Improved methods for bandwidth selection when estimating ROC curves

Peter G. Hall[1] and Rob J. Hyndman[1,2] 6 March 2003

Abstract: The receiver operating characteristic (ROC) curve is used to describe the performance of a diagnostic test which classifies observations into two groups. We introduce new methods for selecting bandwidths when computing kernel estimates of ROC curves. Our techniques allow for interaction between the distributions of each group of observations and give substantial improvement in MISE over other proposed methods, especially when the two distributions are very different.

Key words: Bandwidth selection; binary classification; kernel estimator; MISE; ROC curve.

[1]Centre for Mathematics and its Applications, Australian National University, Canberra ACT 0200, Australia.

[2]Department of Econometrics and Business Statistics, Monash University, VIC 3800, Australia. Corresponding author: Rob Hyndman (Rob.Hyndman@monash.edu.au).

1 INTRODUCTION

A receiver operating characteristic (ROC) curve can be used to describe the performance of a diagnostic test which classifies individuals into either group G_1 or group G_2. For example, G_1 may contain individuals with a disease and G_2 those without the disease. We assume that the diagnostic test is based on a continuous measurement T and that a person is classified as G_1 if $T \geq \tau$ and G_2 otherwise. Let $G(t) = \Pr(T \leq t \mid G_1)$ and $F(t) = \Pr(T \leq t \mid G_2)$ denote the distribution functions of T for each group. (Thus F is the specificity of the test and $1 - G$ is the sensitivity of the test.) Then the ROC curve is defined as $R(p) = 1 - G(F^{-1}(1-p))$ where $0 \leq p \leq 1$.

Let $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ denote independent random samples from G_1 and G_2 respectively, and let $\hat{F}$ and $\hat{G}$ denote their empirical distribution functions. Then a simple estimator of $R(p)$ is $\hat{R}(p) = 1 - \hat{G}(\hat{F}^{-1}(1-p))$, although this has the obvious weakness of being a step function while $R(p)$ is smooth.

Zou, Hall & Shapiro (1997) and Lloyd (1998) proposed a smooth kernel estimator of $R(p)$ as follows. Let $K(x)$ be a continuous density function and $L(x) = \int_{-\infty}^{x} K(u)\,du$. The kernel estimators of F and G are

$$\widetilde{F}(t) = \frac{1}{m}\sum_{i=1}^{m} L\left(\frac{t - X_i}{h_1}\right) \qquad \text{and} \qquad \widetilde{G}(t) = \frac{1}{m}\sum_{i=1}^{m} L\left(\frac{t - Y_i}{h_2}\right).$$

For the sake of simplicity we have used the same kernel for each distribution, although of course this is not strictly necessary. The kernel estimator of $R(p)$ is then

$$\widetilde{R}(p) = 1 - \widetilde{G}(\widetilde{F}^{-1}(1-p)).$$

Qiu & Le (2001) and Peng & Zhou (2002) have discussed estimators alternative to $\widetilde{R}(p)$.

Lloyd & Yong (1999) were the first to suggest empirical methods for choosing bandwidths h_1 and h_2 of appropriate size for $\widetilde{R}(p)$. This problem is also considered by Zhou and Harezlak (2002). However, both of these papers treat the problem as one of estimating F and G separately, rather than of estimating the ROC function R. We shall show that by adopting the latter approach one can significantly reduce the surplus of mean squared error over its theoretically minimum level. This is particularly true in the practically in-

teresting case where F and G are quite different. In the present paper we introduce and describe a bandwidth choice method which achieves these levels of performance.

A related problem, which leads to bandwidths of the correct order but without the correct constants, is that of smoothing in distribution estimation. See, for example, Mielniczuk, Sarda & Vieu (1989), Sarda (1993), Altman & Legér (1995), and Bowman, Hall & Prvan (1998).

2 METHODOLOGY

2.1 Optimality criterion and optimal bandwidths

If the tails of the distribution F are much lighter than those of G then the error of an estimator of F in its tail can produce a relatively large contribution to the error of the corresponding estimator of $G(F^{-1})$. As a result, if the L^2 performance criterion

$$\alpha_1(\mathcal{S}) = \int_{\mathcal{S}} \mathrm{E}\left[\widehat{G}\big(\widehat{F}^{-1}(p)\big) - G\big(F^{-1}(p)\big)\right]^2 dp \tag{2.1}$$

for a set $\mathcal{S} \subseteq [0,1]$, is not weighted in an appropriate way then choice of the optimal bandwidth in terms of $\alpha_1(\mathcal{S})$ can be driven by relative tail properties of f and g. Formula (A.1) in the appendix will provide a theoretical illustration of this phenomenon. We suggest that the weight be chosen equal to $f(F^{-1})$, so that the L^2 criterion becomes

$$\alpha(\mathcal{S}) = \int_{\mathcal{S}} \mathrm{E}\left[\widehat{G}\big(\widehat{F}^{-1}(p)\big) - G\big(F^{-1}(p)\big)\right]^2 f\big(F^{-1}(p)\big)\, dp\,. \tag{2.2}$$

We shall show in the appendix that for this definition of mean integrated squared error,

$$\alpha(\mathcal{S}) \sim \beta(\mathcal{S}) \equiv \int_{F^{-1}(\mathcal{S})} \Big\{\mathrm{E}[\widehat{F}(t) - F(t)]^2 g^2(t) + \mathrm{E}[\widehat{G}(t) - G(t)]^2 f^2(t)\Big\}\, dt \tag{2.3}$$

where $F^{-1}(\mathcal{S})$ denotes the set of points $F^{-1}(p)$ with $p \in \mathcal{S}$. Note particularly that the right-hand side is additive in the mean squared errors $\mathrm{E}(\widehat{F} - F)^2$ and $\mathrm{E}(\widehat{G} - G)^2$, so that in principle h_1 and h_2 may be chosen individually, rather than together. That is, if h_1 and

h_2 minimise

$$\beta_1(\mathcal{S}) = \int_{F^{-1}(\mathcal{S})} \mathrm{E}[\widehat{F}(t) - F(t)]^2 g^2(t)\,dt \quad \text{and} \quad \beta_2(\mathcal{S}) = \int_{F^{-1}(\mathcal{S})} \mathrm{E}[\widehat{G}(t) - G(t)]^2 f^2(t)\,dt,$$

respectively, then they provide asymptotic minimisation of $\alpha(\mathcal{S})$.

To express optimality we take $F^{-1}(\mathcal{S})$ equal to the whole real line, obtaining the global criterion $\gamma(h_1, h_2) = \gamma_1(h_1, h_2) + \gamma_2(h_1, h_2)$ where

$$\gamma_1(h_1, h_2) = \int_{-\infty}^{\infty} \mathrm{E}[\widehat{F}(t) - F(t)]^2 g^2(t)\,dt \quad \text{and} \quad \gamma_2(h_1, h_2) = \int_{-\infty}^{\infty} \mathrm{E}[\widehat{G}(t) - G(t)]^2 f^2(t)\,dt \tag{2.4}$$

Suppose K is a compactly supported and symmetric probability density, and f' is bounded, continuous and square-integrable. Then arguments similar to those of Azzalini (1981) show that

$$\mathrm{E}(\widehat{F} - F)^2 = m^{-1}\left[(1-F)F - h_1\,\kappa\,f\right] + \left(\frac{1}{2}\,\kappa_2\,h_1^2\,f'\right)^2 + o\big(n^{-1}\,h_1 + h_1^4\big)\,,$$

where $\kappa = \int (1 - L(u))L(u)\,du$, $\kappa_2 = \int u^2\,K(u)\,du$. Of course, an analogous formula holds for $\mathrm{E}(\widehat{G} - G)^2$, and so the formulae at (2.4) admit simple asymptotic approximations:

$$\begin{aligned} \gamma_1 &= m^{-1} \int (1-F)\,F\,g^2 + \delta_1 + o\big(m^{-1}\,h_1 + h_1^4\big) \\ \gamma_2 &= n^{-1} \int (1-G)\,G\,f^2 + \delta_2 + o\big(n^{-1}\,h_2 + h_2^4\big) \end{aligned}$$

where

$$\begin{aligned} \delta_1 &= -m^{-1}\,h_1\,\kappa \int f\,g^2 + \frac{1}{4}\,\kappa_2^2\,h_1^4 \int (f'g)^2 \\ \text{and} \qquad \delta_2 &= -n^{-1}\,h_2\,\kappa \int f^2\,g + \frac{1}{4}\,\kappa_2^2\,h_2^4 \int (fg')^2\,. \end{aligned}$$

The asymptotically optimal bandwidths are therefore

$$h_1 = m^{-1/3} c(f, g) \qquad \text{and} \qquad h_2 = n^{-1/3} c(g, f) \tag{2.5}$$

where

$$c(f, g)^3 = \left\{\kappa \int f(u)\,g^2(u)\,du\right\} \Big/ \left\{\kappa_2^2 \int [f'(u)\,g(u)]^2 du\right\}. \tag{2.6}$$

2.2 Normal-reference bandwidth selector

An early, and often very effective, approach to bandwidth selection was to employ the bandwidth that was asymptotically optimal for a Normal population that had the same scale; the latter might be measured by the interquartile range or the variance. See, for example, Silverman (1986, pp. 46–48). Taking the measure of scale to be variance, formulae (2.5) and (2.6) suggest taking $h_f = m^{-1/3} d(f,g)$ and $h_g = n^{-1/3} d(g,f)$, where

$$d(f,g) = \left(\frac{4\sqrt{\pi}\kappa}{\kappa_2^2} \, \frac{\sigma_f^3 (\sigma_f^2 + \sigma_g^2)^{5/2}}{(\sigma_g^2 + 2\sigma_f^2)^{1/2} [\sigma_g^4 + \sigma_g^2 \sigma_f^2 + 2\sigma_f^2 (\mu_f - \mu_g)^2]} \right.$$
$$\left. \times \exp\left[\frac{(\mu_f - \mu_g)^2 \sigma_f^2}{(\sigma_f^2 + \sigma_g^2)(2\sigma_f^2 + \sigma_g^2)} \right] \right)^{1/3}$$

and μ_f and σ_f, and μ_g and σ_g^2, denote the mean and variance of the distributions with densities f and g, respectively. Of course, the empirical forms of $d(f,g)$ and $d(g,f)$ use the sample versions of these moments.

As in the case of conventional density estimation, one can expect the Normal-reference method to perform well provided f and g are not too distant from Normal. Unimodal, moderately light-tailed densities are cases in point. However, when at least one of f and g is markedly multimodal an alternative approach, such as that proposed in the next section, can be necessary.

2.3 Plug-in bandwidth selector

A plug-in rule for choosing h_1 and h_2 may be developed directly from (2.5) and (2.6). First we construct an estimator of $c(f,g)$, as follows. Write H, H_1 and H_2 for new bandwidths, put

$$\hat{g}_{-i}(y \mid h) = \frac{1}{(n-1)\, h} \sum_{j \neq i} K\left(\frac{y - Y_j}{h} \right),$$

and let $\widehat{g^2}$ denote the leave-one-out kernel estimator of g^2:

$$\widehat{g^2}(y \mid H) = \frac{2}{n(n-1) H_2^2} \sum_{1 \le i_1 < i_2 \le n} \sum K\left(\frac{y - Y_{i_1}}{H} \right) K\left(\frac{y - Y_{i_2}}{H} \right).$$

Let $\widehat{f_1^2}$ be the corresponding estimator of $(f')^2$:

$$\widehat{f_1^2}(x \mid H_1) = \frac{2}{m(m-1)H_1^4} \sum\sum_{1 \le i_1 < i_2 \le m} K'\Big(\frac{x - X_{i_1}}{H_1}\Big)\, K'\Big(\frac{x - X_{i_2}}{H_1}\Big)\,.$$

Our estimator of $c(f,g)$ is $\hat{c}(f,g)$, defined by

$$\hat{c}(f,g)^3 = \Big(\frac{\kappa}{\kappa_2^2}\Big) \frac{m^{-1} \sum_{1 \le i \le m} \widehat{g^2}(X_i \mid H)}{n^{-1} \Big| \sum_{1 \le i \le n} \widehat{f_1^2}(Y_i \mid H_1)\, \hat{g}_{-i}(Y_i \mid H_2) \Big|}\,. \tag{2.7}$$

The analogous estimator of $c(g,f)$ is of course defined by switching the roles of the data $\{X_i\}$ and $\{Y_i\}$. Empirical versions of h_1 and h_2 are obtained by substituting $\hat{c}(f,g)$ and $\hat{c}(g,f)$ for $c(f,g)$ and $c(g,f)$ in (2.5).

The question then arises of how to choose the bandwidths H, H_1 and H_2 for constructing $\hat{c}$. The asymptotically optimal bandwidth H for computing the estimator $m^{-1} \sum_{1 \le i \le m} \widehat{g^2}(X_i \mid H)$ of $\int f g^2$ is

$$H = n^{-2/5} \left(\frac{2\, R\, \lambda_1 \int f\, g^3 + \lambda_2 \int f^2 g^2}{2\, \kappa_2^2\, (\int f\, g\, g'')^2} \right)^{1/5}, \tag{2.8}$$

where $R = n/m$, $\lambda_1 = \int K^2$ and $\lambda_2 = \int \{\int K(u)\, K(u+v)\, du\}^2\, dv$.

The asymptotically optimal bandwidth pair (H_1, H_2) for computing the estimator $n^{-1} \sum_{1 \le i \le n} \widehat{f_1^2}(Y_i \mid H_1)\, \hat{g}_{-i}(Y_i \mid H_2)$ of $\int (f'g)^2$ is the minimiser of

$$B(H_1, H_2) = \frac{1}{4}\, \kappa_2^2 \left\{2\, H_1^2\, I(f', g) + H_2^2\, I(g, f')\right\}^2 + \frac{J(f,g)\, \psi(H_1, H_2)}{mnH_1^2 H_2}\,, \tag{2.9}$$

where, given smooth functions a_1 and a_2,

$$I(a_1, a_2) = \int a_1 a_1'' a_2^2\,, \quad J(a_1, a_2) = 64 \int (a_1')^2 a_1 a_2^3\,,$$
$$\psi(H_1, H_2) = \int\int K'(u_1)\, K'(u_2)\, K\big\{(u_2 - u_1)\, H_1 / H_2\big\}\, du_1\, du_2\,.$$

See Appendix A.2 for an outline derivation of (2.9). Result (2.8) follows via a similar, but simpler, argument.

However, $B(H_1, H_2)$ is optimized for $H_1 = 0$. So we constrain the optimization of

B such that $\rho = H_1/H_2$ is fixed to the value that is optimal for distribution function estimation (we use the ratio of bandwidths proposed by Lloyd and Yong, 1999). Then, letting $H_1 = \rho H_2$, we obtain

$$B(H_2) \;=\; \tfrac{1}{4}H_2^4[2\rho^2 I(f',g) + I(g,f')]^2 + \frac{J(f,g)}{mn\sqrt{2\pi}H_2^3(1+2\rho^2)^{3/2}}$$

so that the optimal H_2 is

$$H_2 = \left(\frac{3J(f,g)}{mn\sqrt{2\pi}(1+2\rho^2)^{3/2}[2\rho^2 I(f',g) + I(g,f')]^2}\right)^{1/7}. \tag{2.10}$$

For both (2.8) and (2.9) it is assumed that $m \asymp n$, meaning that m/n is bounded away from zero and infinity as each increases, that f and g each have three continuous derivatives, and that the sixth power of each of the first three derivatives is integrable.

The integral functions of f and g in (2.8) and (2.10) are estimated by replacing f and g by Gaussian densities with means and variances computed from the data. The values of H, H_1 and H_2 obtained from (2.8) and (2.10) are then substituted into (2.7) to obtain an estimate of $c(f,g)$ which is then used to find h_1. An analogous procedure is used to find h_2.

2.4 Combining Normal-reference and plug-in methods

When the Normal-reference method fails it is generally because the density of the sampled distribution is significantly more complex than its Normal approximation. In particular, the former usually has two or more modes, and in such cases the optimal bandwidth is generally smaller than the Normal-reference one. (The larger, Normal-reference bandwidth tends to smooth out peaks and troughs in a density estimate, and so does not preserve modes particularly well.) On the other hand, when the sampled distributions are not far from Normal, the Normal-reference bandwidth's advantage of being substantially less variable than the plug-in rule can be a major asset. In this case it can deliver better performance.

These considerations also apply to the problem of bandwidth choice for estimating an ROC curve. They suggest the following rule for combining the Normal-reference and

plug-in approaches: let $h_x = m^{-1/3}e(f,g)$ and $h_y = n^{-1/3}e(g,f)$, where

$$e(f,g) = \min[c(f,g), d(f,g)]\,.$$

3 SOME SIMULATIONS

We compare the estimates obtained with our bandwidth selection method outlined in Section 2.4 to those obtained by Lloyd and Yong (1999) using their plug-in rule. Let

$$W(p) = E\left[\widetilde{G}\big(\widetilde{F}^{-1}(p)\big) - G\big(F^{-1}(p)\big)\right]^2 f\big(F^{-1}(p)\big) \tag{3.1}$$

denote mean squared error. Thus, mean integrated squared error, introduced at (2.2), is given by $\alpha(\mathcal{S}) = \int_{\mathcal{S}} W(p)\,dp$. The ideal but practically unattainable minimum of $W(p)$, for a nonrandom bandwidth, can be deduced by simulation, and will be denoted by $W_0(p)$. This value will be compared with its analogue, $W_1(p)$, obtained from (3.1) using the combination method described in Section 2.4. We also compare $W_2(p)$ obtained from (3.1) using the values of h_1 and h_2 chosen using the plug-in procedure suggested by Lloyd and Yong (1999).

We consider six examples, with F = N(0,1) in each case and G taking a variety of unimodal, bimodal and skewed shapes:

- Example 1: G = N(1,1);
- Example 2: G = N(2,1);
- Example 3: G = N(2,2);
- Example 4: G = equal mixture of $\mathrm{N}(-2,1)$ and $\mathrm{N}(2,1)$;
- Example 5: G = a $(1/5, 4/5)$ mixture of $\mathrm{N}(-2,1)$ and $\mathrm{N}(2,1)$;
- Example 6: G = equal mixture of $\mathrm{N}(-1,1)$ and $\mathrm{N}(1,1)$.

Figure 1 shows $D_i(p) = W_i(p) - W_0(p)$ for $i = 1$ and $i = 2$. The areas under the curves are $A_i(p) = \int D_i(p)\,dp$ and represent the increase in $\alpha(\mathcal{S})$ due to bandwidth selection. Table 1 shows these values for each example. The table shows that our method performs better than the method of Lloyd and Yong (in terms of the weighted MISE $\alpha(\mathcal{S})$) for all examples; the figure shows this improvement holds for almost all values of p (the most notable exception being small values of p in Example 3).

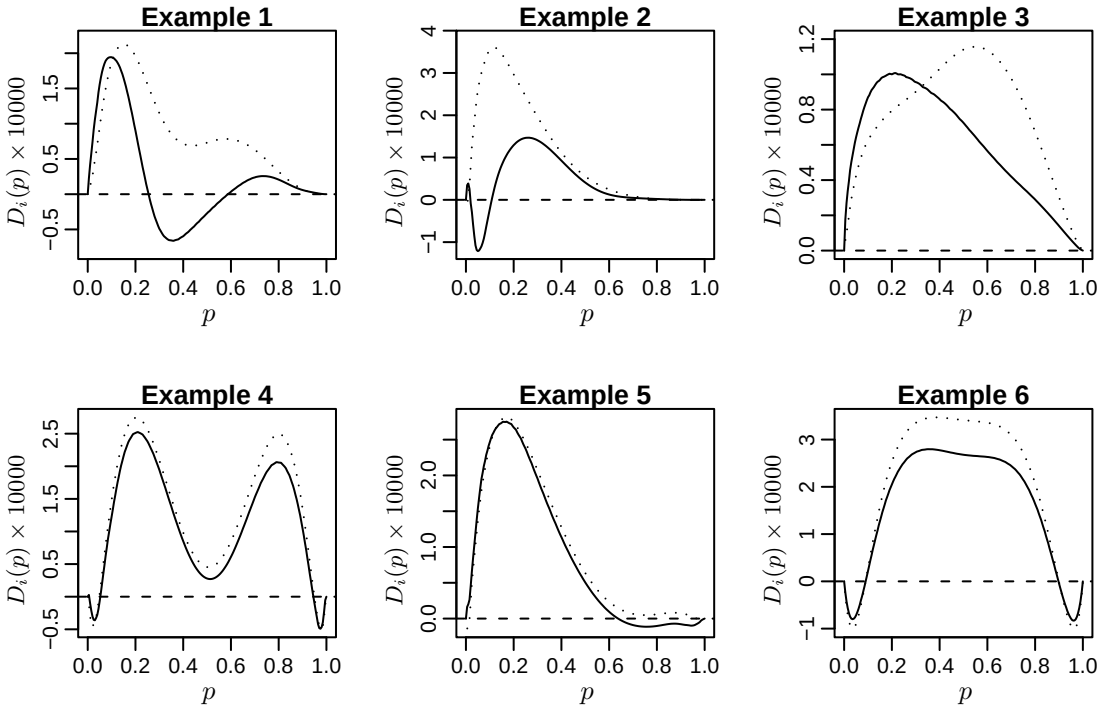

Figure 1: *Solid lines:* $D_1(p) = W_1(p) - W_0(p)$. *Dotted lines:* $D_2(p) = W_2(p) - W_0(p)$.

Example	$A_1(p)$	$A_2(p)$
1	0.83	0.24
2	1.10	0.38
3	0.77	0.60
4	1.40	1.20
5	0.97	0.88
6	2.10	1.60

Table 1: *Values of* $A_i(p) = \int [W_i(p) - W_0(p)]\, dp$ *for* $i = 1$ *(Lloyd and Yong's method) and* $i = 2$ *(the method described in Section 2.4). These values represent the increase in the weighted mean integrated square error due to bandwidth selection.*

APPENDIX

A.1: Derivation of (2.3)

Assume that f and g have continuous derivatives and are bounded away from 0 on $\mathcal{S}$. Put $A = \widehat{F} - F$, $B = \widehat{G} - G$ and $C = \widehat{F}^{-1} - F^{-1}$, and write I for the identity function. Then by Taylor expansion,

$$I = \widehat{F}\big(F^{-1} + C\big) = I + A\big(F^{-1}\big) + C\, f\big(F^{-1}\big) + o_p\big(|A(F^{-1})| + |C|\big)\,,$$

whence it follows that $C = -[A(F^{-1})/f(F^{-1})] + o_p(|A(F^{-1})|)$. Hence,

$$\widehat{G}\big(\widehat{F}^{-1}\big) - G\big(F^{-1}\big) \;=\; B\big(F^{-1}\big) - \frac{g(F^{-1})}{f(F^{-1})}\, A\big(F^{-1}\big) + o_p\big(|A(F^{-1})| + |B(F^{-1})|\big)\,. \quad \text{(A.1)}$$

Note the ratio $g(F^{-1})/f(F^{-1})$ on the right-hand side of (A.1). Since the variance of A equals $(1-F)F$ then the unweighted criterion α_1, defined at (2.1), can be largely determined by the value of $(g/f)^2(1-F)F$ in the tails if this quantity is not bounded.

Using instead the weighted criterion α, defined at (2.2), we may deduce from (A.1), related computations and the independence of the samples that

$$\int_{\mathcal{S}} \mathrm{E}\big[\widehat{G}\big(\widehat{F}^{-1}\big) - G\big(F^{-1}\big)\big]^2\, f\big(F^{-1}\big) \;=\; [1 + o(1)] \int_{F^{-1}(\mathcal{S})} \big[\mathrm{E}\big(B^2\big)\, f^2 + \mathrm{E}\big(A^2\big)\, g^2\big]$$

which is equivalent to (2.3).

A.2: Derivation of (2.9)

Put $\rho_1 = 4\{m(m-1)n(n-1)H_1^4 H_2\}^{-1}$, $\rho_2 = (m-1)\rho_1$, $\rho_3 = m\rho_2$, $\rho_4 = (n-1)\rho_3$, $\mu_1(y) = E[K'\{(y-X)/H_1\}]$, $\mu_2(y) = E[K\{(y-Y)/H_2\}]$, $\mu_3(y) = E[\mu_1(Y)^2\, K\{(Y-y)/H_2\}]$, $\mu_4 = E\{\mu_3(Y)\}$,

$$\Delta_1(i,j) = K'\left(\frac{Y_j - X_i}{H_1}\right) - \mu_1(Y_j)\,, \quad \Delta_3(j) = \mu_1(Y_j)^2\, \mu_2(Y_j) + \mu_3(Y_j) - 2\mu_4\,,$$

$$\Delta_2(j_1, j_2) = \mu_1(Y_{j_1})^2\, K\left(\frac{Y_{j_1} - Y_{j_2}}{H_2}\right) - \mu_1(Y_{j_1})^2\, \mu_2(Y_{j_1}) - \mu_3(Y_{j_2}) + \mu_4\,,$$

$$S = \rho_1 \sum\sum_{1 \le i_1 < i_2 \le m} \sum\sum_{1 \le j_1 < j_2 \le n} K'\left(\frac{Y_{j_1} - X_{i_1}}{H_1}\right) K'\left(\frac{Y_{j_1} - X_{i_2}}{H_1}\right) K\left(\frac{Y_{j_1} - Y_{j_2}}{H_2}\right),$$

$$S_1 = \sum_{1 \le i_1 < i_2 \le m} \sum \sum_{1 \le j_1 < j_2 \le n} \Delta_1(i_1, j_1)\, \Delta_1(i_2, j_1)\, K\left(\frac{Y_{j_1} - Y_{j_2}}{H_2}\right),$$
$$S_2 = \sum_{i=1}^{m} \sum \sum_{1 \le j_1 < j_2 \le n} \Delta_1(i, j_1)\, \mu_1(Y_{j_1})\, K\left(\frac{Y_{j_1} - Y_{j_2}}{H_2}\right), \quad S_3 = \sum \sum_{1 \le j_1 < j_2 \le n} \Delta_2(j_1, j_2)$$

and $S_4 = \sum_j \Delta_3(j)$. In this notation,

$$S = n^{-1} \sum_{i=1}^{n} \widehat{f_1^2}(Y_i \mid H_1)\, \hat{g}_{-i}(Y_i \mid H_2) = E(S) + \rho_1\, S_1 + 2\rho_2\, S_2 + \rho_3\, S_3 + \rho_4\, S_4\,.$$

The variables $S_1, \ldots, S_4$ each have zero mean, and $E(S_i S_j) = 0$ for $i \neq j$. Therefore,

$$\mathrm{Var} S = \rho_1^2\, \mathrm{Var} S_1 + 4\, \rho_2^2\, \mathrm{Var} S_2 + \rho_3^2\, \mathrm{Var} S_3 + \rho_4^2\, \mathrm{Var} S_4\,. \tag{A.2}$$

Assume $m \asymp n$, $H_1 \asymp H_2 \asymp H$ (say), $nH^5 \to 0$ and $nH^3 \to \infty$. Put $\xi = (n^2 h^3)^{-1}$. Lengthy algebraic arguments show that $E(\rho_1\, S_1)^2 + E(\rho_3\, S_3)^2 = o(\xi)$,

$$4\, E(\rho_2\, S_2)^2 = \frac{C_1}{m} + \frac{J(f,g)\, \psi(H_1, H_2)}{mnH_1^2 H_2} + o(\xi)$$

and $E(\rho_4\, S_4)^2 = n^{-1}\, C_2 + o(\xi)$, where $C_1, C_2 > 0$ depend only on f and g. Combining these results with (A.2) we deduce that

$$\mathrm{Var} S = C_1\, m^{-1} + C_2\, n^{-1} + \frac{J(f,g)\, \psi(H_1, H_2)}{mnH_1^2 H_2} + o(\xi)\,. \tag{A.3}$$

More simply, defining $T = \int (f'g)^2$ it may be shown that

$$E(S) - T = \frac{1}{2}\, \kappa_2 \left\{2\, H_1^2\, I(f', g) + H_2^2\, I(g, f')\right\} + o\big(H^2\big)\,, \tag{A.4}$$

Adding (A.3) and the square of (A.4) we deduce that, up to terms that either do not depend on H_1 or H_2 or are of second order, $E(S - T)^2$ equals $B(H_1, H_2)$ defined at (2.9). Furthermore, the values of H_1 and H_2 that result from minimising $B(H_1, H_2)$ are both of size $n^{-2/7}$, from which it follows that $E(S - T)^2 = C_1 m^{-1} + C_2 n^{-1} + O(n^{-8/7})$. This establishes root-$n$ consistency of our estimator, S, of T, in the sense of L_2 convergence.

REFERENCES

Altman, N. and Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *J. Statist. Plann. Inf.* **46**, 195–214.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika* **68**, 326–328.

Bowman, A.W., Hall, P. and Prvan, T. (1998). Cross-validation for the smoothing of distribution functions. *Biometrika* **85**, 799–808.

Lloyd, C.J. (1998). The use of smoothed ROC curves to summarise and compare diagnostic systems. *J. Amer. Statist. Assoc.* **93**, 1356–1364.

Lloyd, C.J. and Yong, Z (1999). Kernel estimators of the ROC curve are better than empirical. *Statist. Prob. Letters* **44**, 221–228.

Mielniczuk, J., Sarda, P. and Vieu, P. (1989). Local data-driven bandwidth choice for density estimation. *J. Statist. Plann. Inf.* **23**, 53–69.

Peng, L. and Zhou, X.-H. (2002). Local linear smoothing of receiver operator characteristic (ROC) curves. *J. Statist. Plann. Inf.*, to appear.

Qiu, P. and Le, C. (2001). ROC curve estimation based on local smoothing. *J. Statist. Comput. and Simul.* **70**, 55–69.

Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *J. Statist. Plann. Inf.* **35**, 65–75.

Silverman, B.W. (1986) *Density estimation for statistics and data analysis*. Chapman and Hall: London.

Zou, K.H., Hall, W.J. and Shapiro, D.E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* **16**, 2143–2156.

Zhou, X.-H., and Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine* **21**, 2045–2055.