ISSN 1440-771X

MONASH UNIVERSITY
AUSTRALIA

DEPARTMENT OF ECONOMETRICS AND BUSINESS STATISTICS

An Improved Method for Bandwidth Selection When Estimating ROC Curves

Peter G Hall and Rob J Hyndman

Working Paper 11/2002

An improved method for bandwidth selection when estimating ROC curves

Peter G. Hall[1] and Rob J. Hyndman[1,2] 13 September 2002

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 a new method for selecting bandwidths when computing kernel estimates of ROC curves. Our technique allows for interaction between the distributions of each group of observations and gives substantial improvement in MISE over other proposed methods, especially when the two distributions are very different.

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

JEL classification: C12, C13, C14.

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 samples of independent data from G_1 and G_2, 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, W.J. 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

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

[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).

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 and Yong (1999) were the first to suggest empirical methods for choosing bandwidths h_1 and h_2 of appropriate size for $\widetilde{R}(p)$, but they treated 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 interesting 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 and Vieu (1989), Sarda (1993), Altman and Legér (1995), and Bowman, Hall and 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

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

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

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

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

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

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

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

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:

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

where

$$\delta_1 = -m^{-1}\, h_1\, \kappa \int f\, g^2 + \tfrac{1}{4}\,\kappa_2^2\, h_1^4 \int (f'g)^2 \tag{2.5}$$

and

$$\delta_2 = -n^{-1}\, h_2\, \kappa \int f^2\, g + \tfrac{1}{4}\,\kappa_2^2\, h_2^4 \int (fg')^2 \tag{2.6}$$

The asymptotically optimal bandwidths are therefore

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

where

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

A conventional plug-in rule for choosing h_1 and h_2 may be developed directly from these formulae. However, it requires selection of pilot bandwidths for estimating f, g and their derivatives. The technique suggested in the next section avoids that difficulty.

2.2 Empirical choice of bandwidth

Let $\widehat{f^2}$ and $\widehat{g^2}$ denote leave-one-out kernel estimators of f^2 and g^2, respectively:

$$\begin{aligned}\widehat{f^2}(x|h_1) &= \frac{2}{m(m-1)h_1^2}\sum\sum_{1\le i_1<i_2\le m} K\Big(\frac{x-X_{i_1}}{h_1}\Big)\,K\Big(\frac{x-X_{i_2}}{h_1}\Big)\\ \widehat{g^2}(y|h_2) &= \frac{2}{n(n-1)h_2^2}\sum\sum_{1\le i_1<i_2\le n} K\Big(\frac{y-Y_{i_1}}{h_2}\Big)\,K\Big(\frac{y-Y_{i_2}}{h_2}\Big)\,.\end{aligned}$$

Let $\hat{f}_{-i}(x|h_1) = \{(m-1)\,h_1\}^{-1}\sum_{j\ne i} K\{(x-X_j)/h_1\}$, and define $\hat{g}_{-i}(y|h_2)$ analogously, and let $\widehat{f_1^2}$ and $\widehat{g_1^2}$ denote the kernel estimators of $(f')^2$ and $(g')^2$, respectively:

$$\begin{aligned}\widehat{f_1^2}(x|h_1) &= \frac{2}{m(m-1)h_1^4}\sum_{i_1=1}^{m}\sum_{i_2=1}^{m} K'\Big(\frac{x-X_{i_1}}{h_1}\Big)\,K'\Big(\frac{x-X_{i_2}}{h_1}\Big)\\ \widehat{g_1^2}(y|h_2) &= \frac{2}{n(n-1)h_2^4}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} K'\Big(\frac{y-Y_{i_1}}{h_2}\Big)\,K'\Big(\frac{y-Y_{i_2}}{h_2}\Big)\,.\end{aligned}$$

Note that the latter two estimators include all terms whereas the other estimators are "leave-one-out" estimators. We include the diagonal terms in the estimators of $(f')^2$ and $(g')^2$ as they act like ridge parameters and produce better empirical performance.

Now let

$$\begin{aligned}\Delta(h_1,h_2) &= -m^{-1}\,h_1\,\kappa\,m^{-1}\sum_{i=1}^{m}\widehat{g^2}(X_i|h_2) + \tfrac{1}{4}\,\kappa_2^2\,h_1^4\,n^{-1}\sum_{i=1}^{n}\widehat{f_1^2}(Y_i|h_1)\,\hat{g}_{-i}(Y_i|h_2)\\ &\quad - n^{-1}\,h_2\,\kappa\,n^{-1}\sum_{i=1}^{n}\widehat{f^2}(Y_i|h_1) + \tfrac{1}{4}\,\kappa_2^2\,h_2^4\,m^{-1}\sum_{i=1}^{m}\widehat{g_1^2}(X_i|h_2)\,\hat{f}_{-i}(X_i|h_1).\end{aligned}$$

We could choose h_1 and h_2 to minimize $\Delta(h_1,h_2)$. To motivate this approach, note that

$$\begin{aligned}\mathrm{E}\{\Delta(h_1,h_2)\} &= -m^{-1}\,h_1\,\kappa\int(\mathrm{E}\hat{g})^2 f + \tfrac{1}{4}\,\kappa_2^2\,h_1^4\int(\mathrm{E}\hat{f}')^2\,(\mathrm{E}\hat{g})\,g\\ &\quad -n^{-1}\,h_2\,\kappa\int(\mathrm{E}\hat{f})^2 g + \tfrac{1}{4}\,\kappa_2^2\,h_2^4\int(\mathrm{E}\hat{g}')^2\,(\mathrm{E}\hat{f})\,f\,,\end{aligned}\qquad(2.7)$$

which indicates that Δ is an almost-unbiased approximation to $\delta = \delta_1 + \delta_2$; compare (2.7) with the sum of the terms at (2.5) and (2.6). The relative size of stochastic error may also be shown to be asymptotically negligible. Indeed, if $m \asymp n$ as $n \to \infty$, if K is compactly supported and has a Hölder-continuous derivative, and if f and g are compactly supported and have three bounded derivatives, then $\Delta(h_1,h_2)/\delta(h_1,h_2)$ converges to 1 with probability 1, uniformly in $n^{-1+\epsilon} \le h_1, h_2 \le n^{-\epsilon}$ for each $0 < \epsilon < \frac{1}{2}$, as $n \to \infty$.

However, minimizing $\Delta(h_1,h_2)$ leads to some numerical instability. Instead, we constrain the minimization so that $h_1 = \rho h_2$ where $\rho = h_1^*/h_2^*$ and h_1^* and h_2^* are the bandwidths selected for estimating F and G using the plug-in rule proposed by Lloyd and Yong (1999). Minimizing $\Delta(h_1,h_2)$ under this constraint provides values of h_1 and h_2 which are suitable for estimating $\widetilde{R}(p)$.

3 SOME SIMULATIONS

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

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

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 values of h_1 and h_2 chosen using the method outlined in Section 2.2; and with $W_2(p)$, obtained from (3.1) using the values of h_1 and h_2 chosen using the plug-in procedure suggested by Lloyd and Yong (1995).

In our first example, illustrated in the first panel of Figure 1, we used Lloyd and Yong's (1999) model, where F and G are $\mathrm{N}(0,1)$ and $\mathrm{N}(1,1)$ respectively. In the second example we chose F and G to be more different; F was $\mathrm{N}(0,1)$ and G was an equal mixture of $\mathrm{N}(-2,1)$ and $\mathrm{N}(2,1)$. In both cases our method offers an improvement, which as expected is greater when the distributions are further apart. The areas under the curves represent the increase in $\alpha(\mathcal{S})$ due to bandwidth selection. In these terms our method improves on that of Lloyd and Yong (1999) by 1.2% and 28.6%, in the respective examples.

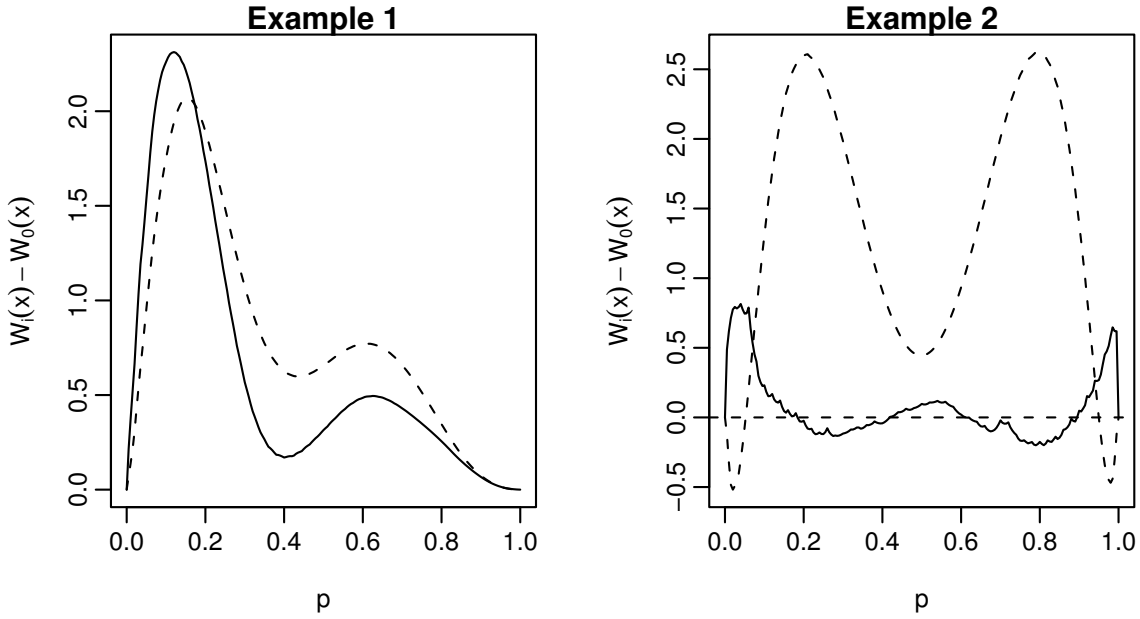

Figure 1: *Solid lines:* $W_1(p) - W_0(p)$. *Dashed lines:* $W_2(p) - W_0(p)$.

APPENDIX: 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,

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

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

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

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

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

which is equivalent to (2.3).

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.

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.