Nonparametric confidence intervals for receiver operating characteristic curves

Peter G. Hall[1], Rob J. Hyndman[2], and Yanan Fan[3] 17 July 2003

Abstract: We study methods for constructing confidence intervals, and confidence bands, for estimators of receiver operating characteristics. Particular emphasis is placed on the way in which smoothing should be implemented, when estimating either the characteristic itself or its variance. We show that substantial undersmoothing is necessary if coverage properties are not to be impaired. A theoretical analysis of the problem suggests an empirical, plug-in rule for bandwidth choice, optimising the coverage accuracy of interval estimators. The performance of this approach is explored. Our preferred technique is based on asymptotic approximation, rather than a more sophisticated approach using the bootstrap, since the latter requires a multiplicity of smoothing parameters all of which must be chosen in nonstandard ways. It is shown that the asymptotic method can give very good performance.

Key words: Bandwidth selection; binary classification; kernel estimator; receiver operating characteristic curve.

JEL classification: C12, C13, C14.

[1]Centre for Mathematics and its Applications, Australian National University, Canberra ACT 0200, Australia. Email: halpstat@maths.anu.edu.au.

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

[3]Department of Mathematics and Computer Science, University of Puerto Rico, PO Box 23355, San Juan, Puerto Rico, 00931-3355. Email: yfan@goliath.cnnet.clu.edu.

1 Introduction

A receiver operating characteristic curve is often used to describe the performance of a diagnostic test which classifies individuals into either group G_1 or group G_2. It is most commonly used with medical data where, 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. Then the receiver operating characteristic curve is defined as $R(p) = 1 - G\{F^{-1}(1-p)\}$ where $0 \leq p \leq 1$.

Zweig and Campbell (1993) discussed the importance and application of receiver operating characteristic plots in clinical medicine. See also Lloyd (1998), who addressed aspects of the plots' estimation and use. There is a rapidly growing literature on methods for estimating the plots, ranging from parametric approaches (e.g. Goddard and Hinberg, 1990) to nonparametric and semiparametric techniques (e.g. Hsieh and Turnbull, 1996; Li, Tiwari and Wells, 1999). Nonparametric methods range from those based on kernel ideas (e.g. Zhou, W. Hall and Shapiro, 1997; P. Hall and Hyndman, 2003) to techniques founded on local linear smoothing (Peng and Zhou, 2002). G. Claeskens and co-authors, in an unpublished 2002 manuscript, have considered empirical likelihood methods for constructing confidence intervals.

Against the background of this growing interest in both point and interval estimation, the present paper shows how the bandwidth used to construct an estimator of R influences the performance of pointwise confidence bands. For example, we demonstrate that bandwidths which are appropriate for point or curve estimation are not of the right size for good coverage accuracy. To achieve good performance in the latter sense, an order of magnitude less smoothing is necessary. This is true no matter whether asymptotic methods, or techniques based on the bootstrap, are used to construct the bands. However, we favour the asymptotic approach, since, as we show, bootstrap methods require a multiplicity of decisions about smoothing, all of them needing nonstandard solutions. This makes the bootstrap relatively unattractive for constructing confidence bands for receiver operating characteristic curves.

It is one thing to determine theoretically that undersmoothing is necessary, and quite another to develop a practicable technique for selecting the appropriate amount of undersmoothing. However, we shall show that the theoretical analysis which leads to our conclusions about undersmoothing, can also be employed to develop an explicit and effective device for selecting the correct amount of smoothing for confidence bands. The performance of this approach is demonstrated using both numerical simulation and theory.

Section 2 discusses point and curve estimators of receiver operating characteristic curves, and introduces the two main approaches to interval, or band, estimators, based on asymp-

totic and bootstrap methods, respectively. Theoretical properties of coverage are summarised in section 3, leading to the conclusions drawn two paragraphs above, and to the methodology mentioned in the previous paragraph. Numerical properties of our confidence bands are summarised in section 4. Technical arguments are outlined the Appendix.

2 Methodology

2.1 Distribution estimators

Suppose we are given independent random samples $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_m\}$ from distributions with respective distribution functions F and G. Let $\widehat{F}_{\rm emp}$ and $\widehat{G}_{\rm emp}$ denote the corresponding empirical distribution functions. For example, $\widehat{F}_{\rm emp}(x) = m^{-1} \sum_i I(X_i \le x)$, where $I(\mathcal{E})$ denotes the indicator function of an event $\mathcal{E}$. We could estimate the function, $R(p)$, by $\hat{R}(p) = 1 - \widehat{G}_{\rm emp}\{\widehat{F}^{-1}_{\rm emp}(1-p)\}$. However, $\widehat{F}_{\rm emp}$ and $\widehat{G}_{\rm emp}$ are discontinuous, and, especially if the sample sizes m and n differ, $\hat{R}(p)$ can have a very erratic appearance.

For this reason, and another given later in this section, it can be advantageous to smooth $\widehat{F}_{\rm emp}$ and $\widehat{G}_{\rm emp}$ prior to calculating the estimator of $R(p)$. To this end, let L be a known, smooth distribution function, let h_1 and h_2 denote bandwidths, and put

$$\widehat{F}(x) = m^{-1} \sum_{i=1}^{m} L\biggl(x - \frac{X_i}{h_1}\biggr), \quad \widehat{G}(y) = n^{-1} \sum_{i=1}^{n} L\biggl(y - \frac{Y_i}{h_2}\biggr). \tag{2.1}$$

Then $\widehat{F}$ and $\widehat{G}$ are smoothed versions of $\widehat{F}_{\rm emp}$ and $\widehat{G}_{\rm emp}$, respectively. Their derivatives, $\hat{f} = \widehat{F}'$ and $\hat{g} = \widehat{G}'$, are conventional kernel estimators of the densities $f = F'$ and $g = G'$, computed using the kernel $K = L'$. Optimal choice of bandwidth for $\widehat{F}$ and $\widehat{G}$ is quite different from that which is appropriate for $\hat{f}$ and $\hat{g}$, and indeed h_1 and h_2 at (2.1) can be often chosen quite small without seriously hindering performance. See, for example, Azzalini (1981), Reiss (1981), Mielniczuk, Sarda and Vieu (1989), Sarda (1993), Altman and Léger (1995), Bowman, P. Hall and Prvan (1998), de Una-Alvarez, Gonzalez-Manteiga and Cadarso-Suarez (2000) and Polansky and Baker (2000).

The kernel estimate of $R(p)$ is $\hat{R}(p) = 1 - \widehat{G}\{\widehat{F}^{-1}(1-p)\}$. Bandwidth choice for $\hat{R}(p)$ has been considered by Lloyd and Yong (1999), P. Hall and Hyndman (2003), Zou and W. Hall (2000), Zhou, W. Hall and Shapiro (1997) and Zhou and Harezlak (2002).

In Sections 2.2 and 2.3 we shall suggest asymptotic and bootstrap methods, respectively, for constructing pointwise confidence intervals for $R(p)$. The former are based on the normal distribution, and so make no attempt to capture the skewness of the estimator of $R(p)$. The latter have an opportunity for capturing skewness, but in both cases optimal performance can only be realised if the distribution function estimators are smoothed.

To appreciate why, note that while $\widehat{F}_{\rm emp}(x)$ can be represented by a sum of independent

and identically distributed random variables, the component variables $I(X_i \le x)$ are lattice-valued. It is known even from very early work on bootstrap approximation (Singh, 1981) that in such cases the bootstrap can fail to capture the main effects of skewness. Likewise, "rounding errors" are present in the coverage of confidence intervals and bands based on asymptotic approximations. Doing a little smoothing, for example using $\widehat{F}$ rather than $\widehat{F}_{\rm emp}$, can overcome these difficulties. To some extent the problems might be alleviated by the amount of smoothing that is implicit in studentisation, but this is unclear and quite awkward to verify. For these reasons, smoothing the distribution function estimator is advantageous from the viewpoint of improving the accuracy of the bootstrap, in addition to enhancing the appearance of an estimate of $R(p)$.

2.2 Asymptotic confidence intervals

It can be shown that, to a first-order approximation, if $0 < t < 1$ then $\widehat{G}\{\widehat{F}^{-1}(t)\} - G\{F^{-1}(t)\}$ is distributed as

$$\begin{aligned}\widehat{G}\{\widehat{F}^{-1}(t)\} - G\{F^{-1}(t)\} \\ \approx \widehat{G}\big\{F^{-1}(t)\big\} - G\big\{F^{-1}(t)\big\} - \frac{g\{F^{-1}(t)\}}{f\{F^{-1}(t)\}}\left[\widehat{F}\big\{F^{-1}(t)\big\} - t\right].\end{aligned} \tag{2.2}$$

See, for example, Hsieh and Turnbull (1993). For the relatively small bandwidths that would be used to construct $\widehat{F}$ and $\widehat{G}$, the quantity on the right-hand side of (2.2) is asymptotically normally distributed with zero mean and variance given by

$$\sigma(t)^2 = n^{-1}\, G\big\{F^{-1}(t)\big\}\left[1 - G\big\{F^{-1}(t)\big\}\right] + m^{-1}\,\frac{g\{F^{-1}(t)\}^2}{f\{F^{-1}(t)\}^2}\,t\,(1-t)\,. \tag{2.3}$$

Here we have made use of the assumption that $\mathcal{X}$ and $\mathcal{Y}$ are independent samples of independent data.

Replacing F, G, f and g at (2.3) by respective estimators $\widehat{F}$, $\widehat{G}$, $\tilde{f}$ and $\tilde{g}$, we obtain an estimator of σ:

$$\hat{\sigma}(t)^2 = n^{-1}\,\widehat{G}\big\{\widehat{F}^{-1}(t)\big\}\left[1 - \widehat{G}\big\{\widehat{F}^{-1}(t)\big\}\right] + m^{-1}\,\frac{\tilde{g}\{\widehat{F}^{-1}(t)\}^2}{\tilde{f}\{\widehat{F}^{-1}(t)\}^2}\,t\,(1-t)\,. \tag{2.4}$$

We might take $\tilde{f}$ and $\tilde{g}$ here to be simply the estimators $\hat{f}$ and $\hat{g}$, noted earlier. However, a substantially different size of bandwidth can be necessary when optimising confidence intervals for coverage accuracy, relative to that which is appropriate when constructing distribution or density estimators with good pointwise accuracy. We recognise this by using "tilde" rather than "hat" notation. For future reference, let h_f and h_g denote the bandwidths used for $\tilde{f}$ and $\tilde{g}$:

$$\tilde{f}(x) = \frac{1}{mh_f}\sum_{i=1}^{m} K\left(\frac{x - X_i}{h_f}\right), \quad \tilde{g}(y) = \frac{1}{nh_g}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_g}\right). \tag{2.5}$$

One-sided, asymptotic, $(1-\alpha)$-level confidence intervals for $R(p)$ are therefore given by $(\hat{R}(p) - z_\alpha\,\hat{\sigma}(1-p), 1)$ and $(0, \hat{R}(p) + z_\alpha\,\hat{\sigma}(1-p))$, where $z_\alpha > 0$ is the upper $1-\alpha$ point of the standard normal variable distribution. A two-sided confidence interval has of course endpoints $\hat{R}(p) \pm z_{\alpha/2}\,\hat{\sigma}(1-p)$. Here, $\hat{R}(p)$ is based on (2.1) and (2.2), but it does not necessarily use the same bandwidths as are used in (2.4). In our numerical examples, we estimate $\hat{R}(p)$ using the bandwidth proposal of P. Hall and Hyndman (2002).

2.3 Bootstrap confidence intervals

An alternative approach to constructing interval estimators is to approximate the distribution of

$$S = \bigl[G\{F^{-1}(t)\} - \widehat{G}\{\widehat{F}^{-1}(t)\}\bigr]/\hat{\sigma}\,,$$

using the bootstrap and Monte Carlo simulation. Specifically, draw data $\mathcal{X}^* = \{X_1^*, \ldots, X_m^*\}$ and $\mathcal{Y}^* = \{Y_1^*, \ldots, Y_n^*\}$ randomly, without replacement, from distributions with respective densities $\check{f}$ and $\check{g}$, where $\check{f}$ and $\check{g}$ are smoothed estimators of f and g and are computed from $\mathcal{X}$ and $\mathcal{Y}$, respectively. Compute the bootstrap versions, $\widehat{F}^*$, $\widehat{G}^*$, $\tilde{f}^*$ and $\tilde{g}^*$ say, of $\widehat{F}$, $\widehat{G}$, $\tilde{f}$ and $\tilde{g}$; let $\hat{\sigma}^*$ denote the version of $\hat{\sigma}$ at (2.4) that is obtained on replacing the latter estimators by their bootstrap forms; write $\check{F}$ and $\check{G}$ for the respective distribution functions corresponding to the densities $\check{f}$ and $\check{g}$; and put

$$S^* = \bigl[\check{G}\{\check{F}^{-1}(t)\} - \widehat{G}^*\{(\widehat{F}^*)^{-1}(t)\}\bigr]/\hat{\sigma}^*\,. \tag{2.6}$$

Then, the distribution of S^*, conditional on the original data $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$, is an approximation to the unconditional distribution of S.

In particular, we may compute $\hat{z}_\alpha = \hat{z}_\alpha(\mathcal{Z})$ as the solution of the equation $P(S^* \le \hat{z}_\alpha \mid \mathcal{Z}) = \alpha$, for $0 < \alpha < 1$, and take one-sided, $(1-\alpha)$-level confidence intervals for $R(p)$ to be $(\hat{R}(p) - \hat{z}_\alpha\,\hat{\sigma}(1-p), 1)$ and $(0, \hat{R}(p) - \hat{z}_{1-\alpha}\,\hat{\sigma}(1-p))$. These are of course percentile-t intervals.

We have introduced a third density estimator, $\check{f}$, rather than use one of the existing estimators $\hat{f}$ or $\tilde{f}$, since it is initially far from clear what the appropriate level of smoothing in the bootstrap resampling step should be. We may of course take $\check{f}$ to have the same form as $\hat{f}$ and $\tilde{f}$, but with a different choice of bandwidth. Likewise, we introduce $\check{g}$ rather than rely on $\hat{g}$ or $\tilde{g}$. Without choosing bandwidth appropriately the bootstrap algorithm may fail to adequately capture the effects of bias on the distribution of S. Indeed, we shall argue in Section 3.2 that it is necessary to choose the bandwidths for $\check{f}$ and $\check{g}$ much larger than those for $\hat{f}$, $\tilde{f}$, $\hat{g}$ and $\tilde{g}$. See Härdle and Bowman (1988) for an early account of the need to resample from a smoothed distribution when constructing confidence intervals where smoothed estimators are involved.

3 Coverage probabilities

3.1 Effect of bandwidth choice on asymptotic intervals

Let $\alpha \in (\frac{1}{2}, 1)$, and define z_α by $\Phi(z_\alpha) = 1 - \alpha$, where Φ denotes the standard normal distribution function. Also, let $y_1 = \widehat{F}^{-1}(t)$ and $y = F^{-1}(t)$. Since $\{\widehat{G}(y_1) - G(y)\}/\hat{\sigma}$ is asymptotically $\mathrm{N}(0, 1)$ then examples of asymptotic confidence intervals for $R(p) = 1 - G\{F^{-1}(1-p)\}$ are given by

$$\big(-\infty\ ,\ \hat{R}(p) + \hat{\sigma}\, z_\alpha\big] \qquad \big[\hat{R}(p) - \hat{\sigma}\, z_\alpha\ ,\ -\infty\big) \qquad \big[\hat{R}(p) - \hat{\sigma}\, z_{\alpha/2}\ ,\ \hat{R}(p) + \hat{\sigma}\, z_{\alpha/2}\big]\,. \quad (3.1)$$

The coverage probability of each converges to α as $n \to \infty$.

In familiar semiparametric problems, for example confidence intervals for a population mean, the three intervals at (3.1) would have coverage errors of sizes $n^{-1/2}$, $n^{-1/2}$ and n^{-1}, respectively. This reflects results in the theory of Edgeworth expansion; see, for example, Bhattacharya and Ghosh (1978) and P. Hall (1992, Chapter 2). In particular, the terms in $n^{-1/2}$ that dominate coverage-error formulae for one-sided intervals cancel, in the two-sided case, through a fortuitous parity property, and then second-order terms, of size n^{-1}, prevail.

In the present setting, however, such a simple account of coverage accuracy is prevented by the fact that $\hat{\sigma}$ involves a nonparametric component, depending critically on the bandwidths h_f and h_g used to construct $\tilde{f}$ and $\tilde{g}$ at (2.5), and employed to compute $\hat{\sigma}$. It can be shown that if h_f and h_g are chosen to be of conventional size, $n^{-1/5}$, appropriate for point estimation of f and g, then the coverage errors of each of the confidence intervals at (3.1) are of size $n^{-2/5}$, which falls short even of the level $n^{-1/2}$ that is available in the one-sided case in a classical setting.

That this is true even for the third, two-sided interval at (3.1) follows from the fact that the leading terms which introduce h_f and h_g to coverage-error formulae do not enjoy the classical parity property. As a result, errors of size $n^{-2/5}$ persist for each of the three intervals at (3.1). They compound, rather than cancel, in passing from one-sided to two-sided intervals. There are, of course, two other bandwidths, h_1 and h_2, used to construct $\widehat{F}$ and $\widehat{G}$ at (2.1). These, however, have only a minor impact, and can be chosen within a wide range without seriously affecting coverage error.

These results motivate a careful analysis of the impact that choosing h_f and h_g has on coverage accuracy. We shall show that it is optimal to select these bandwidths to be constant multiples of $m^{-1/3}$ and $n^{-1/3}$, respectively, and we shall suggest formulae for the constants. With this choice of the bandwidths, the coverage errors of the one-sided intervals at (3.1) are of size $n^{-1/2}$, reducing to $n^{-2/3}$ in the two-sided setting. Thus, accuracy in the one-sided case coincides with that in classical problems, while in the two-sided setting it is a little less than in the classical case, but still better than for one-sided intervals.

Next we describe our main theoretical results. Put $\rho = n/m$, $\kappa = \int K^2$, $\kappa_2 = \int u^2\, K(u)\, du$ and

$$a = \frac{\rho\,\{g(y)/f(y)\}^2\, t\,(1-t)}{G(y)\,\{1-G(y)\} + \rho\,\{g(y)/f(y)\}^2\, t\,(1-t)}\,. \tag{3.2}$$

Note that $0 < a < 1$. Define the even, quadratic polynomials

$$q_f(x) = \kappa\,\bigl(3 - a - a\,x^2\bigr) + 2\,K(0)\,\bigl(a + a\,x^2 - 1\bigr)\,, \quad q_g(x) = \kappa\,\bigl(a - 1 - a\,x^2\bigr)\,.$$

For $\psi = f$ or g, put $p_\psi(x) = ax\,\{q_\psi(x) - \kappa_2\, m h_\psi^3\, \psi''(y)\}/2\,\psi(y)$, an odd, cubic polynomial. Construct $\tilde f$ and $\tilde g$ using the kernel K and the respective bandwidths h_f and h_g. We shall show that, provided h_f and h_g are of respective sizes $m^{-1/3}$ and $n^{-1/3}$,

$$\begin{aligned} P\bigl[\{\widehat G(y_1) - G(y)\}/\hat\sigma \le x\bigr] &= \Phi(x) + n^{-1/2}\, p(x)\,\phi(x) \\ &\quad + \frac{p_f(x)}{2\,m h_f}\,\phi(x) - \frac{p_g(x)}{2\,n h_g}\,\phi(x) + o\bigl(n^{-2/3}\bigr)\,, \end{aligned} \tag{3.3}$$

where p denotes an even, quadratic polynomial, the coefficients of which do not depend on h_f or h_g, and which involve m and n only through the ratio ρ, remaining bounded as long as ρ is bounded away from zero and infinity. Regularity conditions for (3.3) will be given later in this section.

The implications of (3.3) are tied to parity properties of the polynomials p, p_f and p_g. Note that p is even, whereas p_f and p_g are odd, and so (3.3) implies that the two-sided confidence interval, $\mathcal{I} = [\hat R(p) - \hat\sigma\, z_{\alpha/2}, \hat R(p) + \hat\sigma\, z_{\alpha/2}]$, has coverage probability

$$P\bigl\{R(p) \in \mathcal{I}\bigr\} = \alpha + \left\{\frac{p_f(z_{\alpha/2})}{m h_f} - \frac{p_g(z_{\alpha/2})}{n h_g}\right\}\phi(z_{\alpha/2}) + o\bigl(n^{-2/3}\bigr)\,. \tag{3.4}$$

Depending on the values of a, α, m, n, $f^{(j)}(y)$ and $g^{(j)}(y)$ for $j = 1, 2$, it can be possible to choose h_f and h_g at (3.4) so that the quantity within braces there vanishes. This is not always feasible, however, and a simpler approach is to select h_f and h_g separately, to minimise absolute values of the respective terms within braces. Either approach produces bandwidths of size $m^{-1/3}$ and $n^{-1/3}$, respectively; the second approach results in the formulae $h_f = c_f\, m^{-1/3}$ and $h_g = c_g\, n^{-1/3}$, where

$$c_f = \theta\left|\frac{q_f(z_{\alpha/2})}{\kappa_2\, f''(y)}\right|^{1/3}, \quad c_g = \theta\left|\frac{q_g(z_{\alpha/2})}{\kappa_2\, g''(y)}\right|^{1/3}, \tag{3.5}$$

and $\theta = 1$ or $2^{-1/3}$ according as the ratio of the term within modulus signs is positive or negative. This approach to bandwidth choice is also appropriate when constructing the one-sided interval $\mathcal{J} = (-\infty, \hat R(p) + \hat\sigma\, z_\alpha]$. There the formulae at (3.5) remain valid, except that $z_{\alpha/2}$ should be replaced by z_α. For a constant bandwidth over the curve, we integrate the numerator and denominator of (3.5) over y.

By way of regularity conditions for (3.3) we require: (a) f and g have two continuous derivatives in a neighbourhood of y, (b) neither $f(y)$ nor $g(y)$ vanishes, (c) K is a con-

tinuous, symmetric, compactly supported density, (d) the bandwidths h_1 and h_2 used to construct $\widehat{F}$ and $\widehat{G}$, at (2.1), satisfy $h_j = o(n^{-7/12})$ and $nh_j/\log n \to \infty$ as $n \to \infty$, and (e) the sample-size ratio, ρ, is bounded away from zero and infinity as $n \to \infty$.

The regularity conditions (a)–(e) are mild, and it is clear that except possibly for (d) they are usually assured in practical settings. Moreover, (d) is guaranteed, in most cases of interest, if we choose h_1 and h_2 to be as small as possible subject to the jump discontinuities of $\widehat{F}_{\rm emp}$ and $\widehat{G}_{\rm emp}$ being "smoothed away" by $\widehat{F}$ and $\widehat{G}$, respectively, except in the extreme tails. This follows from the fact that, away from the tails, the maximum spacing of order statistics is of size $n^{-1}\log n$, and, across the entire distribution, is an order of magnitude larger provided that at least one tail of each of f and g descends to zero. Choosing a bandwidth that is just sufficiently large to smooth away jumps is the approach that is often followed in practice when using kernel methods to estimate a distribution function.

To implement the asymptotic intervals requires ten different smoothing parameters: very small bandwidths h_1 and h_2 for $\widehat{F}$ and $\widehat{G}$, at (2.1); bandwidths h_f and h_g for $\tilde{f}$ and $\tilde{g}$, at (2.5); bandwidths H_1 and H_2 for estimating $f''(y)$ and $g''(y)$ in (3.5); bandwidths H_f and H_g for estimating $f(y)$ and $g(y)$ in (3.2); and bandwidths H_F and H_G for estimating $F(y)$ and $G(y)$ in (3.2).

In our numerical examples we choose h_1 and h_2 to be 0.25 times the plug-in bandwidths for conditional distribution estimation (Lloyd and Yong, 1999); we choose h_f and h_g using (3.5); H_1 and H_2 are chosen to be optimal assuming f and g are normal, thus $H_1 = (4/7)^{1/9} m^{-1/9} s_x$ where s_x is the standard deviation of the $\mathcal{X}$, and H_2 is chosen analogously; we choose H_f and H_g using the Sheather-Jones (1991) plug-in rule; and an analogous plug-in rule for H_F and H_G. R code to carry out these calculations is available from Rob Hyndman.

3.2 Bootstrap intervals

A bootstrap version of (3.3) is readily developed. It has the form

$$P(S^* \le x \mid \mathcal{Z}) = \Phi(x) + n^{-1/2}\, p(x)\, \phi(x) + \frac{p_f(x)}{2\, m h_f}\, \phi(x) - \frac{p_g(x)}{2\, n h_g}\, \phi(x) + o_p\bigl(n^{-2/3}\bigr), \quad (3.6)$$

where S^* is as defined at (2.6). Recall that $\mathcal{Z}$ denotes the set of all data X_i and Y_j. The right-hand side of (3.6) is identical to its counterpart at (3.2), except that the remainder is now stochastic.

Results (3.6) follows from a close analogue of (3.3), in which the quantities f, g and their derivatives, appearing in formulae for p_f and p_g, are replaced by their counterparts involving $\check{f}$ and $\check{g}$. In order for (3.6) to follow from this particular expansion it is necessary that $\check{f}$ and $\check{g}$ involve sufficient smoothing to ensure that their second derivatives consistently estimate the second derivatives of f and g, respectively. In mathematical terms this means that the bandwidths used to construct $\tilde{f}$ and $\tilde{g}$ should converge to zero more slowly than

$m^{-1/5}$ and $n^{-1/5}$, respectively. Ideally, in the case of sufficiently smooth densities, the bandwidths should be of sizes $m^{-1/9}$ and $n^{-1/9}$. Thus, oversmoothing is required at this level; conventional bandwidth choices are of sizes $m^{-1/5}$ and $n^{-1/5}$. Without oversmoothing, the bootstrap method described in section 2.3 may not lead to improvements over the asymptotic approach. If sufficient oversmoothing is used, however, then it can be deduced from (3.3) and (3.6) that the bootstrap will produce one- and two-sided confidence intervals with coverage error equal to $o(n^{-2/3})$.

Therefore, choice of bandwidth for constructing the smoothed distribution estimators, $\widetilde{F}$ and $\widetilde{G}$, from which bootstrap sampling is done is a critical matter. For proper implementation the bootstrap technique requires six quite different, and all nonstandard, smoothing parameters: very small bandwidths h_1 and h_2 for $\widehat{F}$ and $\widehat{G}$, at (2.1); larger, but still smaller than usual, bandwidths h_f and h_g for $\tilde{f}$ and $\tilde{g}$, at (2.5); and quite large bandwidths for $\check{F}$ and $\check{G}$. This complexity makes the bootstrap approach particularly challenging, and relatively unattractive, to implement.

4 Examples

We compute the actual probability coverage of our confidence intervals using simulations on four examples having a range of density shapes. These are:

1. $F = \beta(2,3)$; $G = \beta(2,4)$;
2. $F = \beta(1.2,3)$; $G = \beta(1.2,2)$;
3. $F = \gamma(2)$; $G = \gamma(3)$;
4. $F = t(5)$; $G = 0.2(t(5)-1) + 0.8(t(5)+1)$.

where $\beta(a,b)$ denotes the Beta distribution with density $f(x) = \Gamma(a+b)\{\Gamma(a)\Gamma(b)\}^{-1}$ $x^{(a-1)}(1-x)^{(b-1)}$, $0 \le x \le 1$; $\gamma(a)$ denotes the Gamma distribution with density $x^{a-1}e^{-x}/\Gamma(a)$, $x > 0$; and $t(v)$ denotes the t distribution with v degrees of freedom.

For each example, we generated 1000 sets of data from F and G, each of size $m = n = 100$. Then the curve $\hat{R}(p)$ was computed with bandwidths chosen using the method of P. Hall and Hyndman (2003). Confidence intervals around the curve were computed using the method outlined in Section 2.2.

The proportion of times the confidence interval contained the true $R(p)$ for each p is plotted in Figure 1. Except in the extreme tails of the distributions, our approach is usually conservative.

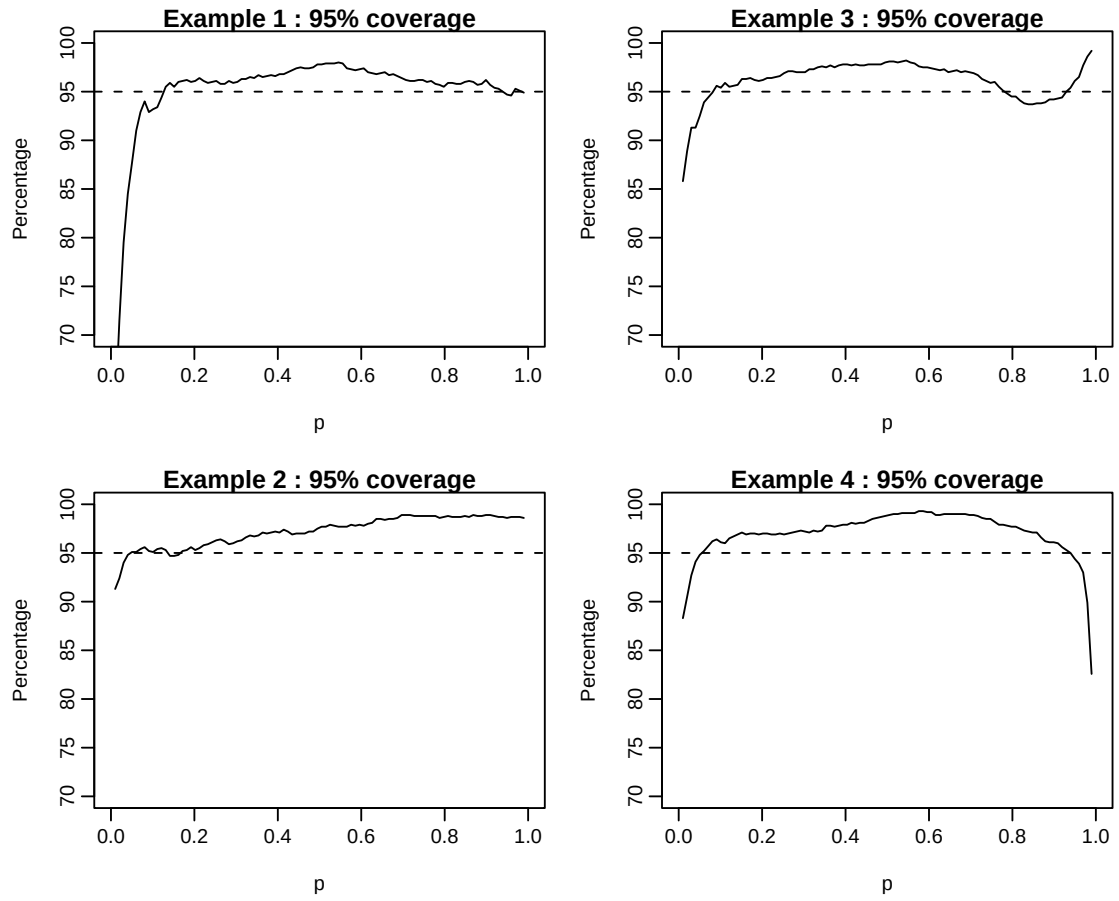

Figure 1: *Actual coverage of asymptotic confidence intervals computed as described in Section 2.2. In each example, the percentage is computed from 1000 simulated sets of data. Sample sizes were* $m = n = 100$*. Nominal coverage was 95%.*

5 Acknowledgements

Rob Hyndman thanks the Centre for Mathematics and its Applications for their hospitality while some of this work was being undertaken.

Appendix: Derivation of (3.3)

Recall that $y = F^{-1}(t)$ and $y_1 = \widehat{F}^{-1}(t)$, and define $B = \widehat{G}(y_1)\,\{1 - \widehat{G}(y_1)\}$, $B_0 = \rho\,\{g(y)/f(y)\}^2\,t\,(1-t)$, $B_1 = \rho\,\{\tilde{g}(y_1)/\tilde{f}(y_1)\}^2\,t\,(1-t)$, $A_j = B + B_j$, $A = B_0/A_0$, $\Delta_f = \{\tilde{f}(y_1) - f(y)\}/f(y)$, $\Delta_g = \{\tilde{g}(y_1) - g(y)\}/g(y)$ and

$$S_1 = A\,(\Delta_g - \Delta_f) + \tfrac{1}{2}\,A\,(3-A)\,\Delta_f^2 + \tfrac{1}{2}\,A\,(1-A)\,\Delta_g^2 + A\,(A-2)\,\Delta_f\,\Delta_g\,. \tag{A.1}$$

Then, omitting cubic and higher-order terms in expansions,

$$\begin{aligned} A_1^{1/2} &= \Big\{\, B + B_0\,\Big(1 + 2\,\Delta_g - 2\,\Delta_f + \Delta_g^2 + 3\,\Delta_f^2 - 4\,\Delta_f\,\Delta_g\Big)\,\Big\}^{1/2} \\ &= A_0^{1/2}\,\Big\{\, 1 + A\,\Big(2\,\Delta_g - 2\,\Delta_f + \Delta_g^2 + 3\,\Delta_f^2 - 4\,\Delta_f\,\Delta_g\Big)\,\Big\}^{1/2} = A_0^{1/2}\,(1 + S_1)\,. \end{aligned} \tag{A.2}$$

Put $b = G(y)\,\{1-G(y)\}$, $a_0 = b + B_0$, $U_1 = \{\widehat{G}(y_1) - G(y)\}/\hat{\sigma} = n^{1/2}\{\widehat{G}(y_1) - G(y)\}/A_1^{1/2}$ and $U_0 = n^{1/2}\{\widehat{G}(y_1) - G(y)\}/A_0^{1/2}$, and let Φ denote the standard normal distribution function. Note too that $a = B_0/a_0$ and $A_0 = a_0 + O_p(n^{-1/2})$. Write S_2 for the version of S_1 that is obtained if, in the definition at (A.1), A is replaced by a. Assume $h_f \sim c_f\,n^{-1/3}$ and $h_g \sim c_g\,n^{-1/3}$, for constants $c_f, c_g > 0$ still to be determined. Then $|\Delta_f| + |\Delta_g| = O_p(n^{-1/3})$. It follows that $S_1 - S_2 = O_p(n^{-(1/2)-(1/3)}) = O_p(n^{-5/6})$, and that the cubic terms that have been omitted from (A.2) are of order n^{-1}.

Let $\delta_f = \{\tilde{f}(y) - f(y)\}/f(y)$ and $\Delta = \{\tilde{g}(y) - g(y)\}/g(y)$. It can be shown that, with $\psi = f$ or g, $E(\Delta_\psi - \delta_\psi)^2 = O(n^{-1})$. Therefore, if we define

$$S_3 = a\,(\Delta_g - \Delta_f) + \tfrac{1}{2}\,a\,(3-a)\,\delta_f^2 + \tfrac{1}{2}\,a\,(1-a)\,\delta_g^2 + a\,(a-2)\,\delta_f\,\delta_g\,,$$

then $S_2 - S_3 = O_p(n^{-5/6})$. Combining this result with those derived in the previous paragraph we deduce that

$$U_1 = \frac{U_0}{1+S_1} + O_p\big(n^{-1}\big) = \frac{U_0}{1+S_2} + O_p\big(n^{-5/6}\big) = \frac{U_0}{1+S_3} + O_p\big(n^{-5/6}\big)\,.$$

Therefore, by the delta method,

$$P(U_1 \le x) = P\big\{U_0 \le x\,(1+S_3)\big\} + o\big(n^{-2/3}\big)\,. \tag{A.3}$$

Note that $U_0 \le x\,(1+S_3)$ is equivalent to $Z \le z$, where $Z = V - W$, $z = v - w$,

$$\begin{aligned} V &= (n/A_0)^{1/2}\,\{\widehat{G}(y_1) - G(y_1)\} + n^{1/2}\,\{G(y_1) - G(y)\}\,\big(A_0^{-1/2} - a_0^{-1/2}\big)\,, \\ W &= x\,\big\{a\,\Delta_g + \tfrac{1}{2}\,a\,(1-a)\,\delta_g^2\big\} + a\,(a-2)\,\delta_f\,\delta_g\big\}\,, \\ v &= x - (n/a_0)^{1/2}\,\{G(y_1) - G(y)\}\,, \quad w = x\,\big\{a\,\Delta_f - \tfrac{1}{2}\,a\,(3-a)\,\delta_f^2\big\}\,. \end{aligned} \tag{A.4}$$

We shall evaluate $P\{U_0 \le x\,(1+S_3)\}$ as $E\{P(Z \le z \mid \mathcal{X})\}$, where $\mathcal{X} = \{X_1, \ldots, X_m\}$, and so we seek initially an approximate formula for $P(Z \le z \mid \mathcal{X})$.

Noting that the random variables V and δ_g are asymptotically independent, that δ_f is a function of $\mathcal{X}$, and that δ_g is independent of $\mathcal{X}$, we show first that

$$\begin{aligned} E\big(e^{iuZ} \,\big|\, \mathcal{X}\big) = E\big\{e^{iuV}\,(1 - iuax\,\Delta_g) \,\big|\, \mathcal{X}\big\} \\ - \tfrac{1}{2}\left\{iuax\,(1-a) + (uax)^2\right\} E\big(\delta_g^2\big)\, E\big(e^{iuV} \,\big|\, \mathcal{X}\big) \\ - iua\,(a-2)\,\delta_f\, E(\delta_g)\, E\big(e^{iuV} \,\big|\, \mathcal{X}\big) + o_p\big(n^{-2/3}\big)\,, \end{aligned} \tag{A.5}$$

where $i = \sqrt{-1}$. Then, after Fourier inversion and some algebra, we obtain:

$$\begin{aligned} P(Z \le z \mid \mathcal{X}) = H(z) + ax\left\{E(\Delta_g \mid \mathcal{X}) + \tfrac{1}{2}\,(1-a)\,E\big(\delta_g^2\big) + (a-2)\,\delta_f\,E(\delta_g)\right\} H'(z) \\ + \tfrac{1}{2}\,(ax)^2\,E\big(\delta_g^2\big)\,H''(z) + o_p\big(n^{-2/3}\big)\,, \end{aligned} \tag{A.6}$$

where $H(z) = P(V \le z \mid \mathcal{X})$. Taking the expected value of both sides, and noting that $E(\Delta_g \mid \mathcal{X}) = E(\delta_g) + (y_1 - y)\,g'(y)\,g(y)^{-1} + o_p(n^{-2/3})$, gives:

$$\begin{aligned} P(Z \le z) = E\{H(v)\} + n^{-1/2}\,\pi_1(x) - ax\,E\{\Delta_f\,H'(v)\} \\ + \tfrac{1}{2}\,E\big(\delta_f^2\big)\left\{ax\,(3-a)\,\phi(x) + (ax)^2\,\phi'(x)\right\} \\ + ax\left\{E(\delta_g) + \tfrac{1}{2}\,(1-a)\,E\big(\delta_g^2\big)\right\}\phi(x) \\ - \tfrac{1}{2}\,(ax)^2\,E\big(\delta_g^2\big)\,\phi'(x) + o\big(n^{-2/3}\big), \end{aligned} \tag{A.7}$$

where, here and below, π_j denotes an even polynomial not depending on h_f or h_g. Now, $E\{H(v)\} = \Phi(x) + n^{-1/2}\,\pi_2(x) + o(n^{-2/3})$, and

$$E\{\Delta_f\,H'(v)\}/\phi(x) = E(\delta_f) + \frac{cx}{n^{1/2}} + \frac{K(0)}{m\,h_f\,f(y)}\left(1 - a + a\,x^2\right) + o\big(n^{-2/3}\big)\,,$$

where c depends on ρ but not on h_f or h_g. Substituting these results into (A.7), and expanding $E(\delta_f^2)$, $E(\delta_g)$ and $E(\delta_g^2)$ in the usual way, we deduce an expansion of $P\{U_0 \le x\,(1+S_3)\} = P(Z \le z)$ which is identical to the right-hand side of (3.3). The latter result now follows from (A.3).

The Fourier inversion which leads to (A.6) requires the small amount of smoothing implicit in the assumption, in (d), that $nh_j/\log n \to \infty$ for $j = 1, 2$. This removes "rounding error" terms, deriving from the lattice nature of the unsmoothed distribution functions $\widehat{F}_{\rm emp}$ and $\widehat{G}_{\rm emp}$. The condition $h_j = o(n^{-7/12})$, in (d), is just sufficient to eliminate bias effects of smoothing these distributions to $\widehat{F}$ and $\widehat{G}$. Bias effects are of size $n^{1/2}\,h_j^2$, which in view of (d) equals $o(n^{-2/3})$.

References

Altman, N. & Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *J. Statist. Plan. Infer.* **46**(2), 195–214.

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

Bhattacharya, R.N. & Ghosh, J.K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434–451.

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

De Uña-Álvarez, J., González-Manteiga, W. & Cadarso-Suárez, C. (2000). Kernel distribution function estimation under the Koziol-Green model. *J. Statist. Plan. Infer.* **87**(2), 199–219.

Goddard, M.J. & Hinberg, I. (1990). Receiver operating characteristic (ROC) curves and non-normal data: an empirical study. *Statist. Med.* **9**, 325–337.

Hall, P.G. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

Hall, P.G, & Hyndman, R.J. (2003) An improved method for bandwidth selection when estimating ROC curves. *Statistics and Probability Letters*, in press.

Härdle, W. & Bowman, A.W. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83**, 102–110.

Hsieh, F.S. & Turnbull, B.W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Statist.* **24**, 25–40.

Li, G., Tiwari, R.C. & Wells, M.T. (1999). Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. *Biometrika* **86**, 487–502.

Lloyd, C.J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *J. Amer. Statist. Assoc.* **93**, 1356–1364.

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

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

Peng, L. & Zhou, X.-H. (2002). Local linear smoothing of receiver operator characteristic (ROC) curves. *J. Statist. Plan. Infer.*, in press.

Polansky, A.M. & Baker, E.R. (2000). Multistage plug-in bandwidth selection for kernel distribution function estimates. *J. Statist. Comput. Simul.* **65**, 63–80.

Reiss, R.-D. (1981). Nonparametric estimation of smooth distribution functions. *Scand. J. Statist.* **8**, 116–119.

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

Sheather, S.J. & Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. Royal Stat. Soc.* **B 53**, 683-690.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9**, 1187–1195.

Zhou, X.H. & Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statist. Med.* **21**, 2045–2055.

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

Zou, K.H. & Hall, W.J. (2000). Two transformation models for estimating an ROC curve derived from continuous data. *J. Appl. Statist.* **27**, 621–631.

Zweig, M.H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots—a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**, 561–577.