SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NY 14853-3801

TECHNICAL REPORT NO. 1026

September 1992

NON- AND SEMI-PARAMETRIC ESTIMATION OF THE RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

by

Fushing Hsieh[1]
and Bruce Turnbull[2]

[1]National Tsing Hua University, Institute of Statistics, Hsinchu, Taiwan 30043. Research supported in part by a fellowship from the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell University.

[2]Research supported in part by Grant GM 28364 from the U.S. National Institutes of Health.

Non- and Semi-parametric Estimation of the Receiver Operating Characteristic (ROC) Curve

Fushing Hsieh
Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan 30043,

Bruce W. Turnbull
School of Operations Research and Industrial Engineering, 227 E&TC Building, Cornell University, Ithaca, NY 14853-3801

September 6, 1992

Abstract

The receiver operating characteristic (ROC) curve describes the performance of a diagnostic test used to discriminate between healthy and diseased individuals based on the variable measured on a continuous scale. The data consist of a training set of m responses $X_1, \ldots, X_m$ from healthy individuals and n responses $Y_1, \ldots, Y_n$ 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 $0 \leq t \leq 1$ or, equivalently, the ordinal dominance curve (ODC) given by $F(G^{-1}(t))$. We first consider nonparametric estimators based on empirical distribution functions and obtain consistency and asymptotic normality results. A common model, termed the "binormal" model, is a semi-parametric one in which it is assumed that the distributions F and G are normal after some arbitrary transformation of the measurement scale. We discuss

the well known estimation algorithm of Dorfman and Alf (1969) which is based on grouping the data, and compare it with an alternative procedure based on an iterative generalized least squares approach. Asymptotic results are obtained; small sample properties are examined via a simulation study. Finally, we develop ROC curve estimators which do not require grouping the data. Two methods are proposed: one based on maximizing the marginal likelihood of the ranks; the other based on a minimum distance criterion. The latter is shown to have a locally asymptotic minimax (LAM) property.

1 Introduction

A diagnostic test giving a measurement on a continuous scale is used to classify patients into either the "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, density g.

The sensitivity of the test is defined as $SE(c) = 1 - G(c)$, which is the probability of correctly classifying a diseased individual when cutoff point c is used. Similarly we define the test's specificity $SP(c) = F(c)$ as the probability of correctly classifying a healthy patient. Clearly these are the complements of the familiar Type I and Type II errors. The receiver operating characteristic (ROC) curve is defined as a plot of the "true positive fraction ", $SE(c)$, on the vertical axis versus the "false positive fraction", $1 - SP(c)$, on the horizontal axis as c varies from $+\infty$ to $-\infty$. Equiva-

lently, it can be viewed as a plot of $ROC(t) = 1 - G(F^{-1}(1-t))$ versus t, $(0 \leq t \leq 1)$. By reversing the axes we obtain the Ordinal Dominance Curve (ODC), $F(G^{-1}(t))$ $(0 \leq t \leq 1)$, which is a plot of $SP(c)$ versus $1 - SE(c)$, or equivalently $F(c)$ versus $G(c)$ for $-\infty \leq c \leq \infty$. Typical curves are illustrated in Figure 1. An desireable diagnostic test has an ODC curve that rises rapidly and then levels out as in the lower graphs of Figure 1. (In fact this definition of the ODC curve is slightly different from that of Bamber (1975) who used $G(F^{-1}(t))$ instead. An advantage of our definition is that the area under the ODC curve is the same as that under the ROC curve, namely $P[X < Y]$, instead of its complement.)

[Figure 1 about here.]

It is well-known that the ROC curve has the following convenient properties:

1. It is invariant under monotone increasing transformations of the measurement scale;

2. If $F(c) \geq G(c)$ for all c, i.e. X is stochastically smaller than Y, then $ROC(t) \geq t$ and the curve lies above the diagonal;

3. If the densities f and g have a monotone likelihood ratio then $ROC(t)$ is concave.

4. The area under the ROC 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. The first developments were in the field of signal detection theory by Peterson, Birdsall and Fox (1954). These were quickly carried over to the field of psychology where Swets (1973) reported that there had been extensive application in perceptual and cognitive studies. Swets (1988) describes applications in such areas as weather forecasting, polygraph lie detection, information retrieval and aptitude testing. Other fields of application have included: radiology and medical imaging (Hanley and McNeil 1982, 1983, Metz 1986, Rockette, Obuchowski and Gur 1990), epidemiology and nutrition (Erdreich and Lee 1981, Brownie, Habicht and Cogill 1986), clinical chemistry (Delong et al. 1988, Goddard and Hinberg 1990) and general medical decision making (Centor 1991). MEDLINE now lists the ROC curve as a separate subject heading.

The ROC curve is important because various measures of performance, or accuracy indices, for a given diagnostic test are based on the curve. For example, Greenhouse and Mantel (1950) and Linnett (1987 Stat. in Med.) use $ROC(t_0)$ for fixed specificity $1-t_0$. Many have used the area under the ROC curve, i.e. $\int_0^1 ROC(t)dt$ as the measure of efficacy of the test. These authors include Swets and Pickett (1982), Hanley and McNeil (1982,1983), Brownie et al. (1986), McClish (1987), DeLong et al. (1988). For a discussion, see Hilden (1991). Since the range of cutoff values c where the specificity is high is more important, Thompson and Zucchini (1989) advocated the use of the partial area $\int_0^P ROC(t)dt$ for given $0 < P < 1$ as a measure of performance, whereas Wieand et al. (1989) proposed a weighted area $\int_0^1 ROC(t)dW(t)$. See also McClish (1989). Measures of performance are important because

they enable several competing diagnostic tests to be compared. In fact many of the papers cited above are concerned with the problem of testing the difference between performance measures of two diagnostic instruments – see, for example, Wieand et al. (1989). However our concern here will be estimating a single ROC curve.

We suppose that a training data set, $X_1, X_2, \cdots, X_m$, of readings from the healthy population is available as is a set, $Y_1, Y_2, \cdots, Y_n$, from the diseased population. Estimation of the ROC curve can be undertaken using nonparametric, parametric or "semi-parametric" methods. In the next section we consider the situation when no parametric assumptions are made concerning the form of F and G. Properties of an empirical estimate are developed. Little work appears to have been done on the empirical ROC curve itself. However the area under the curve can be estimated by 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 et al. (1986) and Goddard and Hinberg (1990). However, a large proportion of the literature takes what we term a "semi-parametric" approach. This approach is based on the so-called "binormal" assumption; that is the existence of some unspecified arbitrary transformation H, say, of the measurement scale that simultaneously converts the F and G distributions to normal ones, $N(\mu_H, \sigma_H^2)$ and $N(\mu_D, \sigma_D^2)$ say. In this

case the ROC curve becomes linear, with slope $b = \sigma_H/\sigma_D$ and intercept $\delta = (\mu_H - \mu_D)/\sigma_D$, when a normal probability scale is used for both axes, i.e. $\Phi^{-1}(1 - G(c)) = \delta + b\Phi^{-1}(1 - F(c))$. (Here Φ denotes the standard normal *cdf.*) Thus fitting a straight line to an empirical ROC curve plotted using normal probability scales yields a graphical test for the goodness of fit of the binormal assumption and graphical estimates of δ and b, assuming the fit is adequate (Swets and Pickett 1982, p.30, Brownie et al. 1986, Figure 2). The binormal assumption was 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 – for example, with a five point rating scale, the values 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 cell boundaries plus two (b and δ). 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. The important parameters to be estimated are b and δ, since these alone determine the ROC curve under the assumption. Hanley (1988) discusses the justifications and applicability of the binormal assunption for ratings data. Other distributional assumptions can be used instead of the normal – Ogilvie and Creelman (1968) use 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 fact in the remaining sections, for notational simplicity, we will work

with the ODC curve, $FG^{-1}(t), 0 < t < 1$, which is a plot of $F(c)$ against $G(c)$ for $-\infty < c < \infty$. Of course properties of estimators of this curve will carry over to corresponding estimators of the ROC curve. In the next section we develop empirical estimators of these curves and derive a uniform strong consistency property and some weak convergence results for them. Application of these results to the nonparametric estimation of $P(X < Y)$ is discussed. In Section 3, we consider the semi-parametric model under the binormal assumption. Using the asymptotic theory of empirical processes, we examine the asymptotic properties of the MLE based on the Dorfman and Alf (1969) procedure for grouped data and then show them asymptotically equivalent to estimates based on a generalized least square approach.

Without grouping the data, we consider two procedures for semi-parametric estimation in Section 4. The first uses a maximum likelihood estimate based on the marginal likelihood of a maximum invariant statistics. The second one employs a minimum distance approach which we will recommend. The optimality of this minimum distance estimate (MDE) in a sense of locally asymptotic minimaxity (LAM) is proved.

2 Empirical ODC curve and $P(X < Y)$ estimation

We denote the empirical cdf's based on the X- and Y-training data sets by $F_m(x)$ and $G_n(y)$, respectively. Also we define $G_n^{-1}(t)$ by $\inf\{y : G_n(y) \geq t\}$. Now we can define the empirical ODC curve by:

$$F_m G_n^{-1}(t), \quad 0 < t < 1. \tag{1}$$

The strong consistency of curve $F_m G_n^{-1}(t)$ is stated in the Theorem 2.1 below. The corresponding weak convergence result follows in Theorem 2.2. We always will assume the sample size ratio $\frac{m}{n} = \lambda_n \longrightarrow \lambda\ (> 0)$ as $n \longrightarrow \infty$.

Theorem 2.1 Let f and g be continuous and slope of the curve $FG^{-1}(t)$, *i.e.* $\frac{f(G^{-1}(t))}{g(G^{-1}(t))}$, be finite for all $t \in (0,1)$. Then

$$\sup_{0 \le t \le 1} \mid F_m G_n^{-1}(t) - FG^{-1}(t) \mid \xrightarrow{a.s} 0 \quad as\ n \to \infty.$$

The proof is given in the Appendix.

For the weak convergence of the empirical ODC curve, we consider the following difference:

$$\begin{aligned} & F_m G_n^{-1}(t) - FG^{-1}(t) \\ = \ & (F_m(G_n^{-1}(t)) - F(G_n^{-1}(t)) + (F(G_n^{-1}(t)) - FG^{-1}(t)) \end{aligned} \tag{2}$$

The uniform $\sqrt{n}$ - consistency of the first term of (2) can be seen from Komlós *et al.* (1975) and Shorack and Wellner (1989 page 491). The second term can be rewritten as below, using a Taylor expansion,

$$\begin{aligned} F(G_n^{-1}(t)) - FG^{-1}(t) &= FG^{-1}(U_n(t)) - FG^{-1}(t) \\ &\simeq \frac{f(G^{-1}(t))}{g(G^{-1}(t))}(U_n(t) - t) \end{aligned} \tag{3}$$

for $t \in (0,1)$, where $U_n(t)$ is the empirical process induced by uniform random variables $G(Y_1), \cdots, G(Y_n)$. The result is summarized in the following theorem.

Theorem 2.2 For any subinterval [a,b] of (0,1), $0 < a < b < 1$, on which the density ratio $f(G^{-1}(t))/g(G^{-1}(t))$ is bounded, then

$$\begin{aligned} & \sqrt{n}(F_m G_n^{-1}(t) - FG^{-1}(t)) \\ \stackrel{a.s}{=} \; & \sqrt{\lambda} B_1(FG^{-1}(t)) + \frac{f(G^{-1}(t))}{g(G^{-1}(t))} B_2(t) + o_p(n^{-1/2}(logn)^2) \end{aligned}$$

uniformly on [a,b] , where $B_1(\cdot)$ and $B_2(\cdot)$ are two independent Brownian bridges.

Again the proof is given in the Appendix.

Now the area under the empirical ODC curve is equal to the Mann-Whitney statistic and is an unbiased estimator of $P(X < Y)$. Explicitly this statistic is given by:

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

The consistency of the Mann-Whitney statistic, $M_{m.n}$, as an estimator of $P(X < Y)$ follows directly from Theorem 2.1. By Theorem 2.2, we have

$$\sqrt{n}(M_{m,n} - P(X < Y)) \stackrel{D}{\Longrightarrow} N(0, \sigma^2)$$

where

$$\sigma^2 = var\left[\sqrt{\lambda} \int_0^1 B_1(FG^{-1}(t)) dt + \int_0^1 \frac{f(G^{-1}(t))}{g(G^{-1}(t))} B_2(t) dt\right]$$

$$= \lambda\, var\left[\int_0^1 B_1(FG^{-1}(t))dt\right] + var\left[\int_0^1 B_2(GF^{-1}(t))dt\right]$$

$$= \lambda \parallel F \cdot G^{-1} \parallel^* + \parallel G \cdot F^{-1} \parallel^* \tag{4}$$

and $\parallel h \parallel^* = \int_0^1 h^2 dt - \left(\int_0^1 h dt\right)^2$. This leads us to:

Theorem 2.3 : If $f(G^{-1}(t))/g(G^{-1}(t))$ is finite for all $0 < t < 1$, then

1. $M_{m,n} \xrightarrow{a.s} P(X < Y)$ as $n \to \infty$.

2. If, in addition, the above density ratio is bounded on [0,1], then $\sqrt{n}\,[M_{m,n} - P(X < Y)]$ converges asymptotically to a normal distribution with mean 0 and variance σ^2 , where σ^2 is defined in (4).

This theorem follows directly from Theorems 2.1 and 2.2. It is worthwhile noting that σ^2 in (4) is the same as the one obtained by considering the projection of $M_{m,n}$ viewed as a U-statistic.

3 The binormal model for grouped data

For the remainder of this article the binormal assumption, as discussed in Section 1, is assumed. Recall that this implies the existence of some arbitrary transformation H, say, that simultaneously converts the F and G distributions to normal ones with parameters μ_H, σ_H and μ_D, σ_D, respectively. Without loss of generality we can take $\mu_H = 0$, $\sigma_H = 1$. Defining $\mu = \mu_D$ and $\sigma = \sigma_D$, under this assumption, the true ODC curve has the form:

$$FG^{-1}(t) = \Phi(\mu + \sigma\Phi^{-1}(t)). \tag{5}$$

Since the ROC and ODC curves depend only on μ and σ, interest revolves mainly about the estimation of these quantities. (However nonparametric estimation of the transformation H is also of interest – see Hsieh (1991, Section 7.6.).) We consider two approaches. The first is the well-known method of Dorfman and Alf (1969), originally designed for discrete rating data. Our other approach is based on a generalized least square method.

In Section 3.1, we first give a description of the Dorfman and Alf (1969) procedure. The model implied in their procedure is shown to be asymptotically equivalent to a certain measurement error model. This leads us to a natural estimator of $(\mu,\ \sigma)$ which has an asymptotic distribution which can be explicitly determined. This distribution is compared to that of another estimator that is based on a generalized least squares approach, which we discuss in Section 3.2.

3.1 Dorfman and Alf analysis

In the analysis of ROC curves with discrete rating data, the method of Dorfman and Alf (1969) is the most commonly used technique. Typically grouped data come from an experiment in which there are two stimuli S_1 and S_2, say, (here, healthy or diseased), and a set of responses $R_i,\ i = 1, \cdots, k+1$, say. These responses take on rating values from "definitely unlikely", "probably unlikely", $\cdots$ "definite likely", say. (Of course, responses on a continuous scale can always be discretized to be in this form, by choosing some arbitrary cell boundaries and grouping the data.) The underlying assumptions for the method of Dorfman and Alf (1969) are as follows:

$< A1 >$ Stimulus S_1 leads to random variable Z distributed as stan-

dard normal $N(0,1)$; stimulus S_2 leads to random variable W distributed as $N(\mu, \sigma^2)$.

$< A2 >$ There exists a set of unknown cutoff points $z_1, \cdots, z_k$ such that, letting y denote the observed value of the random variable, Z or W, if

1. $y < z_1$, the response is R_1

2. $z_i < y \leq z_{i+1}$, the reponse is R_{i+1}

3. $y > z_k$, the response is R_{k+1}

$< A3 >$ Responses from the individual subjects (m with stimulus 1 and n with stimulus 2) are all mutually independent.

It follows that

$$\begin{aligned} P(R_i \mid S_1) &= \Phi(z_i) - \Phi(z_{i-1}) \\ \text{and } P(R_i \mid S_2) &= \Phi\left(\frac{z_i - \mu}{\sigma}\right) - \Phi\left(\frac{z_{i-1} - \mu}{\sigma}\right) \quad i = 1, \cdots k+1. \end{aligned}$$

and that the log likelihood function is given by

$$\sum_{j=1}^{2} \sum_{i=1}^{k+1} \gamma_{ij} log P(R_i \mid S_j) \tag{6}$$

where γ_{ij} is the number of responses R_i from stimulus S_j ($1 \leq i \leq k+1$, $j = 1, 2$).

The computation of the maximum likelihood estimate (MLE) of (μ, σ) based on (6) requires the solution of a system of k+2 nonlinear equations. A χ^2statistic can be used to test the goodness-of-fit to the model given by $< A1 > - < A3 >$. A computer program to carry out this analysis is available in Swets and Pickett (1982, Appendix D). However it should be

pointed out that the computation can be difficult when k is large, and the iterative procedure used in the programs can fail to converge.

It is interesting to note that the Dorfman and Alf (1969) model, $< A1 > - < A3 >$, can be expressed as a measurement error model. Denoting

$$(7) \qquad \alpha_i = \Phi(\frac{z_i - \mu}{\sigma}) = Pr(R_1 \cup R_2 \cdots \cup R_i \mid S_2) \;\; i = 1, \cdots, k$$

we see that

$$\begin{aligned} \Phi(z_i) &= \Phi(\mu + \sigma\Phi^{-1}(\alpha_i)) \\ (8) \qquad &= Pr(R_1 \cup R_2 \cdots \cup R_i \mid S_1), \quad i = 1, \cdots, k \end{aligned}$$

The quantities in (7) and (8) are naturally estimated by the sample proportions $\frac{1}{n}\sum_{\ell=1}^{i}\gamma_{\ell 2}$ and $\frac{1}{m}\sum_{\ell=1}^{i}\gamma_{\ell 1}$, respectively.

Hence we can write:

$$(9) \qquad \Phi^{-1}\left(\frac{1}{m}\sum_{\ell=1}^{i}\gamma_{\ell 1}\right) = \sigma\Phi^{-1}(\alpha_i) + \mu + \varepsilon_{1i}$$

$$(10) \qquad \Phi^{-1}\left(\frac{1}{n}\sum_{\ell=1}^{i}\gamma_{\ell 2}\right) = \Phi^{-1}(\alpha_i) + \varepsilon_{2i} \;\; i = 1, \cdots, k$$

where vectors $\varepsilon_1 = (\varepsilon_{11}, \cdots, \varepsilon_{1k})$ and $\varepsilon_2 = (\varepsilon_{21}, \cdots, \varepsilon_{2k})$ are independent. We see that (9) and (10) define a simple linear regression problem with intercept μ and slope σ, where the independent variable is subject to measurement error (see *e.g.* Fuller (1987, Chapter 1.1)). The asymptotic distributions of ε_1 and ε_2 depend on (μ, σ) and $(\alpha_1, \cdots, \alpha_k)$; they can be derived by the delta method, using Theorem 2.2, as in Lemma 3.1. We develop this further in Theorem 3.3 at the end of Section 3.2.

3.2 The generalized least squares method

We now return to the situation in Sections 1 and 2, where the measurements are on a continuous scale, but now the data are grouped in a different way from that in Section 3.1.. Recall from (5) that, under the binormal assumption, the OD curve has form

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

Hence, we have

$$\Phi^{-1}(FG^{-1}(t)) = \mu + \sigma\Phi^{-1}(t)$$

We now choose k numbers, $0 < \alpha_1 < \alpha_2 \cdots < \alpha_k < 1$ and denote

$$\beta_i = \Phi(\mu + \sigma\Phi^{-1}(\alpha_i)) \tag{11}$$

A natural estimator of β_i is

$$\hat{\beta}_i = F_m(G_n^{-1}(\alpha_i)) \qquad i = 1, \cdots, k. \tag{12}$$

The asymptotic distribution of $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_k)$ is given by the following lemma:

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

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, \Sigma_1 + \Sigma_2) + o_P(1)$$

and

$$\sqrt{n}(\Phi^{-1}(\hat{\beta}) - \Phi^{-1}(\beta)) \xrightarrow{D} N(0, \Sigma^*) + o_P(1) \tag{13}$$

where $\Sigma^* = C[\Sigma_1 + \Sigma_2]C$ with

$$C = \text{diag}(\cdots, \phi^{-1}(\sigma\Phi^{-1}(\alpha_i) + \mu), \cdots)$$

Here Σ_1 has (i,j)th entry equal $\lambda[(\beta_i \wedge \beta_j) - \beta_i \cdot \beta_j]$, and $\Sigma_2 = ABA$ with $A = diag(\cdots, \frac{\sigma\phi(\sigma\Phi^{-1}(\alpha_i)+\mu)}{\phi(\Phi^{-1}(\alpha_i))}, \cdots)$ and B has (i,j)th entry $[(\alpha_i \wedge \alpha_j) - \alpha_i\alpha_j]$.

The proof follows from Theorem 2.2 and use of the delta method.

From (11) and (13), we see there is a simple linear regression relationship of the form:

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

where $\varepsilon' = (\varepsilon_1, \cdots, \varepsilon_k)$, the error vector has the asymptotic covariance structure specified in (13) of Lemma 3.1. This suggests that an iterative generalized least squares algorithm can be used to find estimates of μ and σ. This algorithm proceeds as follows:

1. Fix $k \geq 3$, and choose $0 < \alpha_1 \leq \cdots \leq \alpha_k < 1$.
   (For example, we might choose $\alpha_i = i/(k+1)$ and k=10, say.)

2. Compute $\{\hat{\beta}_i, i = 1, \cdots, k\}$ as given by (11).

3. Obtain initial values of μ and σ by fitting intercept and slope by ordinary least squares with independent variables $\{\Phi^{-1}(\alpha_i); i = 1, \cdots, k\}$ and dependent variables $\{\Phi^{-1}(\hat{\beta}_i); i = 1, \cdots, k\}$.

4. These estimates of μ and σ are then plugged into the formula for Σ^* as given in the statement of Lemma 3.1, yielding $\hat{\Sigma}^*$, say.

5. Updated estimates of μ and σ are then computed using generalized least squares via the formula:

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

where M is the design matrix, *i.e.*

$$M' = \begin{pmatrix} 1, & \cdots, & 1 \\ \Phi^{-1}(\alpha_1), & \cdots, & \Phi^{-1}(\alpha_k). \end{pmatrix} \tag{15}$$

6. Repeat steps 4 and 5.

Following Carroll and Ruppert (1982), we recommend that only one or two iterations need be performed in Step 6.

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 the assumption of Lemma 3.1,

$$\sqrt{n}\begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma} - \sigma \end{pmatrix} \xrightarrow{D} N(0, (M'\Sigma^{*-1}M)^{-1})$$

as $n \to \infty$.

The proof is given in the Appendix.

In fact a similar theorem, but with a different asymptotic covariance matrix, applies to the rating data situation of Section 3.1. Suppose we replace the above preselected $\{\alpha_i, i = 1, \cdots, k\}$ in the IGLS algorithm by the sample proportions $\frac{1}{n}\sum_{\ell=1}^{i} \gamma_{\ell 2}$ as defined below (8). Denote the resulting estimators as $\hat{\mu}^*$ and $\hat{\sigma}^*$. Then we have

Theorem 3.3: With regression structure in (9) and (10) the estimators $\hat{\mu}^*$ and $\hat{\sigma}^*$ have asymptototic distribution given by:

$$\sqrt{n}\begin{pmatrix} \hat{\mu}^* - \mu \\ \hat{\sigma}^* - \sigma \end{pmatrix} \xrightarrow{D} N(0, (M'\Sigma^{*-1}M))^{-1} + M'\Sigma^{*-1} \cdot (C_0 \cdot B \cdot C_0)\Sigma^{*-1}M),$$

where M, B and Σ^* are as given above and $C_0 = diag(\cdots, \frac{1}{\phi(\Phi^{-1}(\alpha_i))}, \cdots)$.
The proof is given in the Appendix.

Just as the Dorfman and Alf (1969) algorithm can be used for continuous data, by grouping, so the IGLS algorithm can be used for discrete data by appropriate choice of k and the $\{\alpha_i, i = 1, \cdots, k\}$. However as the above development suggests, our computational experience has been that this iterative generalized least squares (IGLS) algorithm is to be preferred to the Dorfman and Alf (1969) algorithm.

For continuous data, there remains the problem of the choice of k and $\{\alpha_i, i = 1, \cdots, k\}$. We discuss this further when we describe our simulation results. This is analogous to the problem of choosing the grouping in the Dorfman and Alf (1969) setup when the data are continuous.

3.3 A simulation study and an adaptive procedure

To investigate the small sample performance of the estimators $\hat{\mu}$ and $\hat{\sigma}$ from the IGLS algrithm of Section 3.2, a small simulation experiment was conducted. In the algorithm only one step is used. Six binormal situations were simulated, in which (μ, σ) = (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 (MSEs) of the estimators, the first using $k = 5$ and $(\alpha_1, \ldots, \alpha_5) = (0.1,$ 0.2, 0.3, 0.4, 0.5), the second using $k = 8$ and $(\alpha_1, \ldots, \alpha_8) = (0.1, \ldots, 0.8)$. In some of the cases (*e.g.* $\mu = 0$, $\sigma = 2$), the estimators' performance is unsatisfactory as demonstrated by the large MSEs. 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 zero

or one. The Φ^{-1} transformation will then clearly lead to unstable estimates with moderate sample sizes.

To remedy this situation, an following adaptive method for selecting the $\{\alpha_i\}$ values is proposed:

(1) Fix a positive integer q,

(2) Take

$$\tilde{\alpha}_1 = \min\{i/n \mid F_m G_n^{-1}(i/n) \geq q/m\}$$

(3) and set

$$\tilde{\alpha}_{i+1} = \min\{i/n \mid F_m G_n^{-1}(i/n) - F_m G_n^{-1}(\tilde{\alpha}_i) \geq q/m\}$$

for $i = 1, \ldots, k(q)$ where $k(q)$ is the largest integer such that $\tilde{\alpha}_{k(q)} < 1$.

[Figure 2 and Table 1 about here.]

The IGLS algorithm is now applied now 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. For very steep ODC curves such as when $\mu = 2$ and $\sigma = 1$ or 2 (see Figure 2), it is clearly desireable 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. Bootstrapping could also be used to determine the choice of q.

It should be noted that the Dorfman and Alf (1969) estimates can also be unstable and that algorithm can fail to converge for reasons analogous to those discussed above.

4 The binormal model for continuous data

In the previous section, two procedures for estimating the ODC and ROC curves under the binormal model were proposed. Both involved grouping or discretizing the response data in some way. In this section, two estimation procedures will be discussed, neither of which require that the continuous data be grouped. The procedures are based on a particular semi-parametric model termed a "transformation" model, see *e.g.* Bickel (1986), Clayton and Cuzick (1986), and Dabrowska and Doksum (1988). These authors were concerned with survival data. For our application, this model implies the existence of an unknown rank preserving transformation, H say, such that $H(X)$ is distributed $N(0,1)$. Because of the binormal assumption this also implies that $H(Y)$ is distributed $N(\mu, \sigma^2)$ for some unknown parameters μ and σ. No parametric form is assumed for H; hence the term "semi-parametric" model. In Section 4.1 we discuss an approach based on a marginal likelihood, while in Section 4.2, a minimum distance estimation (MDE) approach is proposed. For the latter a locally asymptotic minimax (LAM) property is proved in Section 4.3.

4.1 A marginal likelihood approach

The transformation model under the binormal assumption can be rephrased as follows. Consider a group $\mathcal{H}$ of rank preserving transformations,

$$\mathcal{H} = \{H \mid H : \Re \to \Re \text{ is increasing and continuous } \}$$

There are underlying latent random variables $Z_1, \cdots, Z_m$ which are i.i.d. N(0,1) and $W_1, \cdots, W_n$ which are i.i.d. $N(\mu, \sigma^2)$. We observe only transformed variables

$$\begin{aligned} X_i &= H^{-1}(Z_i) \qquad i = 1, \cdots, m \\ \text{and } Y_j &= H^{-1}(W_j) \qquad j = 1, \cdots, n \end{aligned}$$

The transformation $H \in \mathcal{H}$ is unknown; however primary attention is paid to the estimation of unknown parameters μ, σ, since by (5), the ROC and ODC curves are determined solely by these two quantities under the binormal assumption.

Define the rank R_j of Y_j relative to $X's$ as the number of $X's$ less than Y_j. The vector $R = (R_1, \cdots, R_n)$ is invariant with respect to the group $\mathcal{H}$ and is in fact the maximal invariant statistic. It is therefore reasonable to estimate (μ, σ) based on the marginal likelihood, $L(R, \mu, \sigma)$, of R . We have

$$\begin{aligned} L(R, \mu, \sigma) &= E\{Pr(R \mid X)\} \\ &= E\left\{\prod_{j=1}^{n} Pr(R_j = r_j \mid X)\right\} \\ &= E\left\{\prod_{j=1}^{n} Pr(X_{(r_j)} \leq Y_j < X_{(r_j+1)} \mid X)\right\} \end{aligned}$$

$$= E\left\{\prod_{j=1}^{n} Pr(Z_{(r_j)} \leq W_j < Z_{(r_j+1)} \mid Z)\right\}$$

$$= E\left\{\prod_{j=1}^{n}\left[\Phi\left(\frac{Z_{(r_j+1)}-\mu}{\sigma}\right) - \Phi\left(\frac{Z_{(r_j)}-\mu}{\sigma}\right)\right]\right\} \quad (16)$$

where the expectation in (16) is taken with respect to $Z_{(1)}, \cdots, Z_{(m)}$, the order statistics of an i.i.d. sample of size m from a standard normal $N(0,1)$ distribution .

However, it is difficult to compute and maximize (16) to find maximum marginal likelihood estimators (MMLE) of μ and σ because of the very high dimension of the integration involved. Alternatively, an approximation to the MMLE can be obtained by using the "likelihood sampler" method of Doksum (1987). A large number, B say, of independent sets of standard normal order statistics are generated, $\{z_{(\ell)}^{(i)};\ 1 \leq i \leq B;\ 1 \leq \ell \leq m\}$, say. Estimates of μ and σ are then obtained by maximizing the quantity:

$$\hat{L}(R,\mu,\sigma) = \frac{1}{B}\sum_{i=1}^{B}\left\{\prod_{j=1}^{n}\left[\Phi\left(\frac{z_{(r_j+1)}^{(i)}-\mu}{\sigma}\right) - \Phi\left(\frac{z_{(r_j)}^{(i)}-\mu}{\sigma}\right)\right]\right\}$$

Of course this maximization is still quite difficult. Also some heuristic arguments and our empirical studies, not shown here, indicate that the MMLE of μ can be severely biased towards zero in some situations.

4.2 Minimum distance estimator

In this section, the minimum distance estimator (MDE) is constructed by finding the ODC curve based on the binormal model, *i.e.* 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.

Minimum distance estimation has been studied extensively beginning with the work of Wolfowitz (1957). Millar (1984) presented a general abstract approach. He showed that the MDE has the locally asymptotic minimax (LAM) property in all the examples he considered. In this paper, we follow his approach to show the asymptotic normality of the MDE, and then prove the LAM results for it in this transformation model. Here we follow a "projection" type of approach; a more direct approach (see *e.g.* Pollard (1980)) is taken by Hsieh (1991).

Let Θ be the set $\{(\mu,\sigma) \mid \mu \in \Re \text{ and } \sigma > 1\}$, and define $\{P_1^m \times P_2^n\}$ as a product measure on the product measure space $(\Omega_m \times \Omega_n, \mathcal{F}_m \times \mathcal{F}_n)$. The restriction that $\sigma > 1$ is not unreasonable if one thinks of the normal response as "noise" and the diseased response as "noise plus signal". (However, we can avoid this restriction if we modify the distance criterion (17) below so that the integral is over a closed interval excluding 0 and 1.) Let B_2 be the separable Hilbert space $L_2(\mu)$, where μ is the uniform measure on [0,1]. Defining

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

and the L_2 -distance measure as

$$\| \xi_{mn}(\theta) \| = \int_0^1 [\xi_{mn}(\theta)]^2 dt \tag{17}$$

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

$$\| \xi_{mn}(\hat{\theta}_{mn}) \| = \inf_{\theta\in\Theta} \| \xi_{mn}(\theta) \| .$$

For discussing asymptotic properties of the MDE, we will suppose that m and n tend to ∞ so that $\frac{m}{n} = \lambda$, where λ is some constant. We then write $\xi_n = \xi_{mn}$ and $\hat{\theta}_n = \hat{\theta}_{mn}$, suppressing the dependence on λ. We will also let $\theta_0 = (\mu_0, \sigma_0) \in \Theta$ denote the true unknown value of θ. We start by proving the asymptotic normality of $\hat{\theta}_n$. We apply the Theorem 3.6 of Millar (1984, Section III). Checking the conditions of that theorem, we see that his "identifiability" condition (3.2) 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 non-random and does not depend on n. Millar's convergence condition (3.4) holds because, from Theorem 2.2, the B_2-valued random process $\sqrt{n}\xi_n(\theta_0)$ converges in L_2 to $W(\theta_0)$, the sum of two Brownian bridges. Millar's differentiability condition follows because there is a continuous linear operator T $(= T_{\theta_0})$ from Θ to B_2, such that

$$\begin{aligned} \xi_n(\theta) &= \xi_n(\theta_0) + T(\theta - \theta_0) + o_p(\theta - \theta_0) \\ &= \xi_n(\theta_0) + \eta_1(\theta_0)(\mu - \mu_0) + \eta_2(\theta_0)(\sigma - \sigma_0) + o_p(\theta - \theta_0) \end{aligned} \tag{18}$$

where

$$\begin{aligned} \eta_1(\theta_0) &= \frac{\partial}{\partial\mu}\Phi(\mu + \sigma\Phi^{-1}(t)) \Big| \begin{matrix} \mu = \mu_0 \\ \sigma = \sigma_0 \end{matrix} \\ &= \phi(\mu_0 + \sigma_0\Phi^{-1}(t)) \end{aligned} \tag{19}$$

and

$$\begin{aligned} \eta_2(\theta_0) &= \frac{\partial}{\partial\sigma}\Phi(\mu + \sigma\Phi^{-1}(t)) \Big| \begin{matrix} \mu = \mu_0 \\ \sigma = \sigma_0 \end{matrix} \\ &= \Phi^{-1}(t)\phi(\mu_0 + \sigma_0\Phi^{-1}(t)) \end{aligned} \tag{20}$$

From (19) and (20), T is non-singular since $\eta_1(\theta_0)$ and $\eta_2(\theta_0)$ are linearly independent.

We let B_η denote the linear space spanned by $\eta_1(\theta_0)$ and $\eta_2(\theta_0)$, and π denote the projection mapping from B_2 onto B_η. The following theorem is Theorem 3.6 of Millar (1984) which gives the asymptotic normality of $\hat{\theta}_n$.

Theorem 4.1 With the identifiability, convergence and differentiability conditions above, then, with probability approaching 1, as $n \to \infty$, $\hat{\theta}_n$ exists and is unique. Moreover

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

In addition,

$$\begin{aligned} \sqrt{n}(\xi_n(\hat{\theta}_n) - \xi_n(\theta_0)] &\Longrightarrow \pi \circ W \quad \text{in } B_2 \\ \sqrt{n}(\hat{\theta}_n - \theta_0) &\Longrightarrow -T^{-1} \circ \pi \circ W \quad \text{in } \Re^2. \end{aligned}$$

With results of this theorem, we can obtain an explicit expression for the asymptotic convariance matrix, $\frac{1}{n}\Sigma$ say, for $\hat{\theta}_n$ as follows. First define the inner product in B_2 by $< h, k >= \int_0^1 h(t) \cdot k(t)dt$. Now define $R(s,t) = E(W(s) \cdot W(t))$ to be the covariance function of W. An explicit expression is given in (21) below. Finally define 2×2 matrices A and C by $C_{ij} =< \eta_i, \eta_j >$ and

$$\begin{aligned} A_{ij} &= E(< \eta_i, W > \cdot < \eta_j, W >) \\ &= \int_0^1 \int_0^1 \eta_i(s) \cdot R(s,t) \cdot \eta_j(t)dsdt \end{aligned}$$

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

Remark: We can compute $R(s,t)$ as follows:

$$R(s,t) = \frac{\lambda}{\sigma^2}\int_0^1\int_0^1 (t\wedge s - ts)\phi\left(\frac{\Phi^{-1}(t)-\mu}{\sigma}\right)\cdot\phi\left(\frac{\Phi^{-1}(s)-\mu}{\sigma}\right)dsdt$$

$$+\int_0^1\int_0^1 (t\wedge s - ts)\phi(\mu+\sigma\Phi^{-1}(t))\phi(\mu+\sigma\Phi^{-1}(s))dsdt$$

$$= \frac{\lambda}{\sigma^2}\left[\int_0^1\phi^2\left(\frac{\Phi^{-1}(t)-\mu}{\sigma}\right)dt - \left(\int_0^1\phi\left(\frac{\Phi^{-1}(t)-\mu}{\sigma}\right)dt\right)^2\right]$$

$$+\left[\int_0^1\phi^2(\mu+\sigma\Phi^{-1}(t))dt - \left(\int_0^1\phi(\mu+\sigma\Phi^{-1}(t))dt\right)^2\right]$$

$$= \frac{\lambda}{\sigma^2}\parallel\phi\left(\frac{\Phi^{-1}(t)-\mu}{\sigma}\right)\parallel^* + \parallel\phi(\mu+\sigma\Phi^{-1}(t))\parallel^* \tag{21}$$

where $\parallel h\parallel^* = \int_0^1 h^2(t)dt - \left(\int_0^1 hdt\right)^2$, as in (4).

This minimum distance approach also provides a natural statistic to test the binormal assumption, namely

$$\parallel\xi_n(\hat{\theta}_n)\parallel = \inf_{\theta\in\Theta}\parallel\xi_n(\theta)\parallel\ .$$

The following corollary gives the asymptotic distribution of this test statistic under the binormal assumption.

Corollary 4.2 :

$$n\parallel\xi_n(\hat{\theta}_n)\parallel \Longrightarrow \int_0^1[(1-\pi)\cdot W(t)]^2dt + o_p(1) \tag{22}$$

The proof follows immediately from Theorem 4.1,

The variance of random variable in the RHS of (22) can be obtained from

the covariance structure $R^*(s,t)$, say, of the process $(1-\pi) \circ W$ by rewriting

$$(1-\pi) \circ W = \sum_1^\infty \sqrt{\lambda_i} Z_i^* f_i$$

where $\{Z_i^*, i = 1, \cdots, \infty\}$ are i.i.d. $N(0,1)$ random variables and $\{f_i,\ i = 1, \cdots, \infty\}$ forms an orthonormal basis for B_2 such that

$$\int_0^1 R^*(s,t) f_i(s) ds = \lambda_i f_i(t) \quad i = 1, \cdots, \infty \quad \text{and}$$

$$\sqrt{\lambda_i} \cdot Z_i^* = \int_0^1 (1-\pi) \circ W(t) \cdot f_i(t) dt$$

The $\{\sqrt{\lambda_i} Z_i^*\}$ are the principal components of $(1-\pi) \circ W$. Hence

$$\int_0^1 [(1-\pi) \circ W]^2 dt = \sum_1^\infty \lambda_i z_i^{*2}$$

This argument is similar to the type found in Shorack and Wellner (1989, Chapter 5). It remains to evaluate $R^*(s,t)$. By definition:

$$\begin{aligned} (1-\pi) \circ W &= W - \pi \circ W \\ &= W + [\eta_1(\theta_0)(\hat{\mu} - \mu) + \eta_2(\theta_0)(\hat{\sigma} - \sigma)] \\ &= W + (\eta_1(\theta_0), \eta_2(\theta_0)) \cdot C^{-1} V \end{aligned} \tag{23}$$

where $V' = \left(\int_0^1 \eta_1(\theta_0) W(t) dt,\ \int_0^1 \eta_2(\theta_0) W(t) dt \right)$.

Let the random process, $\tilde{Z}$, represent the second term in (23). Then

$$\begin{aligned} R^*(s,t) &= cov\left[(W(s) + \tilde{Z}(s)), (W(t) + \tilde{Z}(t)) \right] \\ &= R(s,t) + cov(W(t), \tilde{Z}(s)) + cov(\tilde{Z}(t), \tilde{Z}(s)), \end{aligned}$$

where

$$cov(W(s), \tilde{Z}(t)) = E[W(s) \cdot \tilde{Z}(t)]$$

$$= (\eta_1(\theta_o), \eta_2(\theta_o)) \cdot C^{-1} \cdot E \begin{pmatrix} \int_0^1 \eta_1(\theta_o) \cdot W(\ell) \cdot W(s) d\ell \\ \int_0^1 \eta_2(\theta_o) \cdot W(\ell) \cdot W(s) d\ell \end{pmatrix}$$

$$= (\eta_1(\theta_o), \eta_2(\theta_o)) \cdot C^{-1} \begin{pmatrix} \int_0^1 \eta_1(\theta_o) \cdot R(\ell, s) d\ell \\ \int_0^1 \eta_2(\theta_o) \cdot R(\ell, s) d\ell \end{pmatrix}$$

and $cov(\tilde{Z}(t), \tilde{Z}(s))$ can be calculated in the same manner.

4.3 The LAM property of the MDE

Roughly, the locally asymptotic minimax (LAM) property of an estimator asserts its performance does not deteriorate when the actual distributions of the data depart slightly from those specified by the model. In this section, we prove that our MDE has this robustness property in the face of the binormal assumption. To prove this, we fit our problem into the general abstract framework of Millar (1984, Sections 3 and 5). We fix a point $\theta_0 \in \Theta$ as before and assume that $FG^{-1}(t) = \Phi(\mu_0 + \sigma_0 \Phi^{-1}(t))$ with $\theta_0 = (\mu_0, \sigma_0)$. We will establish the LAM property in a neighborhood of θ_0.

Consider the following abstract Wiener space (τ, H, B) with $\{Q_{h,k}\}$ as its Gaussian shift family. Let $H = H_1 \times H_2$, where H_1 and H_2 are Hilbert spaces defined below.

$$H_1 = \{h \mid \int h^2(s) f(s) ds < \infty \text{ and } \int h(s) f(s) ds = 0\}$$
$$H_2 = \{k \mid \int k^2(s) g(s) ds < \infty \text{ and } \int k(s) g(s) ds = 0\}.$$

Define

$$f(h,t) = f(t)(1 + h(t)) \text{ with } cdf\ F(h, \cdot) \text{ and}$$

$$g(k,t) = g(t)(1 + k(t)) \text{ with } cdf\ G(k, \cdot)\ .$$

Now we define

$$H_{01} = \{h \epsilon H_1 : \text{such that } f(m^{-\frac{1}{2}}h, \cdot) \geq 0 \text{ for large m}\}$$
$$H_{02} = \{k \epsilon H_2 : \text{such that } g(n^{-\frac{1}{2}}k, \cdot) \geq 0 \text{ for large n}\}$$

and let $H_0 = H_{01} \times H_{02}$.

Consider the transformation $\tau : H \rightarrow B$

$$\tau(h,k)(t) = (\int_0^t h(s)f(s)ds, \int_0^t k(s)g(s)ds)$$

Let μ be the uniform measure on the [0,1]. Then $\tau(h,k) \in L^2(\mu)$. B is taken to be the closure on $L^2(\mu)$ of τH. Denote Q_0 as the probability measure of the stochastic process $\{W_1^0(F(t)), W_2^0(G(t))\}$ where $W_1^0(\ )$ and $W_2^0(\ )$ are two independent Brownian bridges on [0,1]. Therefore (τ, H, B) is an abstract Wiener space since Q_0 is countably additive on B. We define

$$Q_{h,k}(A) = Q_0(A - \tau(h,k))$$

to be the Gaussian shift family for (τ, H, B).

Denote P_{1h}^m as the m-fold product measure of $f(m^{-1/2}h, \cdot)$ and P_{2k}^n as the n-fold product measure of $g(n^{-1/2}k, \cdot)$. Then, by a two-sample analog of the one-sample development in Millar (1984, page 391), we have that $\{P_{1h}^m \times P_{2k}^n, (h,k) \in H_0\}$ converges to $\{Q_{h,k}; (h,k) \epsilon H_0\}$, the Gaussian shift family of (τ, H, B) .

To consider the MDE under $P_{1h}^m \times P_{2k}^n$, let

$$\xi_n^*(\theta, P_{1h}^m \times P_{2k}^n) = \sqrt{n}[F_m \cdot G_n^{-1}(t) - \Phi(\mu + \sigma\Phi^{-1}(t))]$$

and

$$\zeta_n(\theta, P_{1h}^m \times P_{2k}^n) = F(m^{-\frac{1}{2}} \cdot h, \cdot) \circ G^{-1}(n^{-\frac{1}{2}}k, t) - \Phi(\mu + \sigma^{-1}\Phi^{-1}(t))$$

Therefore $\xi_n^*(\theta, P_{1n}^m \times P_{2k}^n)$ is a stochastic process with value in B_2 . Denote by $\theta_{n(h,k)}$ that point in Θ such that

$$\inf_\theta \| \zeta_n(\theta, P_{1h}^m \times P_{2k}^n) \| = \| \zeta_n(\theta_{n\cdot(h,k)}, P_{1h}^m \times P_{2k}^n) \|$$

Let $\Psi_n = \xi_n^*(\theta_0, P_1^m \times P_2^n)$. Then

$$\xi_n^*(\theta_0, P_{1n}^m \times P_{2k}^n) = \Phi_n + \tilde{V} \circ \tau(h,k) + o(1) \quad \forall (h,k) \in H_0$$

where

$$\tilde{V}(x_1, x_2) = x_1(G^{-1}(t)) \; - \; \frac{\phi(\mu_o + \sigma_o\Phi^{-1}(t))}{\phi(\Phi^{-1}(t))} x_2(G^{-1}(t)).$$

and

$$\xi_n^*(\theta, P_{1h}^m \times P_{2k}^n) = \xi_n^*(\theta_o, P_{1h}^m \times P_{2k}^n) - \sqrt{n}T \cdot (\theta - \theta_0) + o(1)$$

where T is defined in (18).

From the above relations, we have

$$\sqrt{n}(\theta_{n(h,k)} - \theta_0) = -T^{-1} \circ \pi \circ \tilde{V} \circ \tau(h,k) + o(1)$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta_o) = -T^{-1} \circ \pi \circ (\Psi_n + \tilde{V} \circ \tau(h,k)) + o(1)$$

Therefore

$$\sqrt{n}(\hat{\theta}_n - \theta_{n(h,k)}) \; = \; -T^{-1} \circ \pi \circ \tilde{V} \circ (\xi_n^*(\theta_0, P_1^m \times P_2^n)) + o(1)$$

Let N_c be a sequence of convex subsets of H_0, indexed by c, with $N_c \uparrow H_0$ as $c \to \infty$. Then by Theorems (5.12) and (5.20) of Millar (1984) we obtain:

Theorem 4.3 (LAM property of MDE)

$$\lim_{c\to\infty} \lim_n \inf_{\tilde{\theta_n}} \sup_{(h,k)\in N_c} \int \ell(\sqrt{n}(\tilde{\theta}_n - \theta_{n(h,k)}))dP_{1h}^m \times P_{2k}^n$$
$$= \int \ell(T^{-1} \circ \pi \circ \tilde{V} \circ z)\ Q_0(dz)$$
$$= \lim_{c\to\infty} \lim_n \sup_{(h,k)\in N_c} \int \ell(\sqrt{n}(\hat{\theta}_n -$$

where $\ell(x)$ is any subconvex function of $\| x \|$.

Appendix:

<Proof of Theorem 2.1 > Consider the following inequality

$$\sup_t \mid F_m \circ G_n^{-1}(t) - F \circ G^{-1}(t) \mid \leq \sup_t \mid F_m \circ G_n^{-1}(t) - F(G_n^{-1}(t)) \mid$$
$$+ \sup_t \mid F \circ G_n^{-1}(t) - F \circ G^{-1}(t) \mid \tag{24}$$

The first term on the RHS of (24) converges to zero by the Glivenko-Canteli theorem. We can rewrite the second term as

$$\sup_t \mid F \circ G_n^{-1}(t) - F \circ G^{-1}(t) \mid = \sup_t \mid F \circ G^{-1}(U_n(t)) - F \circ G^{-1}(t) \mid$$

where $U_n(t)$ is the uniform process induced by uniform random variables $\{G(Y_1), \cdot, G(Y_n)\}$. By assumption, $F \circ G^{-1}(\cdot)$ is a continuous and bounded function on [0,1]. Hence $F \circ G^{-1}(\cdot)$ is uniformly continuous. That is, for any $\varepsilon > 0$, there is a $\delta > 0$ such that

$$\sup_{|t_1 - t_2| < \delta} \mid F \circ G^{-1}(t_1) - F \circ G^{-1}(t_2) \mid < \varepsilon.$$

and

$$Pr(\sup_t \mid U_n(t) - t \mid > \delta) \geq Pr(\sup_t \mid F \circ G^{-1}(t) - F \circ G^{-1}(t) \mid > \varepsilon)$$

By the theorem of Dvoretzky, Kiefer and Wolfowitz (1965), we have

$$\begin{aligned}\sum_{n=1}^{\infty} Pr(\sup_t \mid F \circ G^{-1}(U_n(t)) - F \circ G^{-1}(t) \mid > \varepsilon) &\leq \sum_{n=1}^{\infty} Pr(\sup_t \mid u_n(t) - t \mid > \delta) \\ &\leq \sum_{n=1}^{\infty} c \cdot e^{-2\delta^2 n} < \infty\end{aligned}$$

Theorem 2.1 now follows by application of the Borel-Cantelli Lemma.

<Proof of Theorem 2.2 > From (2), (3) and results of Komlós et al (1975) concerning the strong approximation of empirical processes, we have

$$\sqrt{n}(F_m \cdot G_n^{-1}(t) - F \cdot G^{-1}(t))$$

is strongly approximated by $\sqrt{\lambda}B_1(FG^{-1}(t)) + \frac{f(G^{-1}(t))}{g(G^{-1}(t))}B_2(t)$ with error of order $(n^{-\frac{1}{2}}(\log n)^2)$ uniformly on [a,b].

< Proof of Theorem 3.2 > 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 it can be seen that $\sqrt{n}(\hat{\theta}_{op} - \theta)$ is asymptotically distributed as $N(0, (M'\Sigma^{*-1}M)^{-1})$.

By taking the starting estimator, $\hat{\theta}_0$ say, as the ordinary least squares estimator;

$$\begin{aligned}\hat{\theta}_0 &= (M'M)^{-1}M'\Phi^{-1}(\hat{\beta}) \\ &= \theta + (M'M)^{-1}M'\varepsilon.\end{aligned}$$

and we see that $\hat{\theta}_0$ is $\sqrt{n}$- consistent, since ε is $O_p(\frac{1}{\sqrt{n}})$.

Define $\hat{\Sigma}^{*-1}$ as the estimator of Σ^{*-1}, obtained by substituting $\hat{\theta}_0$ for $\theta_0 = (\mu, \sigma)$ in its definition. Thus we have

$$\hat{\Sigma}^{*-1} = \Sigma^{*-1} + \Delta_n, \tag{25}$$

where the Δ_n is of order $O_p(\frac{1}{\sqrt{n}})$ aswell. In order to prove the theorem, it is sufficent to show that the estimator;

$$\hat{\theta}_1 = (M'\hat{\Sigma}^{*-1}M)^{-1}M'\hat{\Sigma}^{*-1}\Phi^{-1}(\hat{\beta}),$$

as defined in the algorithm, is such that

$$\hat{\theta}_1 = \hat{\theta}_{op} + O_p(\frac{1}{n}).$$

With (25),we have

$$\begin{aligned}
\hat{\theta}_1 &= (M'\Sigma^{*-1}M + M'\Delta_n M)^{-1}(M'\Sigma^{*-1} + M'\Delta_n)\Phi^{-1}(\hat{\beta}) \\
&= [(M'\Sigma^{*-1}M)^{-1} - (M'\Sigma^{*-1}M)^{-1}(M'\Delta_n M)(M'\Sigma^{*-1}M)^{-1} \\
&\quad + O_p(\frac{1}{n})](M'\Sigma^{*-1} + M'\Delta_n)\Phi^{-1}(\hat{\beta}) \\
&= \hat{\theta}_{op} + [(M'\Sigma^{*-1}M)^{-1}M'\Delta_n - (M'\Sigma^{*-1}M)^{-1}(M'\Delta_n M)(M'\Sigma^{*-1}M)^{-1}(M'\Sigma^{*-1}) \\
&\quad + O_p(\frac{1}{n})]\Phi^{-1}(\hat{\beta}) \\
&= \hat{\theta}_{op} + O_p(\frac{1}{n})
\end{aligned}$$

since $\Phi^{-1}(\hat{\beta}) = M\theta + O_p(\frac{1}{\sqrt{n}})$. This completes the proof of this theorem.

< Proof of Theorem 3.3 > We define $k \times 2$ matrix $\hat{M}$ as for M in Step 5 of the IGLS algorithm, but with k-vector $\hat{\alpha}$ replacing vector α in the definition.

Recall from Section 3.1, the setup of the measurement error model,

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

Combining the above, we have the regression equation;

$$\Phi^{-1}(\hat{\beta}) = \hat{M}\theta + (\varepsilon_1 - \sigma\varepsilon_2). \tag{26}$$

Here ε_1 and ε_2 are independent error vectors, with $\sqrt{n}(\varepsilon_1 - \sigma\varepsilon_2)$ asymptotically distributed as multivariate normal, $N(0, \Sigma^*)$

From (26), the estimator $\hat{\theta}^*$ can be expressed as

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

where $\hat{M} = M + \delta_n$ with δ_n is a $k \times 2$ matrix with zeroes in its first column and ε_2 as its second column. From ordinary least squares, as in proof of Theorem 3.2, we have similarly that $\hat{\hat{\Sigma}}^{*-1} = \Sigma^{*-1} + \Delta_n^*$ with Δ_n^* being of order $O_p(\frac{1}{\sqrt{n}})$.

Also by calculations similar to those in the proof of Theorem 3.2, we have that:

$$\hat{\theta}_2 = \hat{\theta}_{op} - M'\Sigma^{-1*}\delta_n + O_p(\frac{1}{n}).$$

The covariance matrix of $\hat{\theta}_2$ is given by:

$$\begin{aligned} Cov(\hat{\theta}_2, \hat{\theta}_2) &= Cov(\hat{\theta}_{op}, \hat{\theta}_{op}) + M'\Sigma^{*-1} \cdot Cov(\varepsilon_2, \varepsilon_2)\Sigma^{*-1}M \\ &\quad + \sigma(M'\Sigma^{*-1}M)^{-1}M'\Sigma^{*-1} \cdot Cov(\varepsilon_1, \varepsilon_2)\Sigma^{*-1}M + o_p(\frac{1}{n}) \\ &= \frac{1}{n}[(M'\Sigma^{*-1}M)^{-1}) + M'\Sigma^{*-1} \cdot (C_0 \cdot B \cdot C_0)\Sigma^{*-1}M] + o_p(\frac{1}{n}). \end{aligned}$$

This completes the proof.

Acknowledgment The first author was supported in part by a fellowship from the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell University. The second author was supported in part by Grant GM 28364 from the U.S. National Institutes of Health.

References

1. Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. of Math. Psychology* **12**, 387-415.

2. Bickel, P.J. (1986) Efficient testing in a class of transformation models. *Papers on Semiparametric Model at ISI Centenary Session.* (R.D. Gill and M.V. Voors, eds.) 63-81.

3. Brownie, C., Habicht, J.-P. and Cogill, B. (1986). Comparing indicators of health or nutritional status. *Am. J. of Epidemiology* **124**, 1031-1044.

4. Carroll, R.J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10**, 429-441.

5. Centor, R.M. (1991). Signal detectability: The use of ROC curves and their analyses. *Medical Decision Making* **11**, 102-106.

6. Clayton, D. and Cuzick, J. (1986). The semiparametric Pareto model for regression analysis of suruival times. Proc. ISI. Amsterdam.

7. Dabrowska, D.M. and Doksum, K.A. (1988) Partial likelihood in transformation model with censored data. *Scand. J. Statist.* **15**, 1-23.

8. Delong, E.R., Delong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating curves: a non-parametric approach. *Biometrics* **44**, 837-845.

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

10. 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.

11. 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.

12. Erdreich, S.L. and Lee, E.L. (1981). Use of relative operating characteristic analysis in epidemiology *Am. J. Epidemiology* **114**, 649-662.

13. Fuller, W.A. (1987). *Measurement Error Models.* Wiley, New York.

14. 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.

15. Greenhouse, S.W. and Mantel, N. (1950). The evaluation of diagnostic tests. *Biometrics* **6**, 399-412.

16. Grey, D.R. and Morgan, B.J.T. (1972). Some aspects of ROC curve-fitting: normal and logistic models. *J. Math. Psychology* **9**, 128-139.

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

18. 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.

19. Hanley, J.A. and McNeil, B.J. (1983). A method of comparing the area under two ROC curves derived from the same cases. *Radiology* **148**, 839-843.

20. Hilden, J. (1991). The area under the ROC curve and its competitors. *Medical Decision Making* **11**, 95-101.

21. Hsieh, F.S. (1991). Performance of diagnostic tests in a nonparametric setting. Ph.D. Thesis, Cornell University.

22. Komlós, J., Major, P. and Tusnady, G. (1975). An approximation of partial sums of independent rv's and the sample distribution function, 1. *Z. Wahrsch. Verw. Geb.* **32**, 111-131.

23. Linnett, K.(1987). Comparison of quantitative diagnostic tests: Type I error, power, and sample size. *Statistics in Medicine* **6**, 147-158.

24. McClish, D.K. (1987). Comparing the areas under more than two independent ROC curves. *Medical Decision Making* **7**, 149-155.

25. McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**, 190-195.

26. Metz, C.E. (1986). ROC methodology in radiologic imaging. *Invest. Radiol.* **21**, 720-733.

27. Millar, P.W. (1984). A general approach to the optimality of minimum distance estimators. *Transactions of the American Mathematical Society* **286**, 377-418.

28. Ogilvie J.C. and Creelman, C.D. (1968). Maximum likelihood estimation of ROC curve parameters. *J. of Math. Psychology* **5**, 377-391.

29. Peterson, W.W., Birdsall, T.G. and Fox, W.C. (1954). *Trans. IRE Prof. Group Inf. Theory* **PGIT-4**, 171.

30. Pollard, D. (1980). The minimum distance method of testing. *Metrika* **27**, 43-70.

31. Rockette, H.E., Obuchowski, N.A. and Gur, D. (1990). Nonparametric estimation of degenerate ROC data sets used for comparison of imaging systems. *Investigative Radiology* bf 25, 835-837.

32. Shorack, G.R. and Wellner, J.A. (1986). *Empirical processes with application to statistics.* John Wiley, New York.

33. Swets, J.A. (1973). The relative operating characteristics in psychology. *Science* **183**

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

35. Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**, 1285-1293.

36. Thompson, M.L. and Zucchini,W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine* **8**, 1277-1290.

37. Wieand, 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.

38. Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28**, 75-88.

Table 1: Simulation study results. Estimates of μ and σ (with MSEs) for binormal model, using IGLS and adaptive IGLS methods

	IGLS Method				**Adaptive IGLS Method**					
μ, σ	$k = 5$		$k = 8$		$q = 5$		$q = 10$		$q = 12$	
(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)

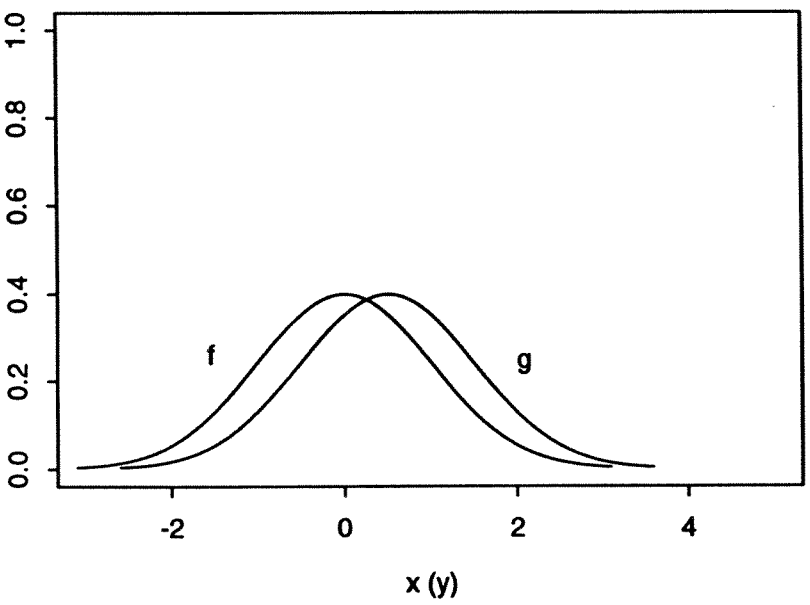
Note: Entries show mean estimates of μ, σ (with MSEs shown in parentheses) for six binormal models with (μ, σ) as given in the left hand column and are based on 100 replications. The situations simulated all use training sets of size $m = n = 100$.

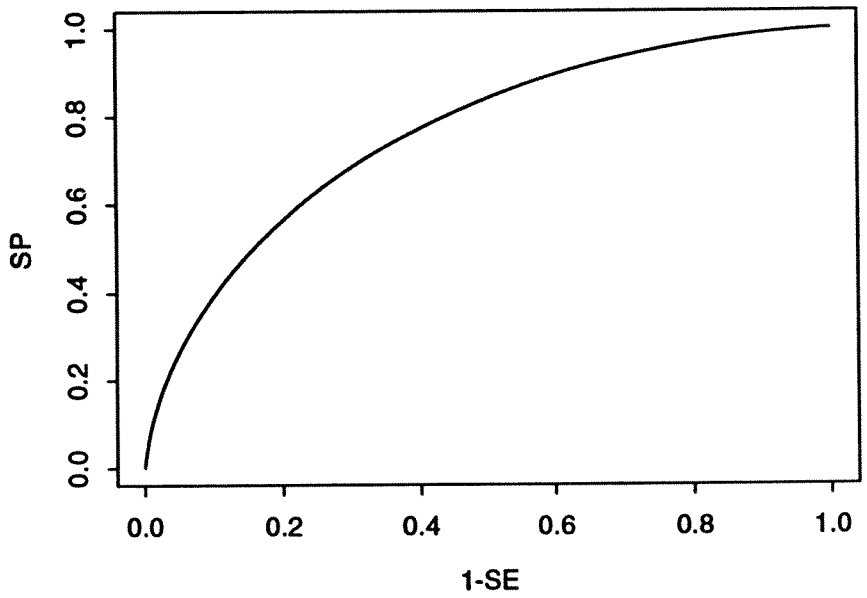
CAPTIONS FOR FIGURES

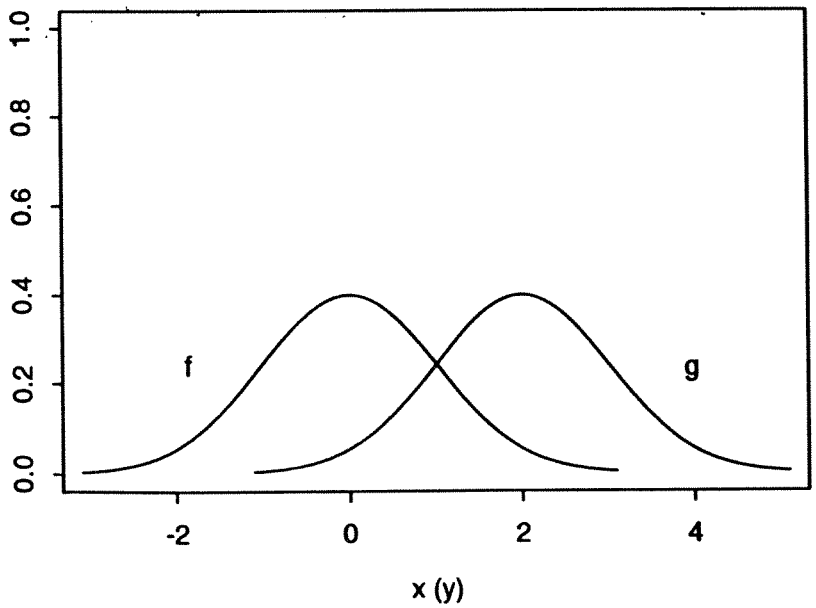
Figure 1. Two examples of densities f, g and their corresponding ODC curves. The diagnostic instrument represented by the lower curves is to be preferred.

Figure 2. Six binormal ODC curves used as models in the simulation study of Section 3.3.

Figure 1


0.0
0.2
0.4
0.6
0.8
1.0
f
g
-2
0
2
4
x (y)


SP
0.0
0.2
0.4
0.6
0.8
1.0
1-SE


0.0
0.2
0.4
0.6
0.8
1.0
f
g
-2
0
2
4
x (y)

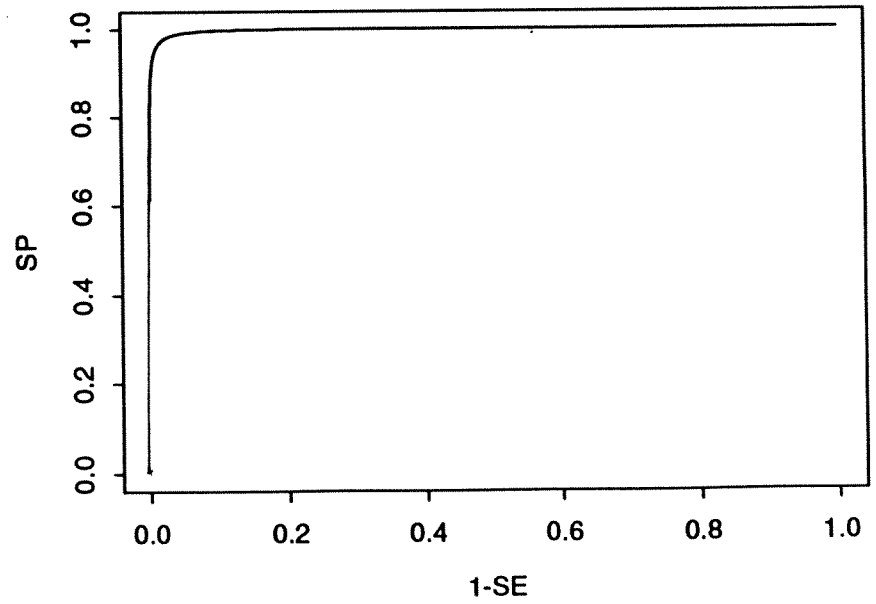
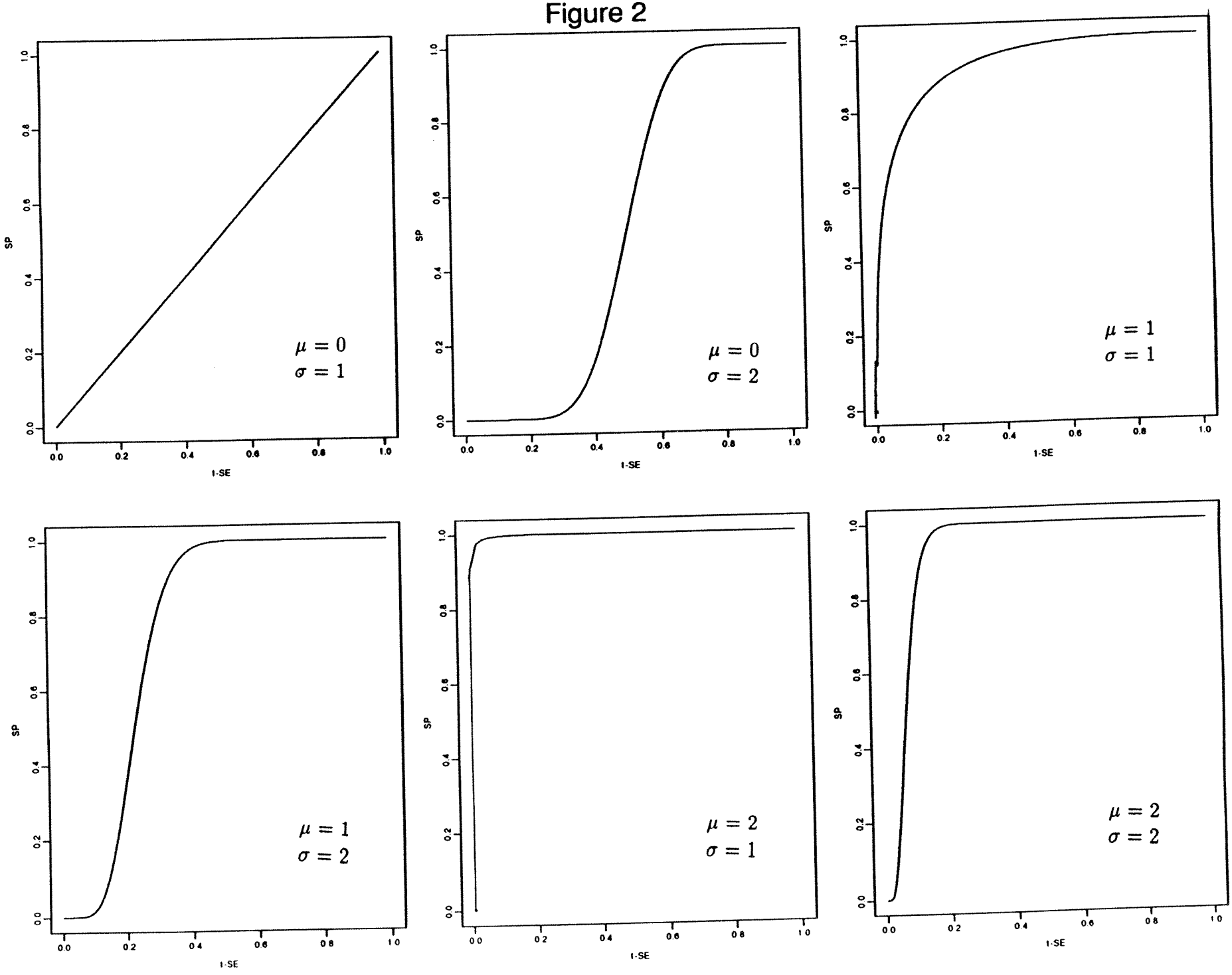

SP
0.0
0.2
0.4
0.6
0.8
1.0
1-SE

Figure 2


SP
1-SE
$\mu = 0$
$\sigma = 1$
$\mu = 0$
$\sigma = 2$
$\mu = 1$
$\sigma = 1$
$\mu = 1$
$\sigma = 2$
$\mu = 2$
$\sigma = 1$
$\mu = 2$
$\sigma = 2$