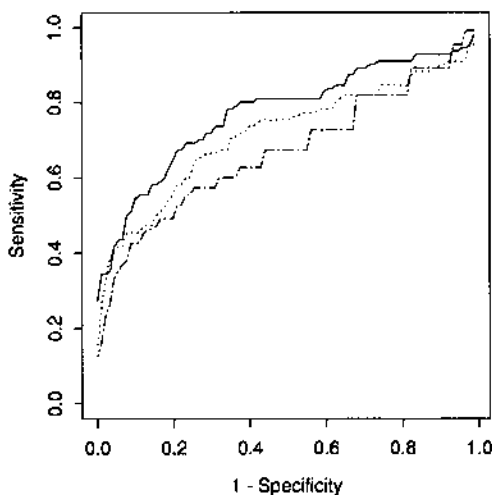


Figure 4. Empirical ROC Curves: Best Linear Combination Under the ROC Criterion. —, combination; ····, CA 19.9; - · - ·, CA 72.4.

Differentiating with respect to \mathbf{a} , a nice result can be obtained:

$$\begin{aligned} \frac{\partial A}{\partial \mathbf{a}} &= E[\phi(\alpha Z + \beta)] \frac{\partial \beta}{\partial \mathbf{a}} + E[Z\phi(\alpha Z + \beta)] \frac{\partial \alpha}{\partial \mathbf{a}} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1+\alpha^2}} \exp\left\{-\frac{\beta^2}{2(1+\alpha^2)}\right\} \left[\frac{\partial \beta}{\partial \mathbf{a}} - \frac{\alpha\beta}{1+\alpha^2} \frac{\partial \alpha}{\partial \mathbf{a}} \right] \\ &= \frac{\sqrt{1+\alpha^2}}{2\sqrt{2\pi}} \exp\left\{-\frac{\beta^2}{2(1+\alpha^2)}\right\} \frac{\partial}{\partial \mathbf{a}} \left(\frac{\beta^2}{1+\alpha^2} \right), \end{aligned}$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. Thus

$$\frac{\partial A}{\partial \mathbf{a}} = 0 \iff \frac{\partial}{\partial \mathbf{a}} \left(\frac{\beta^2}{1+\alpha^2} \right) = 0.$$

So maximizing the area A is equivalent to maximizing

$$\frac{\beta^2}{1+\alpha^2} = \frac{(\mathbf{a}^T \boldsymbol{\mu})^2}{\mathbf{a}^T (I + \boldsymbol{\Sigma}) \mathbf{a}}.$$

It is known that this normalized quadratic form can be maximized by $\mathbf{a} \propto (I + \boldsymbol{\Sigma})^{-1} \boldsymbol{\mu}$ (Seber 1977, p. 388.)

Proof of Theorem 3.1. Again we observe that a common shifting for the two distributions will not affect the solution to the best coefficient of linear combination. Because $\boldsymbol{\Sigma}_x$ is positive definite, we can legitimately write down its "square root," $\boldsymbol{\Sigma}_x^{-1/2}$. Thus $\mathbf{X}_0 = \boldsymbol{\Sigma}_x^{-1/2}(\mathbf{X} - \boldsymbol{\mu}_x)$ is distributed as $N(\mathbf{0}, I)$, and $\mathbf{Y}_0 = \boldsymbol{\Sigma}_y^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}_y)$ is distributed as $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\mu}_0 = \boldsymbol{\Sigma}_x^{-1/2}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x) = \boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_y^{-1/2} \boldsymbol{\Sigma}_y \boldsymbol{\Sigma}_x^{-1/2}$. We apply Lemma 3.2 to see that the best combination vector for \mathbf{X}_0 and \mathbf{Y}_0 is proportional to $(I + \boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{\Sigma}_y \boldsymbol{\Sigma}_x^{-1/2})^{-1} \boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{\mu}$, and thus the conclusion of the theorem follows.

Proof of Corollary 3.1. From previous arguments,

$$\begin{aligned} A &= E(\Phi(\alpha Z + \beta)) \\ &= E(\Pr(Z' \leq \alpha Z + \beta | Z)) \\ &= \Pr(Z' - \alpha Z \leq \beta) = \Phi\left(\frac{\beta}{\sqrt{1+\alpha^2}}\right), \end{aligned}$$

where $\alpha = \sqrt{(\mathbf{a}^T \boldsymbol{\Sigma}_y \mathbf{a}) / (\mathbf{a}^T \boldsymbol{\Sigma}_x \mathbf{a})}$, $\beta = \mathbf{a}^T \boldsymbol{\mu} / \sqrt{\mathbf{a}^T \boldsymbol{\Sigma}_x \mathbf{a}}$, and $Z, Z' \sim N(0, 1)$. The last equality follows from the fact that Z' and Z are independently distributed as $N(0, 1)$. Furthermore, because the best linear coefficient vector from Theorem 3.1 is $\propto (\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)^{-1} \boldsymbol{\mu}$, the conclusion follows from the fact that

$$\mathbf{a}^T \boldsymbol{\mu} = \mathbf{a}^T \boldsymbol{\Sigma}_x \mathbf{a} + \mathbf{a}^T \boldsymbol{\Sigma}_y \mathbf{a} = \boldsymbol{\mu}^T (\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)^{-1} \boldsymbol{\mu}.$$

Proof of Theorem 4.1. It follows from the exposition of Kshirsagar (1972) that $S \sim \text{Wishart}(p, m+n-2, \boldsymbol{\Sigma})$ and $E(S^{-1}) = (m+n-p-3)^{-1} \boldsymbol{\Sigma}^{-1}$. Thus the first result easily follows. The unbiasedness of $\hat{\mathbf{a}}_0$ follows immediately if we notice that $\bar{Y} - \bar{X}$ is independent of S . The latter equation follows from a fact about the Wishart distribution (Kshirsagar 1972, pp. 72, 116):

$$\begin{aligned} E(S^{-1} \boldsymbol{\Sigma} S^{-1}) &= \frac{m+n-3}{(m+n-p-2)(m+n-p-3)(m+n-p-4)} \boldsymbol{\Sigma}^{-1}. \end{aligned}$$

By the independence between $\bar{Y} - \bar{X}$ and S ,

$$\begin{aligned} E\|\hat{\mathbf{a}}_0 - \mathbf{a}_0\|_{\boldsymbol{\Sigma}}^2 &= (m+n-p-3)^2 \\ &\quad \times E((\bar{Y} - \bar{X})^T S^{-1} \boldsymbol{\Sigma} S^{-1} (\bar{Y} - \bar{X})) - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ &\approx \frac{m+n-3}{m+n-p-3} \\ &\quad \times E((\bar{Y} - \bar{X})^T \boldsymbol{\Sigma}^{-1} (\bar{Y} - \bar{X})) - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ &= \frac{p}{m+n-p-3} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ &\quad + \frac{m+n-3}{m+n-p-3} \left(\frac{1}{n} + \frac{1}{m} \right) p. \end{aligned}$$

Hence the theorem is proved.

[Received June 1992. Revised January 1993.]

REFERENCES

Anderson, T. W., and Bahadur, R. R. (1962), "Classification Into Two Multivariate Normal Distributions With Different Covariance Matrices," *The Annals of Mathematical Statistics*, 33, 420-431.

Bamber, D. (1975), "The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph," *Journal of Mathematical Psychology*, 12, 387-415.

Cox, D. R., and Snell, E. J. (1989), *Analysis of Binary Data* (2nd ed.) London: Chapman & Hall.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988), "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, 44, 837-845.

Efron, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, 70, 892-898.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179-188.

Greenhouse, S. W., and Mantel, N. (1950), "Evaluation of Diagnostic Tests," *Biometrics*, 6, 399-412.

Hanley, J. A., and McNeil, B. J. (1982), "The Meaning and Use of the Area Under the Receiver Operating Characteristic (ROC) Curve," *Radiology*, 143, 29-36.

——— (1983), "A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived From the Same Cases," *Radiology*, 148, 839-843.

Kshirsagar, A. M. (1972), *Multivariate Analysis*. New York: Marcel Dekker.

Linnet, K. (1987), "Comparison of Quantitative Diagnostic Tests: Type I Error, Power, and Sample Size," *Statistics in Medicine*, 6, 147-158.

Lachenbruch, P. A. (1975), *Discriminant Analysis*. New York: Hafner Press.

McClish, D. K. (1987), "Comparing the Areas Under More Than Two Independent ROC Curves," *Medical Decision Making*, 7, 149-155.

Ritts, R., and Wieand, H. S. (1992), "Monoclonal Antibody Analysis," in preparation.

Ruiz-Velasco, S. (1991), "Asymptotic Efficiency of Logistic Regression Relative to Linear Discriminant Analysis," *Biometrika*, 78, 235-243.

Seber, G. A. F. (1977), *Linear Regression Analysis*. New York: John Wiley.

Swets, J. A., and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press.

Von Mises, R. (1945), "On the Classification of Observation Data Into Distinct Groups," *Annals of Mathematical Statistics*, 16, 68-73.

Welch, B. L. (1939), "Note on Discriminant Functions," *Biometrika*, 31, 218-220.

Wieand, H. S., Gail, M. H., James, B. R., and James, K. L. (1989), "A Family of Nonparametric Statistics for Comparing Diagnostic Markers With Paired or Unpaired Data," *Biometrika*, 76, 585-592.