The Annals of Statistics
1996, Vol. 24, No. 1, 25–40

NONPARAMETRIC AND SEMIPARAMETRIC ESTIMATION OF THE RECEIVER OPERATING CHARACTERISTIC CURVE

BY FUSHING HSIEH[1] AND BRUCE W. TURNBULL[2]

National Taiwan University and Cornell University

The receiver operating characteristic (ROC) curve describes the performance of a diagnostic test used to discriminate between healthy and diseased individuals based on a variable measured on a continuous scale. The data consist of a training set of m responses $X_1, \dots, X_m$ from healthy individuals and n responses $Y_1, \dots, Y_n$ from diseased individuals. The responses are assumed i.i.d. from unknown distributions F and G, respectively. We consider estimation of the ROC curve defined by $1 - G(F^{-1}(1 - t))$ for $0 \le t \le 1$ or, equivalently, the ordinal dominance curve (ODC) given by $F(G^{-1}(t))$. First we consider nonparametric estimators based on empirical distribution functions and derive asymptotic properties. Next we consider the so-called semiparametric "binormal" model, in which it is assumed that the distributions F and G are normal after some unknown monotonic transformation of the measurement scale. For this model, we propose a generalized least squares procedure and compare it with the estimation algorithm of Dorfman and Alf, which is based on grouped data. Asymptotic results are obtained; small sample properties are examined via a simulation study. Finally, we describe a minimum distance estimator for the ROC curve, which does not require grouping the data.

1. Introduction. A diagnostic test giving a measurement on a continuous scale is used to classify patients into either "healthy" or "diseased" categories. Typically, a cutoff point, c, is selected, and patients with test results greater than this are classified as "diseased," otherwise as "healthy" or "normal." The test score of a healthy patient is represented as a real random variable X with distribution function F and density f. Similarly a diseased patient's score will be denoted by Y with distribution function G and density g; X and Y are independent.

The sensitivity of the test is defined as $\mathrm{SE}(c) = 1 - G(c)$, which is the probability of correctly classifying a diseased individual when cutoff point c is used. Similarly we define the test's specificity $\mathrm{SP}(c) = F(c)$ as the probability of correctly classifying a healthy patient. Clearly these are the complements

Received September 1992; revised June 1995.

[1] Supported in part by a fellowship from the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell University.

[2] Supported in part by NIH Grant CA-66218.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62H30, 62G20.

Key words and phrases. Binormal model, classification, diagnostic test, discrimination, nonparametric estimation, receiver operating characteristic (ROC), semiparametric models.

of the familiar Type I and Type II errors. The receiver operating characteristic curve (ROC) is defined as a plot of the "true positive fraction," SE(c), on the vertical axis versus the "false positive fraction," $1 - \mathrm{SP}(c)$, on the horizontal axis as c varies from $+\infty$ to $-\infty$. Equivalently, it can be viewed as a plot of $\mathrm{ROC}(t) = 1 - G(F^{-1}(1 - t))$ versus t, $0 \le t \le 1$. Bamber (1975) reverses the axes and defines the ordinal dominance curve (ODC) $F(G^{-1}(t))$, $0 \le t \le 1$, which is a plot of SP(c) versus $1 - \mathrm{SE}(c)$, or equivalently $F(c)$ versus $G(c)$ for $-\infty \le c \le \infty$. Typical ODC curves are illustrated in Figure 1. For a desirable diagnostic test the ODC and the ROC curves rise rapidly and then level out as the lower graphs of Figure 1.

It is straightforward to show that the ODC and the ROC curves have the following convenient properties:

1. Invariance under monotone increasing transformations of the measurement scale.
2. X is stochastically smaller than Y, that is, $F(c) \ge G(c)$ for all c, implies that the curves lie above the diagonal in the unit square.

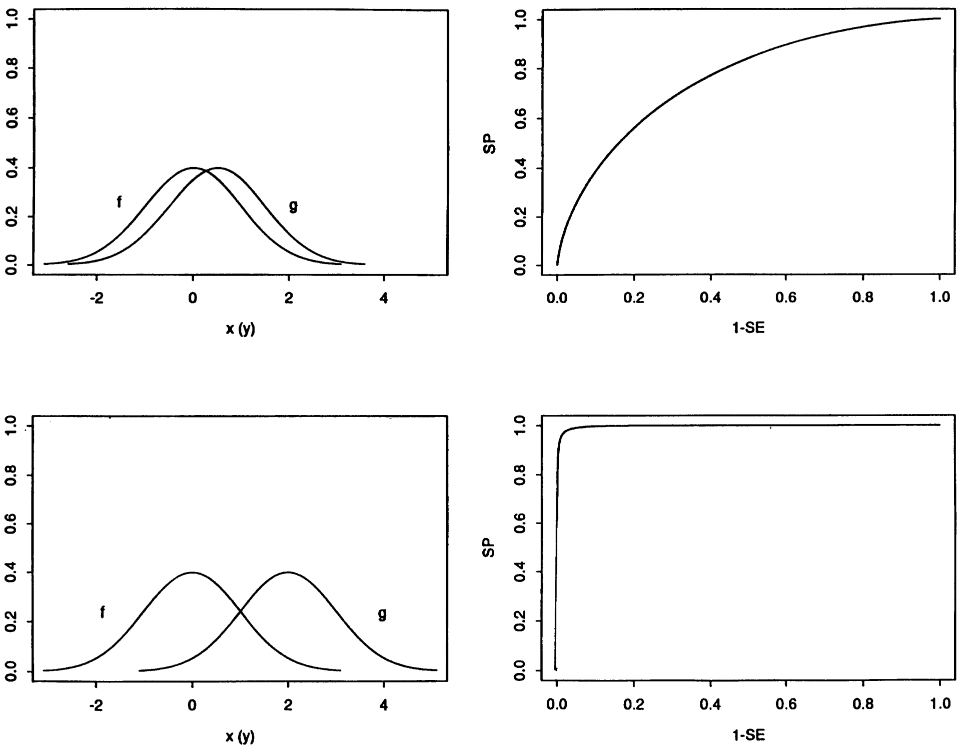

FIG. 1. *Two examples of densities f, g and their corresponding ODC curves. The diagnostic instrument represented by the lower curves is to be preferred.*

3. If the densities f and g have a monotone likelihood ratio, then the curves are concave.
4. The area under either curve is the probability $P[X < Y]$.

There are applications or potential applications of ROC curve analysis in almost every scientific field. Swets and Pickett [(1982), Appendix E] list almost 200 references in a variety of subject areas where ROC curve methods have been used. These and later references span such diverse areas as signal detection, psychology, polygraph lie detection, epidemiology, nutrition, radiology and general medical decision making, among others. The ROC curve is important because various measures of performance, or accuracy indices, for a given diagnostic test are based on the curve, thus allowing comparison of competing tests. Often the area or a weighted area $\int_0^1 \mathrm{ROC}(t)\,dW(t)$ under the curve is used. For a discussion, see Hilden (1991).

We suppose that a training data set $X_1, X_2, \dots, X_m$ of readings from the healthy population is available as is a set $Y_1, Y_2, \dots, Y_n$, from the diseased population. All observations are assumed mutually independent. Empirical ROC and ODC curves can be constructed by replacing F and G in the definitions by their corresponding sample cdfs F_m and G_n. Little work appears to have been done explicitly on the statistical properties of the empirical ROC curve itself. Of course the area under this curve is the Mann–Whitney statistic, the properties of which are well known; see, for example, Hanley and McNeil (1982).

Often some parametric form for F and G is assumed. Typically a normal distribution is assumed for both F and G, possibly after some given transformation of the X and Y scales, such as a logarithmic one. Interest then centers on estimating the small number of parameters that define F, G and, hence, ROC(t). For examples of this approach, see Brownie, Habicht and Cogill (1986) and Goddard and Hinberg (1990).

Another popular approach is to assume a so-called "binormal" model. This is a semiparametric approach that postulates the existence of some unspecified monotonic transformation H, say, of the measurement scale that simultaneously converts the F and G distributions to normal ones. Without loss of generality these can be taken, respectively, to be $N(0,1)$ and $N(\mu, \sigma^2)$, say. In this case the ODC curve has the known parametric form

$$(1) \qquad F(G^{-1}(t)) = \Phi(\mu + \sigma \Phi^{-1}(t)), \qquad t \in (0,1).$$

(Here Φ denotes the standard normal cdf.) Thus fitting a straight line to an empirical ODC curve plotted using normal probability scales on each axis yields a graphical test for the goodness-of-fit of the "binormal" assumption and graphical estimates of μ and σ, assuming the fit is adequate [Swets and Pickett (1982), page 30; Brownie, Habicht and Cogill (1986), Figure 2].

The "binormal" assumption was originally introduced in the field of psychology for use with ordered categorical variables. The measurements X, Y could take on only one of a finite set of values or categories. For example, with a five point rating scale, the categories might be labelled "probably healthy, possibly healthy, equivocal, possibly diseased, probably diseased." The problem is now a parametric one where the number of parameters equals the number of categories plus 1. (These comprise μ, σ and the $k - 1$ unknown cutpoints, where k is the number of categories.) Dorfman and Alf (1969) and Grey and Morgan (1972) described an iterative method for obtaining the maximum likelihood estimates of these parameters under the "binormal" assumption. Hanley (1988) discusses the justifications and applicability of the "binormal" assumption for rating data. Other distributional assumptions can be used instead of the normal: Ogilvie and Creelman (1968) used a logistic distribution. When the measurements are on a continuous scale as considered in this paper, the data must be grouped in order to apply the Dorfman and Alf (1969) procedure. Clearly this will lead to some loss in efficiency.

In the next section we describe asymptotic properties of the empirical ROC and ODC curves. Application of these results to the nonparametric estimation of $P(X < Y)$ is discussed. In Section 3, we consider estimation under the semiparametric "binormal" model. Using the theory of empirical processes, we examine the asymptotic properties of a generalized least squares estimator which we propose, and show it is asymptotically equivalent to the MLE based on the Dorfman and Alf (1969) procedure for grouped data. Finally, in Section 4 we briefly consider a minimum distance approach which does not require grouping of the data. We indicate the robustness of this minimum distance estimate in a sense of locally asymptotic minimaxity (LAM).

2. The empirical ODC curve and estimation of $P(X < Y)$. As in Section 1, we denote the sample cdfs of the X and Y training data sets by $F_m(x)$ and $G_n(y)$, respectively. Also we define the empirical quantile function as $G_n^{-1}(t) = \inf\{y\colon G_n(y) \geq t\}$. Then the empirical ODC curve is defined as $F_m(G_n^{-1}(t))$, $0 < t < 1$, which for convenience we will also write as $F_m G_n^{-1}(t)$. For our asymptotic results we will always be assuming that the sample sizes are such that $m = m(n)$ and $n/m \to \lambda > 0$ as $n \to \infty$. We also assume that cdfs F, G have continuous densities f, g, respectively, and that the slope of the curve $FG^{-1}(t)$, that is, $f(G^{-1}(t))/g(G^{-1}(t))$, is bounded on any subinterval (a, b) of $(0, 1)$, $0 < a < b < 1$. Under the "binormal" model, for example, this condition is satisfied. In fact, if $\sigma > 1$, we can take the interval to be $[0, 1]$. (Of course if $\sigma < 1$, we can reverse the roles of F and G.) The first two theorems state the strong consistency and strong approximation properties for the ODC curve.

THEOREM 2.1. *Under the above conditions*,

$$\sup_{0 \leq t \leq 1} \left| F_m G_n^{-1}(t) - FG^{-1}(t) \right| \to 0 \quad a.s. \text{ as } n \to \infty.$$

PROOF. Consider the inequality

$$\begin{aligned}\sup_t \left| F_m G_n^{-1}(t) - FG^{-1}(t) \right| \leq{} & \sup_t \left| F_m G_n^{-1}(t) - F\left(G_n^{-1}(t)\right) \right| \\ & + \sup_t \left| FG_n^{-1}(t) - FG^{-1}(t) \right|.\end{aligned}$$

If we apply the Glivenko–Cantelli theorem for the first term on the RHS and the theorem of Dvoretzky, Kiefer and Wolfowitz (1956) and then the Borel–Cantelli lemma for the second term, the theorem is proved. □

THEOREM 2.2. *Under the above conditions, there exists a probability space on which one can define sequences of two independent versions of Brownian bridges* $\{B_1^{(n)}, B_2^{(n)}, 0 \leq t \leq 1\}$ *such that*

$$\sqrt{n}\left(F_m G_n^{-1}(t) - FG^{-1}(t)\right) = \sqrt{\lambda}\, B_1^{(n)}\left(FG^{-1}(t)\right) + \frac{f(G^{-1}(t))}{g\left(G^{-1}(t)\right)} B_2^{(n)}(t) + o\left(n^{-1/2}(\log n)^2\right) \quad a.s.$$

uniformly on $[a, b]$.

For notational simplicity, we will omit the superscript n on B_1 and B_2. This theorem follows from Theorem 4.4.1 in Csörgő and Révész (1981) and Theorem 3.2.4 in Csörgő (1983). The details are omitted here but can be found in Hsieh and Turnbull (1992). As an application of Theorem 2.2, consider the area under the empirical ODC curve,

$$M_{m,n} = \int_0^1 F_m G_n^{-1}(t)\, dt = \frac{1}{mn} \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} 1(X_i < Y_j),$$

which equals the Mann–Whitney statistic when there are no ties. Now M_{mn} is a strongly consistent estimator of $P(X < Y)$. In addition, applying Theorem 2.2 with the same conditions we obtain the following theorem.

THEOREM 2.3. *We have*

$$\sqrt{n}\left(M_{m,n} - P(X < Y)\right) = N(0, \sigma^2) + o_p(1),$$

in distribution as $n \to \infty$, *where* σ^2 *is defined as*

$$\begin{aligned} \sigma^2 &= \operatorname{var}\left[\sqrt{\lambda} \int_0^1 B_1\left(FG^{-1}(t)\right) dt + \int_0^1 \frac{f(G^{-1}(t))}{g\left(G^{-1}(t)\right)} B_2(t)\, dt\right] \\ &= \lambda \operatorname{var}\left[\int_0^1 B_1\left(FG^{-1}(t)\right) dt\right] + \operatorname{var}\left[\int_0^1 B_2\left(GF^{-1}(t)\right) dt\right] \qquad (2) \\ &= \lambda \|F \cdot G^{-1}\|^* + \|G \cdot F^{-1}\|^* \end{aligned}$$

and $\|h\|^* = \int_0^1 h^2\, dt - (\int_0^1 h\, dt)^2$.

It is worthwhile noting that the variance σ^2 in (2) is the same as the one obtained by considering the projection of $M_{m,n}$ viewed as a U-statistic [Serfling (1980), page 193], and of course reduces to the usual formula $(1/12)(1/m + 1/n)$ in the null case $F = G$.

3. The binormal model for grouped data. For the remainder of this article we consider the semiparametric "binormal" model as described in Section 1. We discuss first the situation where the data are grouped, which was the setting in the application (psychology) where the model was first used. We present a generalized least squares method, compare it with the well-known procedure of Dorfman and Alf (1969) and obtain asymptotic properties of the resulting estimators of the ODC curve. We also report the results of a simulation experiment to evaluate small sample properties.

3.1. *The generalized least squares method.* Let us fix integer k and let $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_k < 1$ be a given partition of the unit interval. Recall from (1) that the ODC curve is given by $FG^{-1}(t) = \Phi(\mu + \sigma\Phi^{-1}(t))$ for $0 \le t \le 1$, where unknowns μ and σ are to be estimated. Therefore we define

$$\beta_i = \Phi\big(\mu + \sigma\Phi^{-1}(\alpha_i)\big). \tag{3}$$

A natural estimator of β_i is

$$\hat{\beta}_i = F_m\big(G_n^{-1}(\alpha_i)\big), \qquad i = 1, \dots, k, \tag{4}$$

where F_m and G_n are the sample cdfs as in Sections 1 and 2. The asymptotic distribution of $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_k)$ can be derived from the covariance structure of Brownian bridges constructed in Theorem 2.2. This is stated in the following lemma.

LEMMA 3.1. *For fixed* $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_k < 1$, *under the "binormal" assumption, as* $n \to \infty$,

$$\begin{aligned} \sqrt{n}\,\big(\hat{\beta} - \beta\big) &\to_D N(0, \lambda\Sigma_1 + \Sigma_2) \quad \textit{and} \\ \sqrt{n}\,\big(\Phi^{-1}(\hat{\beta}) - \Phi^{-1}(\beta)\big) &\to_D N(0, \Sigma). \end{aligned} \tag{5}$$

Here $\Sigma = C[\lambda\Sigma_1 + \Sigma_2]C$ *with* $C^{-1} = \mathrm{diag}(\dots, \phi(\mu + \sigma\Phi^{-1}(\alpha_i)), \dots)$, Σ_1 *has* (i, j)*th entry equal* $[(\beta_i \wedge \beta_j) - \beta_i\beta_j]$, $\Sigma_2 = A\Sigma_0 A$ *with* $A = \mathrm{diag}(\dots,(\sigma\phi(\mu + \sigma\Phi^{-}(\alpha_i)))/(\phi(\Phi^{-1}(\alpha_i))), \dots)$ *and* Σ_0 *has* (i, j)*th entry* $[(\alpha_i \wedge \alpha_j) - \alpha_i\alpha_j]$. *Also* ϕ *denotes the standard normal density.*

We note that, since the $\{\alpha_i\}$ are chosen to be strictly increasing, so are the $\{\beta_i\}$. Therefore, the two covariance matrices Σ_1 and Σ_0 of the two finite dimensional distributions of Brownian bridges $B_1^{(n)}$ and $B_2^{(n)}$ at index points $\{\beta_i\}$ and $\{\alpha_i\}$, respectively, are positive definite. Hence the matrix Σ is nonsingular.

From (3) and (5), we have a linear regression setup:

$$\Phi^{-1}\big(\hat{\beta}_i\big) = \mu + \sigma\Phi^{-1}(\alpha_i) + \varepsilon_i, \qquad i = 1, \dots, k,$$

where the error vector, $\varepsilon' = (\varepsilon_1, \dots, \varepsilon_k)$, has the asymptotic covariance structure specified in (5) of Lemma 3.1. Since Σ depends on the unknown parameters μ and σ, an iterative procedure is needed to find estimates of μ and σ.

We proceed as follows. Since ε is $O_p(1/\sqrt{n})$, so is the ordinary least squares estimator

$$\begin{pmatrix} \hat{\mu}_0 \\ \hat{\sigma}_0 \end{pmatrix} = (M'M)^{-1} M' \Phi^{-1}(\hat{\beta}), \tag{6}$$

where M is the design matrix given by

$$M' = \begin{pmatrix} 1 & \cdots & 1 \\ \Phi^{-1}(\alpha_1) & \cdots & \Phi^{-1}(\alpha_k) \end{pmatrix} \tag{7}$$

and the vector $\Phi^{-1}(\hat{\beta})' = (\Phi^{-1}(\hat{\beta}_1), \dots, \Phi^{-1}(\hat{\beta}_k))$. We now substitute $\hat{\mu}_0$ and $\hat{\sigma}_0$ for μ and σ in the formula for Σ in (5), calling this estimator $\hat{\Sigma}$, say. Finally, the generalized least squares (GLS) estimator is derived as

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = (M'\hat{\Sigma}^{-1}M)^{-1} M'\hat{\Sigma}^{-1}\Phi^{-1}(\hat{\beta}). \tag{8}$$

We propose to use only this one-step estimator. The procedure could be iterated, but one step is usually adequate and often preferable in GLS estimation situations; for example, see the comments of Carroll and Ruppert [(1988), page 15]. Theorem 3.2 below shows that the asymptotic distribution of $(\hat{\mu}, \hat{\sigma})$ is the same as that if Σ were known.

THEOREM 3.2. *Under assumption of Lemma* 3.1, *as* $n \to \infty$,

$$\sqrt{n}\begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma} - \sigma \end{pmatrix} \to_D N\left(0, (M'\Sigma^{-1}M)^{-1}\right).$$

PROOF. Let

$$\hat{\theta}_{op} = (M'\Sigma^{-1}M)^{-1} M'\Sigma^{-1}\Phi^{-1}(\hat{\beta})$$

be the generalized weighted least squares estimator as if Σ were known. From Lemma 3.1, $\sqrt{n}(\hat{\theta}_{op} - \theta)$ is asymptotically distributed as $N(0, (M'\Sigma^{-1}M)^{-1})$. The ordinary least squares estimator, $\hat{\theta}_0 = (\hat{\mu}_0, \hat{\sigma}_0)$ say, defined in (6) is $\sqrt{n}$-consistent since ε is $O_p(1/\sqrt{n})$. So is the $\hat{\Sigma}^{-1}$ as the estimator of Σ^{-1}. Therefore, the GLS estimator, $\hat{\theta}' = (\hat{\mu}, \hat{\sigma})$ say, defined in (8) is equal to $\hat{\theta}_{op} + O_p(1/n)$. This completes the proof of this theorem. □

REMARK. The "binormal" model is an example of a two-sample transformation model, in which it is posited that there is an unknown transformation $H \in \mathscr{H}$, a transformation group, such that, if $X \sim F$ and $Y \sim G$, then $H(X) \sim N(0,1)$ and $H(Y) \sim N(\mu, \sigma^2)$. If we replace the normal distribution here by another parametric distribution, we can generate other semiparametric families. For example, use of the logistic leads to the proportional odds model; use of the Weibull leads to the proportional hazards model. The empirical ODC curve $F_m(G_n^{-1}(t))$ is a maximal invariant with respect to the group $\mathscr{H}$. It is reasonable to make inferences based on this maximal invariant. The GLS estimating procedure developed here is more convenient than other existing estimators proposed in the semiparametric literature. Estimation and testing in such semiparametric models have been studied by several authors; these include Bickel (1986), Clayton and Cuzick (1986), Doksum (1987) and recently Bickel and Ritov (1995).

3.2. *A simulation study and an adaptive procedure.* To investigate the small sample performance of the estimators $\hat{\mu}$ and $\hat{\sigma}$ from the GLS algorithm of Section 3.1, a small simulation experiment was conducted. In the algorithm only one step is used. Six binormal situations were simulated, in which $(\mu, \sigma) = (0, 1), (0, 2), (1, 1), (1, 2), (2, 1)$ and $(2, 2)$, respectively. The six corresponding true ODC curves are displayed in Figure 2. One hundred training data sets, each with $m = 100$ and $n = 100$, were generated. The first two columns of Table 1 show the estimated means and mean square errors (MSE's) of the estimators: the first using $k = 5$ and $(\alpha_1, \dots, \alpha_5) = (0.1, 0.2, 0.3, 0.4, 0.5)$; the second using $k = 8$ and $(\alpha_1, \dots, \alpha_8) = (0.1, \dots, 0.8)$. In some of the cases (e.g., $\mu = 0$, $\sigma = 2$), the estimators' performance is unsatisfactory as demonstrated by the large MSE's. On reflection, this fact can be explained by observing that a high proportion of the α_i-values fall on the flat part of the curve where the corresponding β-value is close to 0 or 1. The Φ^{-1} transformation will then clearly lead to unstable estimates with moderate sample sizes.

To remedy this situation, we propose the following adaptive method for selecting the $\{\alpha_i\}$ values so that they are concentrated on the steeper part of the ODC curve:

1. Fix a positive integer q.
2. Take $\tilde{\alpha}_1 = \min\{j/n \mid F_m G_n^{-1}(j/n) \geq q/m,\ j = 1, \dots, n\}$.
3. Set $\tilde{\alpha}_{i+1} = \min\{j/n \mid F_m G_n^{-1}(j/n) - F_m G_n^{-1}(\tilde{\alpha}_i) \geq q/m,\ j = 1, \dots, n\}$ for $i = 1, \dots, k(q)$, where $k(q)$ is the largest integer such that $\tilde{\alpha}_{k(q)} < 1$.

The algorithm used for constructing the GLS estimator of (μ, σ) is now applied using the $\{\tilde{\alpha}_i\}$. Simulation results with $q = 5$, 10, 12 are shown in the last three columns, respectively, of Table 1. Recall this implies that the ODC curve is being fitted to $k(q) \approx 100/q$ points, most of which are concentrated at the curved portion of the ODC curve. As expected the adaptive method provides better estimates in terms of bias and of MSE. For very steep ODC curves such as when $\mu = 2$ and $\sigma = 1$ or 2 (see Figure 2), it is clearly desirable to choose smaller values of q (i.e., larger k). On the other hand, we cannot use too small a value for q, since the normal approximation needed in the error structure will not be so accurate. This is a difficult problem.

3.3. *The procedure of Dorfman and Alf.* For estimating ROC curves from discrete or grouped response data, the most commonly used procedure is that proposed by Dorfman and Alf (1969). Here an individual's reading can take on one of a $k + 1$ number of categories, $R_1, \dots, R_{k+1}$, say. Such "ratings" data are described in Section 1. The approach postulates the existence of a latent random variable, W say, and unknown cut points, $-\infty = z_0 < z_1 < \cdots$

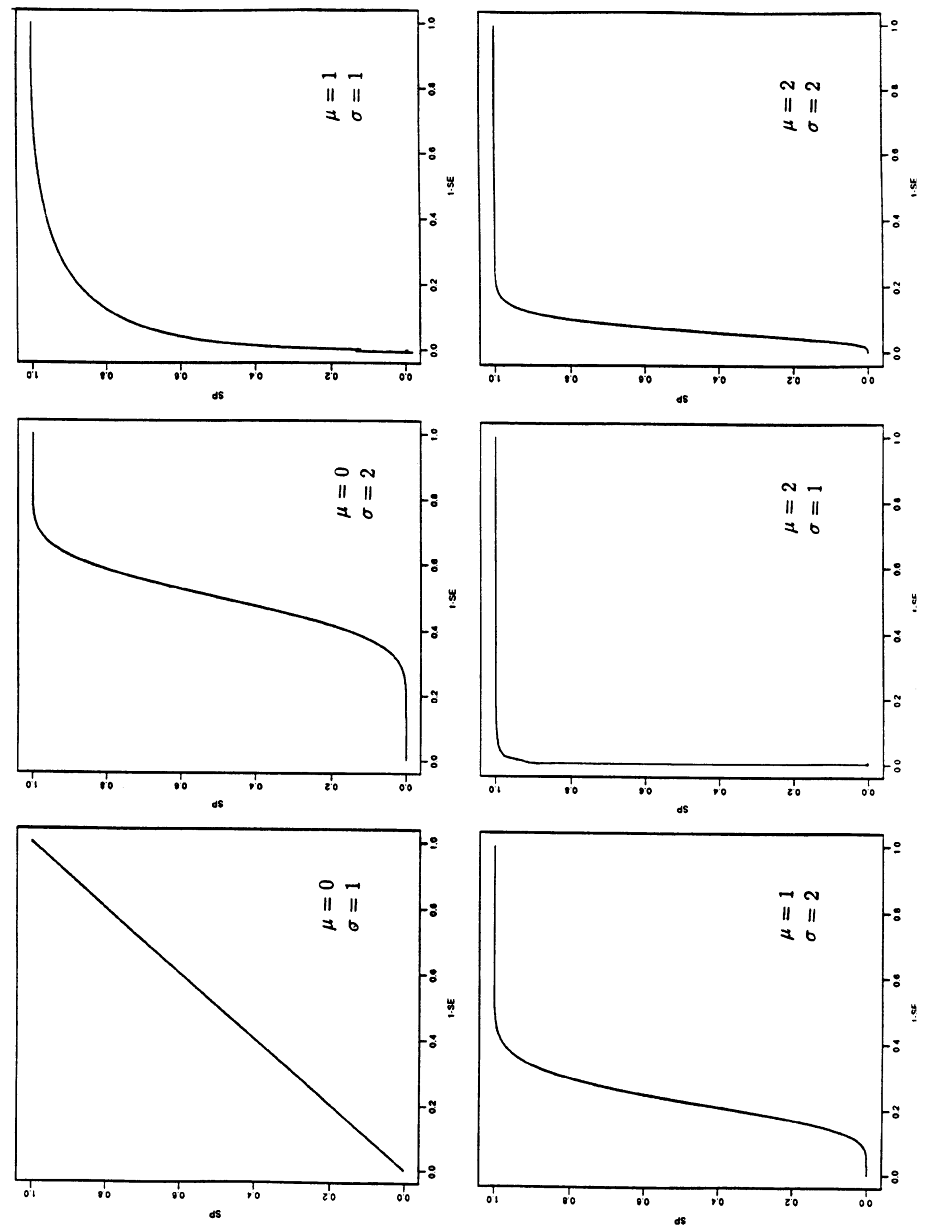

FIG. 2. *Six binormal ODC curves used as models in the simulation study of Section 3.2.*

TABLE 1

Simulation study results: estimates of μ and σ (with MSEs) for binormal model using GLS and adaptive GLS methods

	GLS Method				**Adaptive GLS Method**					
$\boldsymbol{\mu}, \boldsymbol{\sigma}$	$\boldsymbol{k = 5}$		$\boldsymbol{k = 8}$		$\boldsymbol{q = 5}$		$\boldsymbol{q = 10}$		$\boldsymbol{q = 12}$	
(0, 1)	−0.0434	0.9647	−0.0086	0.9903	0.0318	0.9883	0.0277	0.9863	0.0290	1.0030
	(0.0352	0.0549)	(0.0190	0.0285)	(0.0153	0.0175)	(0.0158	0.0231)	(0.0164	0.0245)
(0, 2)	1.2817	8.1000	−0.8562	6.1545	0.0372	1.9940	0.0425	2.0204	0.0429	2.0167
	(3.7549	71.4583)	(2.3054	32.7196)	(0.0516	0.0949)	(0.0558	0.1299)	(0.0542	0.1563)
(1, 1)	1.0157	0.9828	1.4038	1.4733	1.0336	0.9544	1.0269	0.9631	1.0183	0.9533
	(0.0304	0.0340)	(1.9523	3.8235)	(0.0226	0.0188)	(0.0322	0.0273)	(0.0296	0.0263)
(1, 2)	1.0830	2.2603	4.6257	7.6118	1.0419	1.9786	1.0635	2.0293	1.0776	2.0583
	(0.3045	3.2925)	(19.1173	45.2313)	(0.0693	0.1134)	(0.1097	0.2075)	(0.1223	0.2562)
(2, 1)	4.0997	2.8264	7.2842	6.6518	1.9965	0.9131	2.0312	0.9384	2.0715	0.9578
	(25.7160	21.8936)	(36.8637	39.8626)	(0.0531	0.0343)	(0.2360	0.0872)	(0.4001	0.1241)
(2, 2)	3.9831	4.2939	8.4342	8.7022	2.0338	1.9356	1.9970	1.9216	2.0242	1.9364
	(25.3856	35.4289)	(45.1738	46.6128)	(0.1319	0.1424)	(0.1828	0.1912)	(0.2353	0.2274)

Note: Entries show mean estimates of μ, σ (with MSEs shown in parentheses) for six binormal models with (μ, σ) as given in the left-hand column and are based on 100 replications. The situations simulated all use training sets of size $m = n = 100$. For the GLS method, $k = 5$ corresponds to $(\alpha_1, \dots, \alpha_5) = (0.1, \dots, 0.5)$; $k = 8$ corresponds to $(\alpha_1, \dots, \alpha_8) = (0.1, \dots, 0.8)$.

$< z_k < z_{k+1} = \infty$, such that if $z_{i-1} < W \le z_i$, then the response category for the corresponding individual is R_i, $1 \le i \le k + 1$. Let P_{i1} denote the probability of obtaining response category R_i from a healthy individual and similarly P_{i2} for a diseased subject. The Dorfman and Alf model stipulates that

$$\begin{aligned} P_{i1} &= \Phi(z_i) - \Phi(z_{i-1}) \quad \text{and} \\ P_{i2} &= \Phi\left(\frac{z_i - \mu}{\sigma}\right) - \Phi\left(\frac{z_{i-1} - \mu}{\sigma}\right), \qquad i = 1, \dots, k + 1. \end{aligned}$$

It can be seen that this definition corresponds to our previous description of the "binormal" model if we equate the latent variable $W = H(X)$ for a healthy subject and $W = H(Y)$ for a diseased subject, where H is the unknown transformation as defined in Sections 1 and 3. The $\{z_i\}$ can be thought of as $\{H(c_i)\}$ for cut points, $\{c_i\}$ say, on the original latent continuous response variable scale. (In contrast, the grouped data in Section 3.1 are obtained by discretizing with respect to the empirical quantiles $\{G_n^{-1}(\alpha_i)\}$.)

The log likelihood function is given by

$$(9) \qquad \sum_{l=1}^{2} \sum_{i=1}^{k+1} \gamma_{il} \log P_{il},$$

where γ_{i1}, γ_{i2} are the observed numbers of responses in category R_i from the healthy and diseased group, respectively $(1 \le i \le k + 1)$.

The computation of the maximum likelihood estimate (MLE) of (μ, σ) based on (9) requires the solution of a system of $k + 2$ nonlinear equations in $k + 2$ unknowns. A χ^2 statistic can be used to test the goodness-of-fit to the model. A computer program to carry out this analysis is available in Swets and Pickett [(1982), Appendix D]. However, the computation can be difficult when k, the number of nuisance parameters $z_1, z_2, \dots, z_k$, is large, and the iterative procedure used in the program can fail to converge. We now explore alternatives to full maximum likelihood to alleviate this computational problem so that, as with the approach in Section 3.1, only the two unknowns $(\mu, \sigma)' = \theta$, say, are explicitly involved.

We proceed by noting that the Dorfman and Alf model can be viewed approximately as a measurement error regression model. First we denote

$$(10) \qquad \alpha_i = \Phi\left(\frac{z_i - \mu}{\sigma}\right) = P_{12} + P_{22} + \cdots + P_{i2} \quad \text{for } i = 1, \dots, k,$$

which implies that

$$(11) \quad \Phi(z_i) = \Phi\left(\mu + \sigma\Phi^{-1}(\alpha_i)\right) = P_{11} + P_{21} + \cdots + P_{i1} \qquad \text{for } i = 1, \dots, k.$$

The quantities in (10) and (11) are naturally estimated by the sample proportions $(1/n)\Sigma_{l=1}^{i} \gamma_{l2}$ and $(1/m)\Sigma_{l=1}^{i} \gamma_{l1}$, respectively.

Hence a regression model with measurement error can be written as

$$(12) \qquad \Phi^{-1}\left(\frac{1}{m}\sum_{l=1}^{i}\gamma_{l1}\right) = \sigma\Phi^{-1}(\alpha_i) + \mu + \varepsilon_{1i},$$

$$(13) \qquad \Phi^{-1}\left(\frac{1}{n}\sum_{l=1}^{i}\gamma_{l2}\right) = \Phi^{-1}(\alpha_i) + \varepsilon_{2i} \qquad \text{for } i = 1, \dots, k,$$

where $\Phi^{-1}(\alpha_i) = z_i$ and vectors $\varepsilon_1' = (\varepsilon_{11}, \dots, \varepsilon_{1k})$ and $\varepsilon_2' = (\varepsilon_{21}, \dots, \varepsilon_{2k})$ are independent and normally distributed as $N(0, \Sigma_1^*)$ and $N(0, \Sigma_2^*)$, respectively, where $m\Sigma_1^* = C\Sigma_1 C$ and $n\sigma^2\Sigma_2^* = C\Sigma_2 C$ as defined in Lemma 3.1. We define $\hat{\Sigma}$ and $\hat{M}$ as in Lemma 3.1 and (7), respectively, but with sample proportions $\hat{\alpha} = ((1/n)\Sigma_{l=1}^{i}\gamma_{l2})$ and $\hat{\beta}_i = ((1/m)\Sigma_{l=1}^{i}\gamma_{l1})$ replacing α_i and β_i. Then based on (12) and (13), similar to that in Section 3.1, a GLS estimator $(\hat{\mu}^*, \hat{\sigma}^*)$ can be constructed as

$$(14) \qquad \begin{pmatrix} \hat{\mu}^* \\ \hat{\sigma}^* \end{pmatrix} = \left(\hat{M}'\hat{\Sigma}^{-1}\hat{M}\right)^{-1}\hat{M}'\hat{\Sigma}^{-1}\Phi^{-1}(\hat{\beta}).$$

The next theorem shows that $(\hat{\mu}^*, \hat{\sigma}^*)$ has the same asymptotic distribution as $(\hat{\mu}, \hat{\sigma})$ defined in (8).

THEOREM 3.3. *The estimators $\hat{\mu}^*$ and $\hat{\sigma}^*$ as given in* (14) *have asymptotic distribution given by*

$$\sqrt{n}\begin{pmatrix} \hat{\mu}^* - \mu \\ \hat{\sigma}^* - \sigma \end{pmatrix} \to_D N\left(0, \left(M'\Sigma^{-1}M\right)^{-1}\right) \quad \textit{as } n \to \infty.$$

Proof. The measurement error model can be written as

$$\Phi^{-1}(\hat{\beta}) = M\theta + \varepsilon_1 \quad \text{and}$$
$$\Phi^{-1}(\hat{\alpha}) = \Phi^{-1}(\alpha) + \varepsilon_2.$$

Combining the above, we have the regression equation

$$\Phi^{-1}(\hat{\beta}) = \hat{M}\theta + (\varepsilon_1 - \sigma\varepsilon_2). \tag{15}$$

Here $\sqrt{n}(\varepsilon_1 - \sigma\varepsilon_2)$ is asymptotically distributed as multivariate normal $N(0, \Sigma)$. Let $\hat{\Sigma}$ be the estimator of Σ obtained by plugging in $\hat{\alpha}$ for α and replacing σ by $\hat{\sigma}_0$, the second component of $\hat{\theta}_0$, the ordinary least squares estimator obtained from (15) which ignores the error structure. Thus $\hat{\Sigma}$ is $\sqrt{n}$-consistent. Multiplying both sides of (15) by $\hat{M}'\hat{\Sigma}^{-1}$, after some calculations, we have

$$\begin{aligned}\hat{M}'\hat{\Sigma}^{-1}\Phi^{-1}(\hat{\beta}) &= \hat{M}'\hat{\Sigma}^{-1}\hat{M}\theta + \hat{M}'\hat{\Sigma}^{-1}(\varepsilon_1 - \sigma\varepsilon_2)\\ &= \hat{M}'\hat{\Sigma}^{-1}\hat{M}\theta + M'\Sigma^{-1}(\varepsilon_1 - \sigma\varepsilon_2) + O_p\left(\frac{1}{n}\right).\end{aligned}$$

This completes the proof. □

An implication of the equality of the two asymptotic distributions in Theorems 3.2 and 3.3 is that asymptotically all the information about $\theta' = (\mu, \sigma)$ is contained in the empirical ODC (or ROC) curve. Now one could also form an approximate normal likelihood based on equations (12) and (13), and obtain MLE's. It turns out that these have the same asymptotic distribution as the Dorfman–Alf MLE's based on (9) and this again is the same as that in Theorem 3.3. This is stated in the next theorem. Thus all four estimators—(8), (14) and these two MLE's—are equivalent in terms of asymptotic efficiency.

Theorem 3.4. *Under the "binormal" model, the MLE estimators of* $\theta' = (\mu, \sigma)$ *based on* (9) *and the approximate MLE estimators based on the regression setup* (12) *and* (13) *have the same asymptotic distribution as the GLS estimators given in Theorem* 3.3.

A rigorous proof of this theorem can be obtained by taking corresponding derivatives of the normal likelihood of the regression setup (12) and (13), and dropping the several insignificant terms. The we arrive at the same system of score equations as derived from (9). The theorem is proved by taking a profile likelihood approach. A detailed proof can be found in Hsieh and Turnbull (1992).

In case of two samples of continuous data, based on Theorem 3.4, a heuristic argument of the asymptotic efficiency of the GLS estimator is given as follows. By choosing suitable k depending on the sample sizes m and n, the likelihood of (9) or of the regression setup (12) and (13) can be shown to be asymptotically sufficient for θ and the nuisance parameter F in a sense of not losing information. Specifically, the Fisher information of θ, I_n say, tends to the Fisher information I_0 of θ in the original problem. Let the MLE of θ based on (9), or on (12) and (13), be derived from the profile likelihood approach. The projection interpretation of the approach assures that I_0 is the semiparametric Fisher information bound in the same sense as that given in Begun, Hall, Huang and Wellner (1983). Therefore, by Theorem 3.4, the GLS estimator is asymptotically efficient.

To assess goodness-of-fit of the "binormal" model, the test statistic

$$S_n = \left(\Phi^{-1}(\hat{\beta}) - \hat{M}\hat{\theta}\right)' \hat{\Sigma}^{-1}\left(\Phi^{-1}(\hat{\beta}) - \hat{M}\hat{\theta}\right)$$

can be constructed. Here $\hat{\theta} = (\hat{\mu}, \hat{\sigma})'$ is any of the asymptotically equivalent estimators described in this section. Under the "binormal" model assumption, this statistic is distributed as χ^2_{k-2} asymptotically.

REMARK (Estimation of H). We obtain an estimate of the underlying transformation H from the MLE of the $\{z_i\}$ obtained either from (9) or from (12) and (13). Suppose cutpoints $\{c_i\}$ have been used to group the responses on the original continuous scale. Since $\hat{z}_i$ is an estimate of $H(c_i)$, an estimate of H can be obtained by fitting a smooth monotone function to the points $\{(c_i, \hat{z}_i), 1 \le i \le k\}$ using an appropriate smoothing technique.

4. The binormal model for continuous data. In Section 3, procedures for estimating the ODC and ROC curves under the "binormal" model were proposed. They involved grouping or discretizing the response data in some way. In this section, we propose a minimum distance estimator (MDE) of the ODC curve under the "binomial" model which does not require that the continuous data be grouped.

Minimum distance estimation has been studied extensively beginning with the work of Wolfowitz (1957). Millar (1984) presented a general abstract approach. For our problem, the MDE is constructed by finding the ODC curve based on the "binormal" model, that is, of the form $\Phi(\mu + \sigma\Phi^{-1}(t))$, that fits most closely the empirical ODC curve using an L_2 norm criterion. The MDE estimates are defined to be the minimizing values of μ and σ. More precisely, for $\theta = (\mu, \sigma)'$, we define

$$\xi_{mn}(\theta) = \left[F_m G_n^{-1}(t) - \Phi(\mu + \sigma\Phi^{-1}(t))\right]$$

and the L_2-distance measure as

$$\|\xi_{mn}(\theta)\| = \int_0^1 [\xi_{mn}(\theta)]^2\, dt. \tag{16}$$

The MDE, $\hat{\theta}_{mn} = (\hat{\mu}, \hat{\sigma})'$, is defined by

$$\left\|\xi_{mn}\left(\hat{\theta}_{mn}\right)\right\| = \inf_{\theta} \|\xi_{mn}(\theta)\|. \tag{17}$$

As before, for our asymptotic theory we suppose $n/m \to \lambda\ (> 0)$ as $n \to \infty$. From here on, the dependence on λ is suppressed and denote $\xi_n = \xi_{mn}$ and

$\hat{\theta}_n = \hat{\theta}_{mn}$. We let Θ be the set $\{(\mu, \sigma)' \mid \mu \in \Re$ and $\sigma > 1\}$ and suppose $\theta_0 = (\mu_0, \sigma_0)' \in \Theta$ is the true unknown value of θ. The restriction that $\sigma > 1$ is not unreasonable if one thinks of the healthy response as "noise" and the diseased response as "noise plus signal." [However, we can avoid this restriction if we modify the distance criterion (16) above so that the integral is over a closed interval excluding 0 and 1.] Finally, let $\mathscr{B}_2$ be the separable Hilbert space $L_2(\nu)$, where ν is the uniform measure on [0, 1].

To assert the asymptotic normality of the MDE $\hat{\theta}_n$, we will apply Theorem 3.6 of Millar [(1984), Section III]. We need to check the three conditions of the theorem. The "identifiability" condition is satisfied because $\xi_n(\theta) - \xi_n(\theta_0) = \Phi(\mu + \sigma\Phi^{-1}(t)) - \Phi(\mu_0 + \sigma_0^{-1}\Phi(t))$ is nonrandom and does not depend on n. The "convergence" condition holds because, from Theorem 2.2, the $\mathscr{B}_2$-valued random process $\sqrt{n}\,\xi_n(\theta_0)$ converges in L_2 to a process $W(\theta_0)$. Here $W(\theta)$ is the combination of Brownian bridge processes as given in Theorem 2.2 under the "binormal" assumption, that is,

$$W(t;\theta) = \sqrt{\lambda}B_1\big(\Phi(\mu + \sigma\Phi^{-1}(t))\big) + \frac{\sigma\phi\big(\mu + \sigma\Phi^{-1}(t)\big)}{\phi\big(\Phi^{-1}(t)\big)}B_2(t).$$

The third condition of differentiability follows because there is a continuous linear operator T $(= T_{\theta_0})$ from Θ to $\mathscr{B}_2$, such that

$$\begin{aligned}\xi_n(\theta) &= \xi_n(\theta_0) + T(\theta - \theta_0) + o_p(\theta - \theta_0)\\ &= \xi_n(\theta_0) + \eta_1(\theta_0)(\mu - \mu_0) + \eta_2(\theta_0)(\sigma - \sigma_0) + o_p(\theta - \theta_0),\end{aligned}$$

where $\eta_1(\theta_0) = \phi(\mu_0 + \sigma_0\Phi^{-1}(t))$ and $\eta_2(\theta_0) = \Phi^{-1}(t)\phi(\mu_0 + \sigma_0\Phi^{-1}(t))$. Both are partial derivatives of $\Phi(\mu + \sigma\Phi^{-1}(t))$ with respect to μ and σ, respectively, and are evaluated at θ_0. Since $\eta_1(\theta_0)$ and $\eta_2(\theta_0)$ are linearly independent, the operator T is nonsingular.

Let $\mathscr{B}_\eta$ denote the linear space spanned by $\eta_1(\theta_0)$ and $\eta_2(\theta_0)$ and let π denote the projection mapping from $\mathscr{B}_2$ onto $\mathscr{B}_\eta$. We now can apply Millar's Theorem 3.6 to obtain the following asymptotic properties of $\hat{\theta}_n$.

THEOREM 4.1. *With probability approaching* 1 *as* $n \to \infty$, $\hat{\theta}_n$ *exists and is unique. Moreover*

$$\begin{aligned}\xi_n\big(\hat{\theta}_n\big) &= (1 - \pi)\circ\xi_n(\theta_0) + o_p(n^{-1/2}),\\ \hat{\theta}_n - \theta_0 &= -T^{-1}\circ\pi * \xi_n(\theta_0) + o_p(n^{-1/2}).\end{aligned}$$

In addition, we have the following weak convergence results:

$$\begin{aligned}\sqrt{n}\Big(\xi_n\big(\hat{\theta}_n\big) - \xi_n(\theta_0)\Big) &\Rightarrow \pi\circ W \quad \textit{in } \mathscr{B}_2,\\ \sqrt{n}\Big(\hat{\theta}_n - \theta_0\Big) &\Rightarrow -T^{-1}\circ\pi\circ W \quad \textit{in } \Re^2.\end{aligned}$$

With the results of this theorem, we can obtain an explicit expression for the asymptotic covariance matrix, $n^{-1}\Lambda$ say, for $\hat{\theta}_n$ as follows. First define the inner product in $\mathscr{B}_2$ by $\langle h, k\rangle = \int_0^1 h(t)\cdot k(t)\,dt$. Let $R(s,t) = E(W(s)\cdot W(t))$

be the covariance function of W. After some calculations, an explicit expression of $R(s,t)$ is found to be

$$R(s,t) = \frac{\lambda}{\sigma^2}\left\| \phi\left(\frac{\Phi^{-1}(t) - \mu}{\sigma}\right)\right\|^* + \| \phi(\mu + \sigma\Phi^{-1}(t))\|^*,$$

where $\|h\|^* = \int_0^1 h^2(t)\,dt - (\int_0^1 h\,dt)^2$, as in Theorem 2.3. Finally define 2×2 matrices A and C by $C_{ij} = \langle \eta_i, \eta_j \rangle$ and

$$\begin{aligned} A_{ij} &= E\big(\langle \eta_i, W\rangle \cdot \langle \eta_j, W\rangle\big) \\ &= \int_0^1\int_0^1 \eta_i(s)\cdot R(s,t)\cdot \eta_j(t)\,ds\,dt. \end{aligned}$$

Thus the asymptotic covariance matrix of $\hat{\theta}_n = (\hat{\mu}_n, \hat{\sigma}_n)$ is given by $n^{-1}\Lambda$ with $\Lambda = C^{-1}AC^{-1}$.

As a by-product, the minimum distance approach provides a natural statistic for testing the "binormal" assumption, namely,

$$\left\| \xi_n\big(\hat{\theta}_n\big)\right\| = \inf_{\theta\in\Theta} \| \xi_n(\theta)\|.$$

The following corollary, which follows from Theorem 4.1, gives the asymptotic distribution of this test statistic under the "binormal" assumption.

COROLLARY 4.2. $n\|\xi_n(\hat{\theta}_n)\| \Rightarrow \int_0^1[(1-\pi)\circ W(t)]^2\,dt + o_p(1)$.

REMARK 1. In the definition of the distance criterion (16), we could introduce a weight function, such as the inverse of the variance function of processes $\xi_{mn}(\theta)$. Corresponding asymptotic results can be derived. Also distance criteria based on a different process, such as $\xi^*_{mn}(\theta) = [\Phi^{-1}(F_m G_n^{-1}(t)) - (\mu + \sigma\Phi^{-1}(t))]$, could be used. These ideas are the subject for future study, as are the computational problems for finding the MDE.

REMARK 2 (The LAM property of the MDE). Roughly, the locally asymptotic minimax (LAM) property of an estimator asserts that its performance does not deteriorate when the actual distributions of the data depart slightly from those specified by the model—here the "binormal" model. It can be proved that the MDE given by (17) has this property. This is done by showing that the problem can be fit into the general abstract framework of Millar [(1984), Sections 3 and 5]. Details are given in Hsieh and Turnbull (1992).

REFERENCES

BAMBER, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psych.* **12** 387–415.

BEGUN, J. M., HALL, W. J., HUANG, W. M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

BICKEL, P. J. (1986). Efficient testing in a class of transformation models. In *Proceedings of the 45th Session of the International Statistical Institute* 23.3-63–23.3-81. ISI, Amsterdam.

BICKEL, P. J. and RITOV, Y. (1995). Local asymptotic normality of ranks and covariates in transformation models. In *Festschrift for L. LeCam* (D. Pollard and G. Yang, eds.). Springer, New York. To appear.

BROWNIE, C., HABICHT, J.-P. and COGILL, B. (1986). Comparing indicators of health or nutritional status. *American Journal of Epidemiology* **124** 1031–1044.

CARROLL, R. J. and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.

CLAYTON, D. and CUZICK, J. (1986). The semiparametric Pareto model for regression analysis of survival times. In *Proceedings of the 45th Session of the International Statistical Institute* 23.3-1–23.3-18. ISI, Amsterdam.

CSÖRGŐ, M. (1983). *Quantile Processes with Statistical Applications*. SIAM, Philadelphia.

CSÖRGŐ, M. and RÉVÉSZ, P. (1981). *Strong Approximations in Probability and Statistics*. Academic Press, New York.

DOKSUM, K. A. (1987). An extension of partial likelihood method for proportional hazard models to general transformation model. *Ann. Statist.* **15** 325–345.

DORFMAN, D. D. and ALF, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence interval—rating method data. *J. Math Psych.* **6** 487–496.

DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution and of the classical multinomial estimator. *Ann. Math. Statist.* **27** 642–669.

GODDARD, M. J. and HINBERG, I. (1990). Receiver operator characteristic (ROC) curves and non-normal data: An empirical study. *Statistics in Medicine* **9** 325–337.

GREY, D. R. and MORGAN, B. J. T. (1972). Some aspects of ROC curve-fitting: normal and logistic models. *J. Math. Psych.* **9** 128–139.

HANLEY, J. A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical Decision Making* **8** 197–203.

HANLEY, J. A. and MCNEIL, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36.

HILDEN, J. (1991). The area under the ROC curve and its competitiors. *Medical Decision Making* **11** 95–101.

HSIEH, F. S. (1991). Performance of diagnostic tests in a nonparametric setting. Ph.D. dissertation, Cornell Univ.

HSIEH, F. S. and TURNBULL, B. W. (1992). Non- and semi-parametric estimation of the receiver operating characteristic curve. Technical Report 1026, School of Operations Research, Cornell Univ.

MILLAR, P. W. (1984). A general approach to the optimality of minimum distance estimators. *Trans. Amer. Math. Soc.* **286** 377–418.

OGILVIE, J. C. and CREELMAN, C. D. (1968). Maximum likelihood estimation of ROC curve parameters. *J. Math. Psych.* **5** 377–391.

POLLARD, D. (1980). The minimum distance method of testing. *Metrika* **27** 43–70.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Application to Statistics*. Wiley, New York.

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

WOLFOWITZ, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28** 75–88.

DEPARTMENT OF MATHEMATICS
NATIONAL TAIWAN UNIVERSITY
NO. 1., SEC. 4, ROOSEVELT RD.
TAIPEI
TAIWAN

SCHOOL OF OPERATIONS RESEARCH AND
INDUSTRIAL ENGINEERING
227 RHODES HALL
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-3801