Statistica Sinica **6**(1996), 47-62

NONPARAMETRIC METHODS FOR EVALUATING DIAGNOSTIC TESTS

Fushing Hsieh and Bruce W. Turnbull

National Taiwan University and Cornell University

Abstract. We consider the performance of a diagnostic test based on continuous measurements in its ability to distinguish between healthy and diseased individuals. For a performance criterion we use Youden's (1950) index which is essentially the sum of the sensitivity and specificity. Based on available training set data, two types of nonparametric estimators for the optimal cutoff level and for the index are proposed. The first type is constructed from empirical distribution functions, the other from kernel smoothed density estimates. We compare their asymptotic properties, including rates of convergence. Finite sample properties are investigated by means of a small simulation study. Finally, the methods are applied to results of a glucose tolerance test for diabetes in a sample of 578 individuals from the NHANES-II study.

Key words and phrases: Classification, consistency, convergence rates, diagnostic markers, discrimination, empirical distribution function, empirical processes, kernel density estimate, sensitivity, specificity, Youden index.

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". 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. Typically the supports of X and Y will overlap, but we assume that:

(A1) there exists a value θ such that $g(\theta) = f(\theta)$, $g(t) < f(t)$ for $t < \theta$, and $g(t) > f(t)$ for $t > \theta$.

This is satisfied if, for example, the likelihood ratio is monotone. The assumption implies that X is stochastically smaller than Y, i.e. $F(t) \geq G(t)$ for all t.

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. A simple measure of the merit of a diagnostic test is the sum $SP(c) + SE(c)$, which under assumption (A1) is maximized by choosing $c = \theta$. We have

$$\begin{aligned} \max_c [SE(c) + SP(c)] &= SE(\theta) + SP(\theta) = 1 + F(\theta) - G(\theta) \\ &= 1 + \max_c [F(c) - G(c)]. \end{aligned} \tag{1}$$

Youden (1950) proposed $\eta = F(\theta) - G(\theta) = \max_c [F(c) - G(c)]$ as an index of performance of the diagnostic test and he listed a number of its desirable features. This index or measure assumes false positives and false negatives are equally undesirable. Gail and Green (1976) discussed a generalization whereby the index is a weighted sum of sensitivity and specificity. For simplicity we consider only Youden's original unweighted index, although our results can easily be extended. In any case, the relative cost of a false positive to a false negative is often difficult to ascertain. Brownie, Habicht and Cogill (1986) have used Youden's index for rating indicators of nutritional status (e.g. skin fold thickness, arm circumference, weight/height etc.) for a population of rural Bangladeshi children. The value θ is also clearly of interest as the value that yields the maximum in (1). In certain circumstances, θ also approximates the optimal choice of cutoff value for estimating the prevalence of the disease in a population (cf. Habicht and Brownie (1982), Brownie and Habicht (1984)).

When the distributions F and G are unknown, we wish to estimate the value of Youden's index η and the optimal cutoff value θ. We suppose that a training data set $X_1, \ldots, X_m$ of readings from the healthy population is available as is a set $Y_1, \ldots, Y_n$, from the diseased population. Our approach will be nonparametric and in the next section we consider estimators of η and θ , based on empirical distribution functions F_m and G_n for F and G, respectively. We will be concerned about asymptotic properties of the estimators as $m,\ n \to \infty$ with $m/n \to \lambda^2$, say, with $\lambda > 0$. In particular we give a theorem about the convergence in distribution of these estimators ($\hat{\eta}$, $\hat{\theta}$, say) with rates $n^{-1/2}$ and $n^{-1/3}$, respectively. In Section 3, we discuss alternative "smoothed" estimators, $\tilde{\theta}, \tilde{\eta}$ say, based on kernel density estimates of f and g and demonstrate their consistency and convergence properties. The rate of convergence of $\tilde{\theta}$ is shown to be the same as that of the density estimate and depends on the smoothness of the underlying densities, f and g. The estimator $\tilde{\eta}$ is shown to be $\sqrt{n}$ mean square consistent and has considerably lower mean square error than the empirical estimator $\hat{\eta}$. Due to space limitations, only heuristic outlines of the proofs of the results will be presented here. The detailed proofs, which are somewhat lengthy and technical in nature, are available from the authors upon request — Hsieh and Turnbull (1992). In Section 4, simulation results for comparing estimators discussed in

Section 2 and 3 are reported. Also we apply our methods to a glucose tolerance test for the diagnosis of diabetes based on data from the Second National Health and Nutrition Examination Survey (NHANES-II, 1976-1980).

There have been other approaches to the problem of assessment of the merit of a diagnostic test. Altham (1973) used a weighted sum of differences $\sum u_j[F(\xi_j) - G(\xi_j)]$ for given rating levels ξ_j and weights u_j, $1 \leq j \leq r$ for what she terms a measure of "signal discriminability". Greenhouse and Mantel (1950) proposed that a test be acceptable if there existed a cutoff point c such that $SE(c) > \alpha$ and $SP(c) > \beta$ for some prespecified fractions α and β . They went on to describe a hypothesis testing approach for determining whether a diagnostic test was acceptable under this criterion given an available training data set. Schäfer (1989) described a procedure where the cutoff value is chosen to be a specified sample quantile from the X sample or, alternatively, an upper confidence limit for $F^{-1}(p)$, for specified p. He illustrated his method with an application to a marker for bone marrow metastases in patients with small cell lung cancer. Miller and Siegmund (1982) estimated the cutoff point θ by choosing that value θ that maximized the Pearson chi-square statistic based on the 2×2 table formed when the healthy and diseased individuals in the training data set are classified as having test values either above or below θ. Halpern (1982) presented simulation results comparing this maximum chi-square-based statistic, one based on the maximum square of a standardized log cross-product ratio, and the statistic proposed by Gail and Green (1976). Yet another approach involves measures based on the receiver operating characteristic (ROC) curve, given by $1 - G(F^{-1}(1-t))$. For recent papers, see Swets (1988), Wieand et al. (1989), Goddard and Hinberg (1990).

Statistical evaluation of diagnostic tests has been important in many fields, including medicine, nutrition, epidemiology, psychology, electrical engineering and polygraph testing. We shall not attempt to give a review of the large amount of literature on the subject; much of it relates to binary or discrete responses rather than ones on a continuous scale, which is our concern. The reader is referred to the book by Swets and Pickett (1982), also the paper by Gastwirth (1987) with accompanying discussion.

2. An Empirical Estimate of η and θ

A natural estimate of η is obtained by replacing cdf's F and G in the definition by their empirical distribution functions, F_m and G_n, i.e.

$$\hat{\eta} = \max_x (F_m(x) - G_n(x)). \tag{2}$$

Analogously we can use the location of the maximum of (2) as an estimate of θ.

Since this may not be unique, we define the empirical estimator, $\hat{\theta}$, by

$$\hat{\theta} = \text{median}\{x_0 \mid F_m(x_0) - G_n(x_0) = \max_x (F_m(x) - G_n(x))\}. \tag{3}$$

(Alternatively, in the definition (3), we could use the maximum or minimum value instead of the median.) These estimators $\hat{\eta}$ and $\hat{\theta}$ are nonparametric generalized maximum likelihood estimators in the sense of Kiefer and Wolfowitz (1956).

The problem of estimating θ is similar to that of estimating the mode of a density function. Chernoff (1964) provided an estimator of the mode of a density with an $O_p(n^{-1/3})$ rate of convergence, whose distribution was expressed by means of a functional of Brownian motion with quadratic drift. More general development on this cube root asymptotics via functional limit theorems for empirical processes indexed by a class of functions can be found in Kim and Pollard (1990).

We will assume that θ is unique in the following sense:
(A1′) For any $\delta > 0$, there exists ε (> 0) , such that

$$\sup_{|x-\theta|>\delta} [F(x) - G(x)] < F(\theta) - G(\theta) - \varepsilon.$$

Note that (A1′) is slightly weaker than (A1). We shall be concerned with the asymptotic properties of our estimators, $\hat{\eta}$ and $\hat{\theta}$, as defined in (2) and (3). We assume that the sample sizes are increasing such that $m/n \to \lambda^2 (> 0)$, say.

The Glivenko-Cantelli theorem guarantees the strong uniform convergence of F_m and G_n to F and G, respectively. Then it can easily be shown that $\hat{\theta} \to \theta$ and $\hat{\eta} \to \eta$ almost surely. Details are given in Hsieh and Turnbull (1992).

Now we define a functional H by $H(H_1,\ H_2,\ x,\ \theta) = (H_1(x) - H_2(x)) - (H_1(\theta) - H_2(\theta))$ for any two functions H_1 and H_2 . Let $\mathcal{C}^{(k)}(C)$ denote the class of functions with a continuous k-th derivative on interval C, $C \subset \Re$. From the strong approximation of empirical processes (Csörgö and Révész (1981), Theorem 4.41), we have, almost surely:

$$\begin{aligned} &H(F_m, G_n, x, \theta) - H(F, G, x, \theta) \\ =&\frac{1}{\sqrt{m}} [B_1(F(x)) - B_1(F(\theta))] - \frac{1}{\sqrt{n}}[B_2(G(x)) - B_2(G(\theta))] + O(n^{-1} \log n). \end{aligned} \tag{4}$$

Here $\{B_1\}$ and $\{B_2\}$ are two independent Brownian bridge processes on [0,1]. Further, we assume F and G satisfy (A2) and (A3) below:
(A2) F and G are in $\mathcal{C}^{(2)}(a_0, b_0)$, for some a_0, b_0, with $\theta \in (a_0, b_0)$. F and G have connected intervals as their supports with intersection containing (a_0, b_0).
(A3) $|f'(\theta) - g'(\theta)| = a,\ a > 0$.

From (A2), if x is close to θ, we see that (4) is approximately distributed as

$$n^{-\frac{1}{2}}[\lambda^{-2}f(\theta)+g(\theta)]^{\frac{1}{2}}Z((x-\theta)), \tag{5}$$

where $Z(\cdot)$ is a two-sided standard Brownian motion, i.e. Brownian motion on $(-\infty,\infty)$ with $Z(0)=0$ (Chernoff (1964, page 35)). Also, the assumptions imply

$$H(F,G,x,\theta)\approx\frac{1}{2}(f'(\theta)-g'(\theta))(x-\theta)^2. \tag{6}$$

From (4),(5) and (6), we have

$$\max_x(H(F_m,G_n,x,\theta))=\max_x(H(F_m,G_n,x,\theta)-H(F,G,x,\theta)+H(F,G,x,\theta))$$

converges in distribution to

$$\max_x\{\frac{1}{\sqrt{n}}[\lambda^{-2}f(\theta)+g(\theta)]^{\frac{1}{2}}Z(x-\theta)-\frac{a}{2}(x-\theta)^2\}=Cn^{-\frac{2}{3}}\max_z(Z(z)-z^2), \tag{7}$$

where $z=(x-\theta)/\gamma$ with $\gamma=(\frac{4K}{na^2})^{\frac{1}{3}}$, $K=(\lambda^{-2}f(\theta)+g(\theta))$, $C=\frac{a}{2}(\frac{4K}{a^2})^{\frac{2}{3}}$ and a is as defined in (A3). As above, $Z(z)$ is defined as a two-sided standard Brownian motion process. Hence we have that

$$\sqrt{n}(\hat{\eta}-\eta)=\sqrt{n}[F_m(\theta)-G_n(\theta)-(F(\theta)-G(\theta))]\ +\ \max_x\{\sqrt{n}H(F_m,G_n,x,\theta)\}$$

converges in distribution to

$$\lambda^{-1}B_1(F(\theta))-B_2(G(\theta))+O_p(n^{-\frac{1}{6}}), \tag{8}$$

where B_1 and B_2 are two independent Brownian bridges.

The above results are summarized in the Theorem 2.1 below. A rigorous proof may be obtained by a modification of the proof of the main theorem in Kim and Pollard (1990).

Theorem 2.1. *Let F and G satisfy* (A1′), (A2) *and* (A3). *Then we have*
1. *$\hat{\eta}$ converges to η almost surely and $\sqrt{n}(\hat{\eta}-\eta)$ converges in distribution to $\lambda^{-1}B_1(F(\theta))-B_2(G(\theta))+O_p(n^{-\frac{1}{6}})$.*
2. *$\hat{\theta}$ converges to θ almost surely and $(\frac{a^2}{4K})^{\frac{1}{3}}n^{\frac{1}{3}}(\hat{\theta}-\theta)$ converges in distribution to the distribution of the random variable which maximizes process $(Z(z)-z^2)$; $z\in\Re$.*

Remark 1. From (7), it is clear that

$$\text{Bias}(\hat{\eta})\doteq Cn^{-\frac{2}{3}}E\{\max_z(Z(z)-z^2)\}$$

is always positive. Hsieh (1991) considered nonsmoothed bootstrap estimates of η which can reduce the bias, but the bootstrap bias-correction introduces extra variation and the simulation results given there indicate that bootstrapping does not lower the mean square error (MSE).

Remark 2. The MSE of $\hat{\eta}$ can be obtained by squaring (8) and taking the expectation.

$$nE(\hat{\eta}-\eta)^2 = \lambda^{-2}F(\theta)(1-F(\theta)) + G(\theta)(1-G(\theta)) + O(n^{-\frac{1}{3}}). \tag{9}$$

Theorem 2.1 shows that $\hat{\eta}$ is first order efficient in estimating η, doing as well asymptotically as if the true θ were known. However, in Section 3, we show that, under stricter smoothness conditions on F and G, another estimator of η can be constructed which yields a lower mean square error. Theorem 2.1 shows that $\hat{\theta}$ converges to θ at rate $n^{-1/3}$. Also in Section 3 we show that a better rate of convergence can be obtained if a smoother condition than (A2) is assumed. However under (A2), Hsieh and Turnbull (1995) have shown that $n^{-1/3}$ is the best rate in the sense of being local asymptotically minimax.

3. Smoothed Estimators of η and θ

The estimators of η and θ, $\tilde{\eta}$ and $\tilde{\theta}$ say, considered here are obtained by substituting kernel smoothed estimates in their definitions (2), (3). Their properties are compared with those of the estimators in Section 2; in particular, we show that $\tilde{\eta}$ has an asymptotic mean square error which is smaller than that of $\hat{\eta}$.

3.1. Estimation of θ

We now define kernel density estimates f_m and g_n of f and g, respectively and we will show that the estimator, $\tilde{\theta}$, defined as a solution of $f_m(x)-g_n(x)=0$, converges to θ at a certain rate.

First suppose $\gamma > 2$, let α be the largest integer less than γ and set $\beta = \gamma - \alpha$. Define $\mathcal{F}(\gamma,\gamma_1)$ to be the class of distribution functions $Q(x)$ of Hölder continuity of order γ. That is, they satisfy the following conditions:
(i) There exists (a_0,b_0), such that $Q(x) \in \mathcal{C}^{(\alpha)}(a_0,b_0)$ with $\theta \in (a_0,b_0)$.
(ii) $\sup |x_1-x_2|^{-\beta}|Q^{(\alpha)}(x_1)-Q^{(\alpha)}(x_2)| < \gamma_1$, over $(x_1,x_2) \in (a_0,b_0)$.
From here on, we assume that
(A2′) F and G are in $\mathcal{F}(\gamma,\gamma_1)$, for some γ_1 and $\gamma(>2)$:

In order to construct smooth density estimators of f and g, we need to introduce the kernel function $k(\cdot)$. This function can be taken to satisfy the following conditions:
(B1) $k(\cdot)$ is bounded and has a bounded continuous first derivative of bounded variation. Also, for some $\delta\,(>0)$, $|k(\cdot)|^{2+\delta}$ is integrable, and $\int k(z)dz = 1$, $I(k) =$

$\int k^2(z)dz < \infty$ and $H(r,k) = \int |z|^{\gamma-1}|k(z)|dz < \infty$. If not specified otherwise all integrals are over the domain $(-\infty, \infty)$. Finally, for any $\delta > 0$,

$$\frac{1}{h_n^j}\int_{\{z:|z|>\delta/h_n\}} |k^{(j)}|dz \to 0 \quad \text{for} \quad j = 0, 1 \quad \text{as} \quad h_n \to 0.$$

(B2) $k(\cdot)$ is an αth-order kernel. That is

$$\int z^j k(z)dz = 0, \quad j = 1, 2, \ldots \alpha - 1 \quad \text{and} \quad \int z^\alpha k(z)dz \neq 0.$$

Kernel density estimates, $f_m(x)$ and $g_n(x)$, of $f(x)$ and $g(x)$ are given by

$$f_m(x) = \frac{1}{m}\sum_1^m \frac{1}{h_m}k(\frac{x - x_i}{h_m}), \qquad g_n(x) = \frac{1}{n}\sum_1^n \frac{1}{h_n}k(\frac{x - y_i}{h_n})$$

with bandwidths $h_m = cm^{-1/(2\gamma-1)}$ and $h_n = cn^{-1/(2\gamma-1)}$ for an appropriate constant c.

For convenience, we now assume θ is the unique solution of the equation $f(x) = g(x)$ on $(a_0\ b_0)$ and maximizes $F(x) - G(x)$. Under the above convention, the condition (A1′) is equivalent to the following assumption (A1″):
(A1″) For $\delta > 0$, sufficiently small, there exists an $\varepsilon > 0$ such that

$$\inf|f(x) - g(x)| \ > \ \varepsilon, \text{ for } |x - \theta| \ > \ \delta \text{ and } x \in (a_0, b_0).$$

We define the estimator $\tilde{\theta}$ as follows:

$$\tilde{\theta} = \text{median}\{x | x \in (a_0, b_0), \text{and } f_m(x_0) = g_n(x_0)\}. \tag{10}$$

We now have the following theorem.

Theorem 3.1. *Let F and G satisfy* (A1″) *and* (A2′), *and kernel $k(\cdot)$ satisfy* (B1). *Then $\tilde{\theta}$ converges to θ almost surely. Further if* (A3) *is assumed, the equation $f_m(x) = g_n(x)$ has a unique solution almost surely as $m, n \longrightarrow \infty$.*

The proof of this strong consistency of $\tilde{\theta}$ is given in Hsieh and Turnbull (1992). Recall that, for our asymptotic theory, $m/n \to \lambda^2$. The next theorem shows that the rate of convergence of $\tilde{\theta}$ is $n^{-(\gamma-1)/(2\gamma-1)}$.

Theorem 3.2. *Assume that the underlying distribution functions F and G satisfy conditions* (A1″), (A2′) *and* (A3), *kernel function $k(\cdot)$ satisfies conditions* (B1) *and* (B2). *Then*

$$(nh_n)^{\frac{1}{2}}(\tilde{\theta} - \theta) \to Z + c^* \quad (\textit{in distribution})$$

as $n \to \infty$, where Z is normally distributed with mean 0 and variance σ^2 given by

$$\sigma^2 = [(\lambda^{\frac{4\gamma}{2\gamma-1}})f(\theta) + g(\theta)]I(k)/(f'(\theta) - g'(\theta))^2,$$

and

$$c^* = (\lambda^{\frac{2(2\gamma-2)}{2\gamma-1}})[C(\gamma, f, \theta) - C(\gamma, g, \theta)]c^{\frac{2\gamma-1}{2}} H(\gamma, k)/[g'(\theta) - f'(\theta)].$$

Here $I(k)$ and $H(\gamma, k)$ are as in Assumption (B1) *and $C(\gamma, f, \theta)$ is defined by*

$$\begin{aligned} & C(\gamma, f, \theta)h_m^\beta \int |z\,|^{r-1}|\; k(z)|\, dz\, (1 + o(1)) \\ = \; & \frac{(-1)^{\alpha-1}}{(\alpha-1)!} \int z^{(\alpha-1)}(f^{(\alpha-1)}(\theta - h_m z) - f^{\alpha-1}(\theta))k(z)dz \end{aligned}$$

and similarly for $C(\gamma, g, \theta)$.

The detailed proof is given in Hsieh and Turnbull (1992). From Theorem 3.2, it follows that the rate of convergence $n^{-(\gamma-1)/(2\gamma-1)}$ of $\tilde{\theta}$ is the same as the optimal rate for estimation of the density function under the same smoothness conditions (see e.g. Farrell (1972)). It is shown in Hsieh and Turnbull (1994) that this rate is indeed optimal in the sense of being local asymptotically minimax for estimating θ as well.

3.2. Estimation of η

To estimate η, we need first to construct kernel smoothed estimates, $\tilde{F}_m$ and $\tilde{G}_n$, say, of the distribution functions F and G. Because we are now estimating distribution functions rather than densities as in Section 3.1, we use a kernel function $\tilde{k}(\cdot)$ of order $\alpha + 1$, rather than α as above. (This can be seen from the Taylor expansion of the bias in (11) below.) Define kernel distribution $\tilde{K}(z) = \int_{-\infty}^{z} \tilde{k}(u)du$. Now we construct kernel smoothed estimates of F and G with bandwidths $h_m = cm^{-1/(2\gamma-1)}$ and $h_n = cn^{-1/(2\gamma-1)}$,

$$\tilde{F}_m(t) = \frac{1}{m}\sum_{i=1}^{m} \tilde{K}(\frac{t - x_i}{h_m}) \quad \text{and} \quad \tilde{G}_n(t) = \frac{1}{n}\sum_{j=1}^{n} \tilde{K}(\frac{t - y_j}{h_n}).$$

Then, we have the expectations

$$\begin{aligned} E(\tilde{F}_m(t)) &= F(t) + (-1)^\alpha \frac{h_m^\alpha}{\alpha!} \int z^\alpha [F^{(\alpha)}(t - h_m z) - F^{(\alpha)}(t)]\tilde{k}(z)dz \\ &= F(t) + C_1(\gamma, F, t)h_m^\gamma (1 + o(1)), \quad \text{say,} \end{aligned} \tag{11}$$

and similarly, $E(\tilde{G}_n(t)) = G(t) + C_1(\gamma, G, t)h_n^{\gamma}(1 + o(1))$. Variances are given by

$$\mathrm{Var}(\tilde{F}_m(t)) = \frac{1}{m}F(t)(1 - F(t)) - \frac{h_m}{m}f(t)d_0(1 + o(1)), \tag{12}$$

$$\mathrm{Var}(\tilde{G}_n(t)) = \frac{1}{n}G(t)(1 - G(t)) - \frac{h_n}{n}g(t)d_0(1 + o(1)), \tag{13}$$

where

$$d_0 = 2\int z\tilde{k}(z)\tilde{K}(z)dz. \tag{14}$$

From the above expressions, we choose the kernel $\tilde{K}$ such that d_0 defined above is positive in order that the variances in (12) and (13) are reduced. This we list as Assumption (B3).

(B3) $\tilde{K}$ is chosen so that d_0 in (14) is positive.

From (11) and (12) and by choosing suitable bandwidth constants in constructing the smoothed distribution estimators, it follows that the MSE of $\tilde{F}_m(t)$ is

$$E(\tilde{F}_m(t) - F(t))^2 = \frac{1}{m}(F(t)(1 - F(t))) - d^*\frac{h_m}{m}(1 + o(1)), \tag{15}$$

where $d^* > 0$ depends on d_0 and the bandwidth constant. That is, the smoothed distribution function, $\tilde{F}_m(t)$, has a MSE smaller than that of $F_m(t)$ by an amount of order $m^{-2\gamma/(2\gamma-1)}$. (In fact, this rate of improvement upon $F_m(t)$ can be shown to be the optimal one by using the argument of Hsieh and Levit (1991).)

We can now define the smoothed estimator, $\tilde{\eta}$, as follows:

$$\tilde{\eta} = \tilde{F}_m(\tilde{\theta}) - \tilde{G}_n(\tilde{\theta}), \tag{16}$$

where $\tilde{\theta}$ is defined in (10). We might expect that $\tilde{\eta}$ will improve upon $\hat{\eta}$ by a term that is of the same magnitude as the improvement upon MSE of $\tilde{F}_m(t)$ and $\tilde{G}_n(t)$ over $F_m(t)$ and $G_n(t)$. The following theorem says just this. Again the detailed proof can be found in Hsieh and Turnbull (1992).

Theorem 3.3. *We impose the same conditions on F, G and kernel $k(\cdot)$ as assumed in Theorem* 3.2. *Let $\tilde{k}$ be a kernel function of order $\alpha + 1$, uniformly continuous and of bounded variation. Also assume $\tilde{K}$ is bounded and satisfies* (B3). *Then, choosing a bandwidth of order $n^{-1/(2\gamma-1)}$ with appropriate bandwidth constants for kernels k and $\tilde{k}$, the MSE expansion of $\tilde{\eta}$ is*;

$$nE(\tilde{\eta} - \eta)^2 = \lambda^{-2}F(\theta)(1 - F(\theta)) + G(\theta)(1 - G(\theta)) - d_0^*h_n(1 + o(1)),$$

where $d_0^* > 0$ also depends on d_0 and the bandwidth constant.

Comparing this expression to (9) we see that the improvement in MSE by using $\tilde{\eta}$ over $\hat{\eta}$ can be substantial. Using the same methods mentioned above (Hsieh and Levit (1991)) it can be proved that this rate is optimal under the assumed conditions on F and G. It is also clear from (12) and (13) that a "good" kernel $\tilde{k}$ will be the one that gives a large value of d_0.

4. Simulations

Here we report the results of a small simulation study comparing the bias and mean square error (MSE) of various estimators of η and θ to see how they perform with finite samples. Simulated training sets of $m = 200$ X-values and $n = 200$ Y-values were generated where X is distributed as $\mathcal{N}(0,1)$ and Y as $\mathcal{N}(2\theta,1)$. Four values of θ were chosen, namely $\theta = 0.5, 1.0, 1.5$ and 2.0. Table 1 shows the mean values (with mean square errors in parentheses) for five different estimators of η based on 1000 simulations. The first estimator is $\hat{\eta}_1 = F_m(\overline{\theta}) - G_n(\overline{\theta})$, where $\overline{\theta} = \frac{1}{2}(\overline{X}_n + \overline{Y}_n)$. This estimator is a natural one to use if f and g are symmetric and differ only by a translation, as is the case simulated here. Of course, in practice, typically we would not know this. However, $\overline{\theta}$ serves as a convenient "gold standard" by which to judge the nonparametric estimators. The second estimator $\hat{\eta}_2 = \hat{\eta} = \max(F_m(x) - G_n(x))$ is that based on the empirical cdf's, as described in Section 2. The remaining estimators all require specification of a bandwidth constant c. A large number of values for c were investigated, but here we display results for only three choices, namely $c = 1.06,\ 0.5,\ 1.5$.. The value 1.06 was chosen following the suggestion by Silverman (1986, page 45), noting that the standard deviation σ is 1 here. The other two values straddle this value. The next two estimators are of the form $\tilde{\eta} = \tilde{F}_m(\tilde{\theta}) - \tilde{G}_n(\tilde{\theta})$. In both cases the argument $\tilde{\theta}$ is defined as in (10) with bandwidth $h = cn^{-1/5}$ and Gaussian kernels k for f_m and g_n. For the estimates of functions $\tilde{F}_m$, $\tilde{G}_n$, a Gaussian kernel $\tilde{k}$ was also used. However, for $\hat{\eta}_3$ we use bandwidth $h = cn^{-1/5}$, while for $\hat{\eta}_4$, the bandwidth is $h = cn^{-1/3}$. Here of course $n = 200$. The final estimator, $\hat{\eta}_5$, is defined as $\max(\tilde{F}_m(x) - \tilde{G}_n(x))$ using a Gaussian kernel $\tilde{k}$ for $\tilde{F}_m$ and $\tilde{G}_n$ with bandwidth $h = cn^{-1/3}$.

Table 2 shows results from the same simulation study for three estimators of the crossing point θ. The first estimator is $\overline{\theta} = \frac{1}{2}(\overline{X}_n + \overline{Y}_n)$, as discussed above. The second estimator is $\hat{\theta} = \arg\max(F_m(x) - G_n(x))$ as given in Section 2. The third estimator is $\arg\max(\tilde{F}_m(x) - \tilde{G}_n(x))$ using the same Gaussian kernel with bandwidth $h = cn^{-1/3}$. The last estimator is $\tilde{\theta}$ as defined in Theorem 3.1 as the solution to $f_m(x) = g_n(x)$.

It would be expected, in this simple situation where F and G are symmetric and differ only by location, that the "gold standard" estimators $\hat{\eta}_1$, $\overline{\theta}$, which use this information, perform the best. They did, both in terms of bias and mean square error (MSE). The non-smoothed estimators $\hat{\eta}_2$, $\hat{\theta}$, generally performed poorly. On the other hand, the smoothed estimators, with $c = 1.06$, performed equally as well as the "gold standard" estimators. However if the smoothing constant is too large, *e.g.* $c = 1.5$, there is "over-smoothing". The estimated

Table 1. Means and mean square errors (in parentheses) for various estimators of Youden's index η based on a simulation experiment.

bandwidth constant c	θ $\eta = \max(F(x) - G(x))$	$\theta = 0.5$ 0.38292	$\theta = 1.0$ 0.68269	$\theta = 1.5$ 0.86639	$\theta = 2.0$ 0.95450
not applicable	$\hat{\eta}_1$	0.38295 (0.00214)	0.68328 (0.00139)	0.86738 (0.00060)	0.95504 (0.00023)
	$\hat{\eta}_2$	0.41052 (0.00254)	0.70256 (0.00157)	0.88004 (0.00070)	0.96312 (0.00024)
1.06	$\hat{\eta}_3$	0.36337 (0.00182)	0.65372 (0.00177)	0.84231 (0.00102)	0.94078 (0.00035)
	$\hat{\eta}_4$	0.38143 (0.00172)	0.67762 (0.00114)	0.86214 (0.00052)	0.95262 (0.00018)
	$\hat{\eta}_5$	0.38300 (0.00168)	0.67853 (0.00112)	0.86279 (0.00051)	0.95310 (0.00018)
0.5	$\hat{\eta}_3$	0.38370 (0.00170)	0.67937 (0.00112)	0.86344 (0.00051)	0.95349 (0.00018)
	$\hat{\eta}_4$	0.39127 (0.00185)	0.68706 (0.00119)	0.86915 (0.00054)	0.95686 (0.00019)
	$\hat{\eta}_5$	0.39299 (0.00186)	0.68836 (0.00119)	0.87017 (0.00054)	0.95753 (0.00019)
1.5	$\hat{\eta}_3$	0.34226 (0.00299)	0.62421 (0.00431)	0.81646 (0.00293)	0.92411 (0.00108)
	$\hat{\eta}_4$	0.37248 (0.00187)	0.66692 (0.00137)	0.85411 (0.00066)	0.94769 (0.00021)
	$\hat{\eta}_5$	0.37361 (0.00182)	0.66752 (0.00135)	0.85455 (0.00065)	0.94810 (0.00020)

Notes:

1. $\hat{\eta}_1 = \max(F_m(x) - G_n(x))$, $\hat{\eta}_2 = F_m(\overline{\theta}) - G_n(\overline{\theta})$,
2. $\hat{\eta}_3 = {}_5\tilde{F}_m(\tilde{\theta}) - {}_5\tilde{G}_n(\tilde{\theta})$, $\hat{\eta}_4 = {}_3\tilde{F}_m(\tilde{\theta}) - {}_3\tilde{G}_n(\tilde{\theta})$, $\hat{\eta}_5 = \max({}_3\tilde{F}_m(x) - {}_3\tilde{G}_n(x))$.
   Normal kernels used with bandwidth constant c.
3. $\tilde{\theta}$ is defined in (10) with $h = cn^{-1/5}$ and $\overline{\theta} = \frac{1}{2}(\overline{X}_n + \overline{Y}_n)$.
4. ${}_k\tilde{F}_m$ and ${}_k\tilde{G}_n$ are smoothed distibution functions with bandwidth of order $n^{-1/k}$.

densities, f_m, g_n overlap more and so η is underestimated. In the other direction, if c is too small, *e.g.* $c = 0.5$, the estimated densities, f_m, g_n, become "rougher". This results in higher variability in the estimate of the crossing point θ, as reflected in the higher MSE values. Of course, for more general situations

Table 2. Means and mean square errors (in parentheses) for various estimators of the crossing point θ based on a simulation experiment.

bandwidth constant c	Estimator	$\theta = 0.5$	$\theta = 1.0$	$\theta = 1.5$	$\theta = 2.0$
not applicable	$\bar{\theta}$	0.50103 (0.00243)	1.00103 (0.00243)	1.50103 (0.00243)	2.00103 (0.00243)
	$\hat{\theta}$	0.48703 (0.05287)	0.98849 (0.03295)	1.46092 (0.03083)	1.93784 (0.03690)
1.06	$\ddot{\theta}$	0.50246 (0.03384)	1.00509 (0.01577)	1.49818 (0.01256)	2.00312 (0.01595)
	$\tilde{\theta}$	0.50109 (0.01795)	1.00158 (0.00725)	1.50142 (0.00589)	2.00008 (0.00772)
0.5	$\ddot{\theta}$	0.49992 (0.04498)	1.01184 (0.02573)	1.49574 (0.02257)	2.00528 (0.02594)
	$\tilde{\theta}$	0.49805 (0.03208)	1.00507 (0.01627)	1.49842 (0.01300)	2.00285 (0.01671)
1.5	$\ddot{\theta}$	0.50511 (0.02852)	0.99779 (0.00942)	1.49715 (0.00919)	2.00111 (0.01238)
	$\tilde{\theta}$	0.50053 (0.01359)	0.99887 (0.00496)	1.49827 (0.00458)	1.99943 (0.00561)

Notes:
1. $\bar{\theta} = \frac{1}{2}(\overline{X}_n + \overline{Y}_n)$.
2. $\hat{\theta}$ is the unsmoothed estimator defined in (3).
3. $\ddot{\theta}$ = location of maximum of ${}_3F_n(x) - {}_3G_n(x)$ with bandwidth constant c.
4. $\tilde{\theta}$ is defined in (10) with $h = cn^{-1/5}$.

where F and G were not known to be symmetric nor differ simply by location, the estimates, $\hat{\eta}_1$, $\bar{\theta}$, would be no longer applicable, and the smoothed estimators would be preferred. This simulation study, though of necessity limited in scope, does enable us to see the potential benefits of using the smoothed estimates.

Hsieh (1991) also carried out simulations to compare a smoothed bootstrap approach (De Angelis and Young (1992)) to obtain bias corrected estimates of η and θ. Although successful in reducing bias, the mean square errors were not significantly reduced and so the extra computation needed did not seem worthwhile when compared to the performance of the smoothed estimators used in Tables 1 and 2.

5. Application to NHANES Data

In this section, we apply the methods discussed in Sections 2 and 3 to a

training data set from the NHANES-II survey involving glucose tolerance measurements for the diagnosis of diabetes. For each individual, the data consist of three responses, namely fasting glucose level L_0, one-hour glucose level L_1 and two-hour glucose level L_2. These glucose levels are measured in the following fashion; the fasting glucose level is taken after this subject has been fasting for 12 hours. A 75-gram dose of oral glucose is then administered. The one- and two-hour glucose measurements are then taken after the corresponding intervals. For sample sizes we have $n = 96$ individuals in the diabetic group excluding 6 individuals with missing responses; for the healthy group we have $m = 482$, chosen from the first five hundred and excluding 18 individuals with missing responses. The data are available from the authors upon request. Usually, linear combinations of marker values offer improved performance (Su and Liu (1993)). A fourth diagnostic response variable L_3 can be constructed from a linear combination of the three glucose levels as given by,

$$L_3 = 0.5(L_0 + L_2) + L_1.$$

The weights are chosen such that this linear combination is the area under the polygon connecting the three glucose levels by line segments. (Note that an interesting problem, which is not addressed here, would be to construct the *optimal* way of combining the information from the three responses. However, the nonparametric methods described in this paper could certainly be used to evaluate various proposed discriminant functions.)

The nonsmoothed estimators $\hat{\eta}$, $\hat{\theta}$ and smoothed estimators $\tilde{\eta}$, $\tilde{\theta}$ for this data set are displayed in Table 3. For the smoothed estimators in this table $\tilde{F}_m$ and $\tilde{G}_n$ were constructed using a Gaussian kernel with bandwidths, $\hat{\sigma}_x m^{-1/3}$ and $\hat{\sigma}_y n^{-1/3}$, respectively, where $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are sample standard deviations. Here $\tilde{\theta}$ is the solution of equation $g_n(x) = f_m(x)$, also constructed with a Gaussian kernel, but with bandwidth $\hat{\sigma}_x m^{-1/5}$ and $\hat{\sigma}_y n^{-1/5}$ respectively.

Table 3. Comparison of diagnostic tests (Youden index and cutoff point) for diabetes based on NHANES-II Data

Tests	Fasting	1-hour	2-hour	L_3
$\hat{\eta}$	0.4174	0.5469	0.5300	0.5925
$\hat{\theta}$ $(mg/d\ell)$	160.0	187.0	141.0	306.5
$\tilde{\eta}$	0.4203	0.5298	0.5184	0.5634
$\tilde{\theta}$ $(mg/d\ell)$	142.2	198.5	145.7	311.5

From Table 3 it can be seen that the diagnostic variable L_3 has the highest

Youden index value η. It is interesting to note the following recommendation for classification and diagnosis of diabetes from the National Diabetes Data Group (1979, page 1040).

> "8. The diagnosis of diabetes in non-pregnant adults be restricted to (a) those with the classic symptoms of diabetes and unequivocal hyperglycemia; (b) those with fasting venous plasma glucose (PG) concentrations greater than or equal to 140 $mg/d\ell$ on more than one occasion; and (c) those who, if fasting plasma glucose is less than 140 $mg/d\ell$ exhibit sustained elevated venous PG values during the oral glucose tolerance test greater than or equal 200 $mg/d\ell$, both at 2-hours after ingestion of the glucose dose and also at some other time point between time 0 and 2-hr."

The table shows that the smoothed estimator of θ recovers the above recommendations on fasting and one-hour glucose levels. However, both the non-smoothed and smoothed method give much lower optimal cutoff values for a 2-hour glucose level than 200 $mg/d\ell$ as recommended.

Acknowledgements

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 a grant from the U.S. National Institutes of Health.

References

Altham, P. M. E. (1973). A non-parametric measure of signal discriminability. British. J. Math. Statist. Psych. **26**, 1-12.

Brownie, C. and Habicht, J.-P. (1984). Selecting a screening cut-off point or diagnostic criterion for comparing prevalences of disease. Biometrics **40**, 675-684.

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

Chernoff H. (1964). Estimation of the mode. Ann. Inst. Statist. Math. **16**, 31-41.

Csörgö, M. and Révész, P. (1981). Strong Approximations in Probability and Statistics. Academic Press, New York.

De Angelis D. and Young G. A. (1992). Smoothing the bootstrap. Internat. Statist. Rev. **60**, 45-56.

Farrell, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. Ann. Math. Statist. **43**, 170-180.

Gail, M. H. and Green, S. B. (1976). A generalization of the one-sided two sample Kolmogorov-Smirnov Statistic for evaluating diagnostic tests. Biometrics **32**, 561-570.

Gastwirth, J. L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDS Antibodies Test Data (with discussion). Statist. Sci. **2**, 213-238.

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.

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

Habicht, J.-P. and Brownie C. (1982). Reply to letter by Bairagi on best cutoff point for nutritional monitoring. Amer. J. Clinical Nutrition **35**, 369-371.

Halpern, J. (1982). Maximally selected chi square statistics for small samples. Biometrics **38**, 1017-1023.

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

Hsieh F. S. and Levit B. (1991). On the optimal rates of improvement of the sample median. Technical Report 684, Department of Mathematics, University of Utrecht. The Netherlands.

Hsieh F. S. and Turnbull W. B. (1992). Nonparametric methods for evaluating diagnostic tests. Technical Report No.1024, School of Operations Research, Cornell University.

Hsieh F. S. and Turnbull W. B. (1995). A note on the locally asymptotically minimax rate for estimating a crossing point in a diagnostic marker problem. Statist. Probab. Lett. **24**, 181-185.

Kiefer J. and Wolfowitz J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Statist. **27**, 887-906.

Kim J. and Pollard D. (1990). Cube root asymptotics. Ann. Statist. **18**, 191-219.

Mantel, N. (1951). Evaluation of a class of diagnostic tests. Biometrics **7**, 240-246.

Miller, R. and Siegmund, D. (1982). Maximally selected chi square statistics. Biometrics **38**, 1011-1016.

National Diabetes Data Group (1979). Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. Diabetes **28**, 1039-1057.

Pollard D. (1984). Convergence of Stochastic Processes. Springer-Verlag, New York.

Pollard D. (1990). Empirical Processes: Theory and Applications. NSF-CBMS Regional Conference series in Probability and Statistics Vol. 2., Institute of Mathematical Statistics, Hayward, Calif.

Romano, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. Ann. Statist. **16**, 629-647.

Schäfer, H. (1989). Constructing a cut-off point for a quantitative diagnostic test. Statistics in Medicine **8**, 1381-1391.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. Ann. Statist. **6**, 177-184.

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. J. Amer. Statist. Assoc. **88**, 1350-1355.

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

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

Youden, W. J. (1950). Index for rating diagnostic tests. Cancer **3**, 32-35.

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.

Department of Mathematics, National Taiwan University, Taipei, Taiwan.

School of Operations Research and Industrial Engineering, 227 Rhodes Hall, Cornell University, Ithaca, NY 14853-3801, U.S.A.

(Received August 1992; accepted February 1995)