Smoothing Parameter Selection for Power Optimality in Testing of Regression Curves

K.B. Kulasekera*

and

J. Wang

Department of Mathematical Sciences

Clemson University

Clemson, SC 29634

Abstract

We consider selection of smoothing parameters to obtain optimal power in tests of regression curves. Three tests are examined and empirical smoothing parameters are proposed to maximize the power in each test. We also

*K.B. Kulasekera is Associate Professor, and J. Wang is Graduate Student, Department of Mathematical Sciences, Clemson University, Clemson, SC 29634. This research is supported by grant 1R15 GM51106-01 from NIH. The authors wish to thank an associate editor and two referees for helpful comments.

show that the data based smoothing parameters converge to the optimal smoothing parameters as sample sizes get larger. A simulation study is conducted for various classes of alternatives showing the effectiveness of the the proposed procedures.

Key Words: Kernel estimator; Design variables; Nonparametric test.

1. INTRODUCTION

The comparison of regression functions for two or more populations is a widely discussed issue. Nonparametric methods for such comparisons are used in many instances due to difficulties such as model specification that are required with parametric procedures. Use of nonparametric methods for inferences on regression curves involve smoothing parameters that are specified by the experimenter. The performance of a test with regard to its power can be significantly affected by the smoothing parameter. In this article, we are interested in examining the effect of smoothing parameters on the power of nonparametric tests of the hypothesis

$$H_0 : f = g \quad vs \quad H_1 : f \neq g \tag{1}$$

when we observe data in the form $\{(Y_{1i}, x_{1i}), i = 1, ..., n\}$ and $\{(Y_{2j}, x_{2j}), j = 1, ..., m\}$ with

$$\begin{aligned} Y_{1i} &= f(x_{1i}) + \epsilon_i, \quad i = 1, ..., n, \\ Y_{2j} &= g(x_{2j}) + \eta_j, \quad j = 1, ..., m. \end{aligned} \tag{2}$$

Here $\epsilon_i, i = 1, ..., n,$ and $\eta_j, j = 1, ..., m,$ are independent, random errors for the two groups. It is further assumed that within each group the errors are identically distributed with mean zero, but the distributions of the ϵ's and η's may be different. Also, let $Var(\epsilon_1) = \sigma^2$ and $Var(\eta_1) = \tau^2$. In this testing problem, the two functions f and g are specified only up to some smoothness conditions. The domain of the covariate (x) measurements, is taken as $[0, 1]$. The covariate values (design points) for the two samples may be different, and the sample sizes can be unequal.

Testing hypotheses of the above type in a nonparametric setting has been discussed by Hall and Hart (1990), Härdle and Marron (1990), King, Hart and Wehrly (1991) and Kulasekera (1995). Hall and Hart (1990) discuss a bootstrap procedure for testing hypotheses (1) under the assumptions that $m = n$ and that there are common design points (i.e., $x_{1i} = x_{2i}$ for all i). The test involves a smoothing parameter p, which influences the power against many alternatives. The behavior of the power with respect to p does not satisfy any monotonicity properties as shown by Hall and Hart (1990). King et al. (1991) propose two tests, both for common design points, one for normal errors and the other for nonnormal errors. These tests are based on a kernel estimator of the unknown difference function $f - g$. The bandwidth for the kernel estimators have been shown to have significant effect on power for all sample sizes and they also observe the non-monotonic behavior of the power against many alternatives. Kulasekera (1995) proposed three tests that are all free of the common design assumption and the equal sample size assumption. All three tests involve linear estimates, such as kernel estimates of the unknown regression functions, which influence the power through the bandwidth of the linear smoother.

Although the significance of selecting a suitable smoothing parameter in conducting a test has been noted, no attempt has been made to select a data based value for the appropriate smoothing parameter that will preserve the size (at least asymptotically) while maximizing the power against a given alternative. King et al. (1991) note in their simulation study that the bandwidth which gives good power appears to be slightly larger than the asymptotically optimal bandwidth, which minimizes the mean squared error in estimation of the regression function. In contrast, all tests in Kulasekera (1995) require bandwidths of type $o(n^{-1/4})$ for asymptotic validity of the null distribution of the tests.

We address the smoothing parameter selection issue in a test of type (1) using tests by King et al. (1991), Hall and Hart (1990) and Kulasekera (1995) (referred to as KHW, HH and KU hereafter). We observe that the bandwidth which maximize the power in a size α test in KHW depends not only on the difference $f-g$, but on the second derivative of this function as well. In particular, we use the fact that for moderate to large sample sizes, the power of the two tests in KHW typically depends on a parameter

$$\Delta = \left[E_{H_a}\left(\| \widehat{f-g} \|\right) - E_{H_0}\left(\| \widehat{f-g} \|\right)\right] / \sigma(T),$$

where $\| f-g \| = \int_0^1 (f-g)^2$, $\widehat{f-g}$ is a suitable estimator of $f-g$, and $\sigma(T)$ is the null standard deviation of the test statistic T being used. Here E_{H_a} indicates the expectation under the alternative hypothesis and E_{H_0} indicates the expectation under the null hypothesis. It is well known that in second-order kernel-type estimators of $f-g$ with a smoothing parameter h,

$$E(\widehat{f-g}) \approx (f-g) + Ch^2(f-g)'',$$

where C is a constant. Hence for such estimators

$$E_{H_a}\left(\| \widehat{f-g} \|\right) - E_{H_0}\left(\| \widehat{f-g} \|\right) \approx \| f-g \| + 2Ch^2 \int_0^1 (f-g)(f-g)''.$$

Thus, for difference functions $f-g$ with $\int_0^1 (f-g)(f-g)'' > 0$, the power tends to increase with h for moderate sample sizes if $\sigma(T)$ is non-increasing in h. This is the case in KHW and this partially explains the behavior of the power function with respect to the smoothing parameter as seen in the simulation results of the power given in their paper. In KU, the test $\mathcal{W}_3$ exhibit the same characteristics. For the tests $\mathcal{W}_1$ and $\mathcal{W}_2$ in KU, the power depends mainly on a parameter $\Delta = \int_0^1 \left[\int_0^t (f-g)(u)du\right]^2 dt$. In HH, although the role of the smoothing parameter p is somewhat different from a bandwidth of a kernel estimator, the behavior of the power with respect to p is similar to the tests $\mathcal{W}_1$ and $\mathcal{W}_2$.

We propose to select the smoothing parameter in each test (KHW, HH, and KU) to maximize estimators of the corresponding function of Δ, thus optimizing the power of a test for a given set of data while keeping the asymptotic size to be at a desired level α. It will also be shown that the selection procedures to be introduced shortly are asymptotically optimal in the sense that the ratio of the smoothing parameter maximizing the exact Δ to the smoothing parameter selected by the data based method tends to 1 in probability as the sample sizes increase. However, the data based smoothing parameters can inflate the size of a test and the discrepency from the actual size depends largely on the variability of the responses and the sample size. The differences seem to diminish rapidly with increasing sample size as shown by simulations of all the tests. Smoothing parameter selection (specifically the number of terms) in Neyman type smooth tests for goodness-of-fit is somewhat related to the issues discussed here. A discussion relative to that can be found in Ledwina (1994) and the references therein.

We organize the paper in the following fashion. In section 2 we give notation and describe the data based smoothing parameter selection criteria followed by the optimality results for tests KHW, HH and KU. In Section 3 we give an empirical study and two real data examples involving Drilling Data (Hand, Daly, Lunn, McConway and Ostrowski, 1993) and Ethanol data (Cleveland, 1994), followed by a discussion in Section 4. Finally, in the Appendix we sketch the proofs.

2. DATA BASED SMOOTHING PARAMETER FOR EACH TEST

In this section we give the data based smoothing parameter selection procedure for each test: KHW, HH and KU. We also give the asymptotic behavior of the selection rule and discuss the effect of such data based smoothing parameter values on the size of a test. The covariate values are considered to be fixed in all cases, although the extension of the procedures to random the covariate case is not difficult under mild conditions on the conditional distributions. Also, the covariate values are assumed to be ordered without loss of generality throughout the article.

We may need some of the following assumptions at various points.

$\mathcal{A}_1$: The kernel function K is a known probability density function symmetric about 0, it has a compact support $[-1, 1]$, $\kappa_1 = \int_{-1}^{1} u^2 K(u) du < \infty$ and $\kappa_2 = \int_{-1}^{1} K^2(u) du$

$\mathcal{A}_2$**:** The functions f and g are twice continuously differentiable.

$\mathcal{A}_3$**:** The sequences of design points $\{x_{1i},\ i = 1, ..., n\}$ and $\{x_{2j},\ j = 1, ..., m\}$ become dense in $[0, 1]$ as $\min(m, n) \to \infty$.

$\mathcal{A}_4$**:** The bandwidth sequence h used in estimating f and g satisfies $\min(m, n)h \to \infty$ and $\min(m, n)h^4 \to 0$ as $\min(m, n) \to \infty$.

$\mathcal{A}_5$**:** The bandwidth sequence h used in estimating f and g satisfies $\min(m, n)h \to \infty$ and $\min(m, n)h^5 \to 0$ as $\min(m, n) \to \infty$.

In all situations, we take $[0, 1]$ as the domain of the covariate and we use linear smoothers such as kernel estimators (Eubank 1988; Müller 1987) for estimation of unknown regression functions. For example, in estimating f using the sample $(Y_{1i}, x_{1i}), i = 1, ..., n$ we define a kernel estimator as

$$\hat{f}(x) = \sum_{i=1}^{n} Y_{1i} w_{1i}(x), \tag{3}$$

where $w_{1i}(x) = \int_{s_{1,i-1}}^{s_{1,i}} \left(K\left((x-u)/h\right)/h\right) du,\ s_{1,0} = 0, s_{1,i} = (x_{1i} + x_{1\,i+1})/2$ and $s_{1,n} = 1$. Here, h is the bandwidth of the kernel estimator. The bandwidth is typically a function of the sample size. We assume, for a sample size n, that the bandwidth h will fall in an interval $H_n = [\theta n^{-\delta}, \beta n^{-\delta}]$ for arbitrary, but fixed θ and β, where $0 < \theta < \beta$ and $\frac{1}{4} < \delta < 1$.

2.1 KHW PROCEDURE

KHW assume that $m = n$ and $x_{1i} = x_{2i}, i = 1, .., n$. The test statistic for testing hypotheses (1) is

$$T = \frac{D'W'WD/n}{D'G'GD}$$

where D is the vector of $D_i = Y_{1i} - Y_{2i}, i = 1, ..., n$, W is the matrix $w_{1i}(x_{1j})$ of the kernel estimator of $\mu = f - g$ at each design point x_{1j} with a bandwidth h, ($\hat{\mu}(x_{1j}) = \sum_{i=1}^{n} D_i w_{1i}(x_{1j})$), and $G'G$ is the matrix of the quadratic form for the estimator s^2 of $Var(D_i) = \gamma^2 = \sigma^2 + \tau^2$ (Hall, Kay, and Titterington 1990). KHW use

$$s^2 = \frac{1}{2n} \sum_{i=1}^{n-1} (D_{i+1} - D_i)^2 . \tag{4}$$

Under a normal error structure, the exact null distribution of T for any sample size was derived by KHW. This null distribution of course depends on the particular bandwidth being used. When the error structure is not normal, for large n, the distribution of the properly normalized version of T can be approximated by a standard normal distribution. Namely, when $n \to \infty, h \to 0, nh \to \infty$,

$$\frac{T - E(T)}{\sqrt{\frac{1}{n^2 h} B \gamma}} \overset{D}{\approx} N(0, 1), \tag{5}$$

where $B = 2 \int_{-2}^{2} \left(\int_{-1}^{1} K(z) K(z + y) dz \right)^2 dy$, and $N(0, 1)$ is a standard normal variate (KHW, 1991). Let $T_1 = D'W'WD/n$. The statistic T is a scaled version of T_1. Under the above conditions on n and h it is not difficult to verify (King, 1988) that under the null hypothesis,

$$E(T_1) \approx \frac{\kappa_2 \gamma^2}{nh},$$

while under the alternative hypotheses

$$E(T_1) \approx \frac{\kappa_2 \gamma^2}{nh} + \frac{1}{n} \sum_{i=1}^{n} [\mu(x_{1i}) + \kappa_1 h^2 \ \mu''(x_{1i})]^2.$$

Thus, ignoring higher order terms, we observe that the power of the test increases with

$$\Delta_1(h) = \sqrt{n^2 h}\left[\frac{1}{n}\sum_{i=1}^{n}[\mu^2(x_{1i}) + 2\kappa_1 h^2 \mu(x_{1i})\mu''(x_{1i})]\right] \tag{6}$$

regardless of the error distributions when n is assumed large. We use this fact in determining a bandwidth for the test with optimal power. Suppose $h^* = \inf\{\arg(\sup_{h\in H_n} \Delta_1(h))\}$. We propose to obtain an estimator $\hat{h} \in H_n$ of h^* by maximizing

$$\hat{\Delta}_1(h) = \sqrt{n^2 h}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_i^2 - \frac{\kappa_2 s^2}{nh}\right].$$

over H_n. Here $\hat{\mu}_i = \hat{\mu}_(x_{1i})$ is the kernel estimator of $\mu(x_{1i})$ as defined above. In particular, define $\hat{h} = \inf\{\arg(\sup_{h\in H_n} \hat{\Delta}_1(h))\}$. The motivation for this is that since $\hat{\Delta}_1$ is a nearly unbiased estimator of Δ_1, the continuity of the functions Δ_1 and $\hat{\Delta}_1$ in h causes the maximizing values of $h \in H_n$ for the two functions to be close to one another. The following theorem is a precise statement of this result.

Theorem 2.1 *Let assumptions $\mathcal{A}_1 - \mathcal{A}_4$ be satisfied. Then $\frac{\hat{h}}{h^*} \to 1$ in probability as $n \to \infty$.*

Now, suppose $\pi(h) = P[T > c_\alpha(h)]$ where $c_\alpha(h)$ is the critical value for a size α test for the design $x_{1i}, i = 1, ..., n$ with a normal error structure and $\mu(x) = \nu(x)/n^{1/2-\delta/4}$ where $\int_0^1 \nu^2(x)dx \neq 0$, and let $h^{**} \in H_n$ maximize $\pi(h)$. Then we have the following result giving the asymptotic optimality of the procedure.

Theorem 2.2 *Let assumptions $\mathcal{A}_1 - \mathcal{A}_4$ be satisfied. Then, $\frac{\hat{h}}{h^{**}} \to 1$ in probability as $n \to \infty$.*

Proofs of these Theorems are given in the Appendix.

Remark 2.1 Under a normal error structure, one finds the exact critical point $c_\alpha(h)$ for a level α test for a given design and a given fixed bandwidth h. Thus, when we select the bandwidth $\hat{h}$ based on data, and use the critical value $c_\alpha(\hat{h})$, the test may no longer have the desired size α in a finite sample. On the other hand, under H_0, $h^* = \theta n^{-\delta}$ and therefore $c_\alpha(\theta n^{-\delta})$ will be an asymptotic critical value for a test using $\hat{h}$. If the error structure is not normal, then one uses the asymptotic result (5) to obtain the critical points, and the asymptotic optimality of $\hat{h}$ follows from Theorem 2.1 itself. In this case the size of the test could be affected from using the asymptotic distribution as well as the stochastic bandwidth.

Remark 2.2 One can use a bootstrap procedure similar to the procedure in HH in getting critical values. Set $D_i^0 = D_i - \bar{D}$, where $\bar{D} = \sum_{i=1}^n D_i/n$ and re-sample from $\{D_i^0, i = 1, ..., n\}$. For the jth sample, $j = 1, ..., M$, one calculates an $\hat{h}_j$ and a corresponding $T_j(\hat{h}_j)$. The upper αth percentile of these T_j's will be the bootstrap critical point. Arguments similar to those in HH can be used in showing that the the critical points obtained by this method converge to the exact critical points.

2.2 HH PROCEDURE

Although HH consider the unequal designs case, their main focus is on the case where designs and sample sizes are the same. In this article we only consider the equal designs case for smoothing parameter selection since the modification HH suggest for unequal designs case may not be very effective for many alternatives (Kulasekera, 1995).

Using the same notation as in KHW, the test statistic in HH, which is a function of a smoothing parameter p, $0 < p < 1$, can be written as

$$S(p) = \frac{1}{n^2} \sum_{i=1}^{n-1} \left[\sum_{j=i}^{i+[np]} D_j / s \right]^2$$

where s^2 is an estimate of the variance of D_i. Here also we may use the same estimator s^2 given in (4) and $D_i = D_{i-n}, i > n$. In HH, the smoothing parameter p is kept fixed in contrast to other nonparametric techniques. The excellent power properties are a consequence of this. HH show that, under the null hypothesis, S converges in distribution to

$$S_0(p) = \int_0^1 (W(t+p) - W(t))^2 \, dt,$$

where W is the standard Weiner process on $[0,1]$ and $W(t) = W(t-1), t > 1$. Asymptotic critical points can be obtained by simulating S_0. HH propose a bootstrap procedure to obtain accurate approximations to critical points.

Assume that $p \in P = [c, d]$ for given $0 < c < d < 1$ and assume that the design density $r(x) = 1, 0 < x < 1$. Now, from the distributional results in HH, the power of the test for a given alternative $f(x) - g(x) = \mu(x)$ increases with $n^{-2} \sum_{i=1}^{n-1} \left[\sum_{j=i}^{i+[np]} \mu(x_{1j}) \right]^2$ which diverges to ∞ at a rate of n. We define

$$\Delta_2(p) = \frac{1}{n^3} \sum_{i=1}^{n-1} \left[\sum_{j=i}^{i+[np]} \mu_j \right]^2, \tag{7}$$

where $\mu_i = \mu(x_{1i})$, $\mu_i = \mu_{i-n}, i > n$, and maximize Δ_2 in P. Suppose that the maximum is achieved at $p^* \in P$, where $p^* = \inf\{\arg(\sup_{p \in P} \Delta_2(p))\}$. As in the previous section, we obtain an estimate $\hat{p}$ of p^* by maximizing the function $\hat{\Delta}_2(p)$, where

$$\hat{\Delta}_2(p) = \frac{1}{n^3}\sum_{i=1}^{n-1}\left[\sum_{j=i}^{i+[np]}\hat{\mu}_j\right]^2, \tag{8}$$

and $\hat{p} = \inf\{\arg(\sup_{p\in P}\hat{\Delta}_2(p))\}$. Here, $\hat{\mu}_i$ is a kernel estimator of $\mu(x_{1i})$ with an appropriate bandwidth h. For simplicity, we may take a deterministic bandwidth $h = n^{-1/5}$, which satisfy the asymptotic optimality in the mean squared error sense. By definition, the two functions Δ and $\hat{\Delta}$ have discontinuities at $p = i/n$ for integers i. However, we consider continuous versions of these functions by linear interpolation and use the same notation for simplicity.

The sequence $\hat{p}$ is also asymptotically optimal as given by the following theorem.

Theorem 2.3 *Let assumptions $\mathcal{A}_1 - \mathcal{A}_3$ be satisfied. Then $\frac{\hat{p}}{p^*} \to 1$ in probability as $n \to \infty$.*

The parameter Δ_2 is defined for a difference function $\mu(x) = (f-g)(x)$. One can view Δ_2 as a function of $(f-g)(x_i)$ with n^2 instead of n^3 as the divisor in (7) and consider local alternatives $(f-g)(x_i) = \mu(x_i)/\sqrt{n}$, so that Δ_2 converges to

$$\Delta_2^* = \int_0^1 \left(\int_t^{t+p}\mu(x)dx\right)^2 dt$$

as $n \to \infty$. Then, the corresponding sample version

$$\hat{\Delta}_2(p) = \frac{1}{n^2}\sum_{i=1}^{n-1}\left[\sum_{j=i}^{i+[np]}\hat{\mu}_j/\sqrt{n}\right]^2, \tag{9}$$

can be shown to converge in distribution to $\Delta_2^* + \psi(p)$, where

$$\psi(p) = 2\int_0^1 \left(\int_t^{t+p} \mu(x)dx\right) (W(t+p) - W(t))\, dt.$$

This is due to large covariances of the kernel type estimators of $\mu(x)/\sqrt{n}$ at different points. Therefore, for local alternatives converging at a $1/\sqrt{n}$ rate to the null hypothesis, the asymptotic properties given in Theorem 2.3 do not hold. However, one can examine the asymptotic behavior of power when a data based p is used in the test in the following manner.

Let $\pi_0(p) = P[S_0(p) + \Delta_2^* > c_\alpha(p)]$ where $c_\alpha(p)$ is such that $P[S_0(p) > c_\alpha(p)] = \alpha$. Here we indicate the dependence of S_0 on p explicitly. Our goal is to maximize the asymptotic power $\pi_0(p)$ with respect to p under the appropriate local alternative. Since this is not easily done, we obtain $p^{**} \in P$ that maximizes Δ_2^* and define $\pi^* = P[S_0(\frac{1}{2}) + \Delta_2^*(p^{**}) > c_\alpha(\frac{1}{2})]$. This can be thought of as a conservative approximation for $\sup_{p\in P} \pi_0(p)$ since $E(S_0(p)) = 2p(1-p)$ and $E(S_0^k(p)) = O((p(1-p))^k), k > 1$. An estimator of π^* can be defined as

$$\hat{\pi}^* = P\left[S_{n0}\left(\frac{1}{2}\right) + \sup_{p\in P} \hat{\Delta}_2 > c_\alpha\left(\frac{1}{2}\right)\right],$$

where

$$S_{n0} = \frac{1}{n^2}\sum_{i=1}^{n-1}\left[\sum_{j=i}^{i+[n/2]} (\eta_j + \epsilon_j)/s\right]^2,$$

and $\hat{\Delta}_2$ is now defined by (9). Then by the decomposition of the estimator $\hat{\mu}_i/\sqrt{n}$ into terms of $\mu_i/\sqrt{n}$ and the error terms $\eta_i + \epsilon_i$, we can show using results for weak convergence on metric spaces (Billingsley 1968),

$$S_{n0}(1/2) + \sup_{p\in P} \hat{\Delta}_2 \xrightarrow{\mathcal{D}} S_0(1/2) + \sup_{p\in P} \{\Delta_2^* + \psi(p)\}. \tag{10}$$

Thus, $\hat{\pi}^* \to P[S_0\left(\frac{1}{2}\right) + \sup_{p \in P}\{\Delta_2^* + \psi(p)\} > c_\alpha\left(\frac{1}{2}\right)]$. That is, by maximizing $\hat{\Delta}_2$, we asymptotically maximize $P[S_0\left(\frac{1}{2}\right) + \{\Delta_2^* + \psi(p)\} > c_\alpha\left(\frac{1}{2}\right)]$ rather than π^*. Therefore, the actual probability that is maximized depends on the behavior of the random variable $\psi(p)$. In fact, using standard results for Weiner processes, we can show that $\psi(p)$ has a normal distribution with mean 0 and a variance

$$\sigma^2(p) = 2\left[I_1 + I_2 + I_3\right],$$

where

$$I_1 = \int_0^{1-p}\int_0^t \Lambda(t,p)\Lambda(s,p)(s-t+p)dsdt,$$

$$I_2 = \int_{1-p}^1\int_{1-p}^t \Lambda(t,p)\Lambda(s,p)\left\{2s+p-1-\min\{s,t+p-1\}\right\}dsdt,$$

and

$$I_3 = \int_0^{1-p}\int_{1-p}^1 \Lambda(t,p)\Lambda(s,p)\left\{t-\min\{t,s+p\}\right\}dsdt.$$

Here $\Lambda(s,p) = \int_s^{s+p}\mu(x)dx$ and $\mu_x = \mu_{x-1}, x > 1$. One can then examine the behavior of $\sup_{p \in P}\{\Delta_2^* + \sigma_p N(0,1)\}$, where $N(0,1)$ is a standard normal variable, in order to obtain $\pi_0^* = \sup_{p \in P} P[S_0\left(\frac{1}{2}\right) + \{\Delta_2^* + \psi(p)\} > c_\alpha\left(\frac{1}{2}\right)]$. However, this is tedious, as $\sigma(p)$ depends on the unknown μ in a very complex fashion.

Another approach is to use the Cauchy-Schwartz inequality on $\psi(p)$ to obtain

$$\sup_{p \in P} P\left[S_0(1/2) + \{\Delta_2^* + \psi(p)\} > c_\alpha(1/2)\right] \geq \sup_{p \in P} P\left[S_0(1/2) + B(p) > c_\alpha(1/2)\right],$$

where $B(p) = \Delta_2^* - 2\left[S_0(p)\Delta_2^*\right]^{1/2}$. Since the kth moment of $S_0(p)$ is of order $[p(1-p)]^k$, one could then argue that a reasonable "lower bound" on the right-hand side of the above inequality is $\sup_{p \in P} P\left[S_0(1/2) + B_0(p) > c_\alpha(1/2)\right]$, where $B_0(p) = \Delta_2^* - 2\left[S_0(1/2)\Delta_2^*\right]^{1/2}$. Now,

$$\sup_{p \in P} P\left[S_0(1/2) + B_0(p) > c_\alpha(1/2)\right] = \sup_{p \in P}\left[I(p) + II(p)\right] \tag{11}$$

where

$$I(p) = P\left[S_0^{1/2}(1/2) \leq \sqrt{\Delta_2^*} - \sqrt{c_\alpha\,(1/2)}\right]$$

and

$$II(p) = P\left[S_0^{1/2}(1/2) \geq \sqrt{\Delta_2^*} + \sqrt{c_\alpha\,(1/2)}\right]$$

The quantity $II(p)$ is smaller than α. Therefore, as an approximate lower bound on π_0^* we may use

$$\sup_{p \in P} I(p) = P\left[S_0^{1/2}(1/2) \leq \sqrt{\Delta_2^*(p^{**})} - \sqrt{c_\alpha\,(1/2)}\right].$$

This can be much more conservative than the approximation π^* to the optimal asymptotic power π_0, depending on the value of Δ_2^*.

Remark 2.3 The HH test is an asymptotic test and the critical points for this asymptotic test are chosen by bootstrapping the centralized differences. The method we propose above is for selecting an appropriate p for the asymptotic test. In implementing the bootstrap method to obtain a critical value, we use $\hat{p}$ in all bootstrap samples. The consequences of the data based choice of p on the exact size of the test are similar to those for KWH test.

2.3 KU Procedure

The three tests of KU for the above hypotheses, are all based on quasi residuals $e_{1i} = Y_{1i} - \hat{g}(x_{1i}), i = 1, ..., n$ and $e_{2j} = Y_{2j} - \hat{f}(x_{2j}), j = 1, ..., m$, where $\hat{f}$ and $\hat{g}$ are suitable estimators of f and g, such as kernel estimators given in (3), using the corresponding samples and bandwidths h_1 and h_2 respectively. In this article, we only discuss the tests $\mathcal{W}_1$ and $\mathcal{W}_2$ because of their similarities to KHW and HH. The tests $\mathcal{W}_1$ and $\mathcal{W}_2$ are constructed through partial sums

$$\mu_n(t) = \sum_{i=1}^{[nt]} \frac{e_{1i}}{\sqrt{n}} \tag{12}$$

and

$$\mu_m(t) = \sum_{j=1}^{[mt]} \frac{e_{2j}}{\sqrt{m}}, \tag{13}$$

for $0 < t < 1$. These partial sums, which under the null hypothesis act as partial sums of mean zero random errors, can be used to construct statistics

$$T_1 = \sum_{k=1}^{n} \frac{\mu_n^2(k/n)}{n s_n^2}, \quad T_1' = \sum_{k=1}^{m} \frac{\mu_m^2(k/m)}{m s_m^2}. \tag{14}$$

for test $\mathcal{W}_1$. Here, s_n^2 and s_m^2 are consistent estimators of σ^2 and τ^2 (e.g., difference type estimators defined in (4)). Then a suitable function of T_1 and T_1' such as $\min(T_1, T_1')$ is used as the test statistic. For test $\mathcal{W}_2$, the statistics

$$T_2 = \int_0^1 \frac{\mu_n^2(t-)}{s_n^3} d\mu_n(t), \quad T_2' = \int_0^1 \frac{\mu_m^2(t-)}{s_m^3} d\mu_m(t) \tag{15}$$

where $\mu_n(t-)$ and $\mu_m(t-)$ are left-hand limits at t, are combined in a suitable manner to get a test. In all cases, the statistics $T_i, T_i', i = 1, 2$ diverge to ∞ at a rate of n under the alternative hypothesis. For example, with a difference $\mu = f - g$, we observe that

$$E(T_1) \approx \frac{1}{n^2} \sum_{j=1}^{n} \left[\sum_{i=1}^{j} \mu_{1i} \right]^2 + \frac{\kappa_2 \tau^2}{m h_2}$$

where

$$\mu_{1i} = f(x_{1i}) - \sum_{l=1}^{m} g(x_{2l}) w_{2l}(x_{1i}),$$

and

$$E(T_1') \approx \frac{1}{m^2} \sum_{j=1}^{m} \left[\sum_{l=1}^{j} \mu_{2l} \right]^2 + \frac{\kappa_2 \sigma^2}{n h_1}$$

with

$$\mu_{2l} = g(x_{2l}) - \sum_{i=1}^{n} f(x_{1i}) w_{1i}(x_{2l}).$$

Here, h_1 and h_2 and $w_{1i}, i = 1, ..., n$, $w_{2j}, j = 1, ..., m$ are bandwidths and weights in the linear smoothers that estimate f and g using Y_{1i}'s and Y_{2j}'s respectively. In both cases, the divergence of the test statistic under the alternative hypothesis depends on the first term on the right-hand side of each expectation. In view of these approximations, we propose to choose the bandwidths h_1 and h_2 so as to maximize

$$\Delta_3(h_1, h_2) = \frac{1}{n^3} \sum_{i=1}^{n} \left[\sum_{j=1}^{i} \mu_j \right]^2 \tag{16}$$

over $H_{n,m} = H_n \times H_m$, where $H_m = [\theta m^{-\delta}, \beta m^{-\delta}]$ and

$$\mu_j = \mu(x_{1j}) = \frac{1}{2} \left\{ \sum_{i=1}^{n} [f(x_{1i}) - g(x_{1i})] w_{1i}(x_{1j}) + \sum_{l=1}^{m} [f(x_{2l}) - g(x_{2l})] w_{2l}(x_{1j}) \right\}.$$

In this formulation the weights w_{1i}'s are functions of $x_{1i}, i = 1, ...n$ and h_1 and w_{2l}'s are functions of $x_{2j}, j = 1, ...m$ and h_2. We also require the weights to sum to 1 so that $\mu_j = c$ under $H_a : f - g = c$, where c is a constant. Use of local linear smoothers (Fan 1992) will accomplish this and handle the boundary problems simultaneously.

Let $\mathbf{h}^* \in H_{n,m}$ maximize Δ_3. We define the sample version of Δ_3 as

$$\hat{\Delta}_3(h_1, h_2) = \frac{1}{n^3} \sum_{i=1}^{n} \left[\sum_{j=1}^{i} \hat{\mu}_j \right]^2 \tag{17}$$

where

$$\hat{\mu}_j = \frac{1}{2}\left\{\sum_{i=1}^{n}[\hat{f}(x_{1i}) - \hat{g}(x_{1i})]w_{1i}(x_{1j}) + \sum_{l=1}^{m}[\hat{f}(x_{2l}) - \hat{g}(x_{2l})]w_{2l}(x_{1j})\right\}.$$

where $\hat{f}$ and $\hat{g}$ are given by (3). The estimates of the bandwidths $\hat{\mathbf{h}} = (\hat{h}_1, \hat{h}_2)$ are obtained by maximizing $\hat{\Delta}_3$ over $H_{n,m}$. We propose to select the bandwidths for the test $\mathcal{W}_2$, which is based on T_2, T_2', using the same criterion since the divergence of the statistics T_2 and T_2' depends on the same quantity as in T_1 and T_1'. The asymptotic property of this procedure is given in Theorem 2.4.

Theorem 2.4 *Let assumptions $\mathcal{A}_1 - \mathcal{A}_3$ be satisfied. Then $\frac{\hat{\mathbf{h}}}{\mathbf{h}^*} \to \mathbf{1}$ in probability as $\min(m,n) \to \infty$.*

The result in Theorem 2.4 does not hold for alternatives of type $f(x) - g(x) = \mu(x)/\sqrt{n}$ due to the fact that $\hat{\Delta}_3$ does not converge to Δ_3 with such functions. The authors have not been able to develop a suitable approximation for such situations.

Remark 2.4 In applications of these tests one usually uses asymptotic critical points if the error distributions are unknown. Our empirical study shows that the preservation of size is not very good with asymptotic critical points when the empirical smoothing parameters are used with small to moderate sample sizes. If the error distributions were known, then we could simulate the critical points using $f = g \equiv 0$ given the values of $\hat{h}_1$ and $\hat{h}_2$. When one uses $f(x) = g(x) \equiv 0$ in simulating critical points for known error distributions, where actually $f(x) = g(x) \not\equiv 0$, since the value of $\hat{\mathbf{h}}$ heavily depends on the smoothness of f, the simulated critical values could have extra variation due to the common function

f under H_0. This is not the case if one were simulating critical points in KHW or HH.

Remark 2.5 In cases where the error distributions are unknown, we may use a bootstrap procedure similar to that of HH in obtaining critical points. First we define centered observations $y_{1i} = Y_{1i} - \overline{Y_1}, i = 1, ..., n$, and $y_{2j} = Y_{2j} - \overline{Y_2}, j = 1, ..., m$ where $\overline{Y_1} = \sum_{i=1}^{n} Y_{1i}/n$, $\overline{Y_2} = \sum_{j=1}^{m} Y_{2j}/m$ for the two samples. Then we re-sample the y_{1i}'s and y_{2j}'s separately getting samples of n and m responses respectively, along with the associated covariate values, to calculate partial sums and the test statistics for given values of h_1 and h_2 in the bootstrap procedure. The properties of the critical points thus obtained would be similar to those presented in HH when $f(x) = g(x) \equiv 0$. We can use the above critical points as good approximations in large samples because of the convergence results in KU and the comments on Remark 2.4

3. EMPIRICAL RESULTS AND EXAMPLES

In this section we give some simulation results regarding the proposed procedures and the application of the smoothing parameter selection criteria on two real data sets.

3.1 SIMULATIONS

In order to examine the power performance and the possible effect of the data based smoothing parameters on the size of exact tests, we used a set of detailed simulations. For power comparisons, we looked at four types of alternatives,

1. C_1: $\mu(x) = a \cos k\pi x + b \sin j\pi x$ where a and b are real numbers and k and j are integers;

2. C_2: $\mu(x) = a \cos^2 k\pi x + b \sin^2 j\pi x$ where a and b are real numbers and k and j are integers;

3. C_3: $\mu(x) = ax^b + c$ where a, b and c are real numbers;

4. $C_1 \oplus C_3$ and $C_2 \oplus C_3$ where $\oplus$ means the direct sum.

Here $\mu = f - g$. These types cover monotone functions, periodic functions with high frequency and combinations of the two. As for error distributions, we used

$$\begin{aligned}
&(\mathrm{i})\{0.5N_1, 0.5N_2\}, \\
&(\mathrm{ii})\left\{c\left(|N_1| - (2/\pi)^{1/2}\right), c\left(|N_2| - (2/\pi)^{1/2}\right)\right\}.
\end{aligned}$$

Here $N_i, i = 1, 2$ are standard normal variates and $c = (\pi/2(\pi - 2))^{1/2}$. These distributions give an idea of how the level and power is affected for symmetric and skewed error distributions when empirical smoothing parameters are used in each test. The choice of error distributions make $\gamma = 1$.

Since the methods in KHW and HH are applicable only with matched designs, we restricted ourselves to equispaced equal designs in almost all cases. A very small set of simulations with unequal designs was conducted for examining the KU procedures exclusively, where the designs for these were generated using uniform $(0, 1)$ random variables for the simulations. In KU we examined $\mathcal{W}_1 = \min(T_1, T_1')$ and $\mathcal{W}_2 = \min(|T_2|, |T_2'|)$. The sample sizes were taken to be equal with a level 0.05 for all tests.

The equispaced bandwidths $h_l, l = 1, ..., 50$ from the set $H_n = [0.05, 0.25]$ was used in KHW. In KU we used $h_1, h_2 \in [0.05, 0.25]$ such that the h_i's $(i = 1, 2)$ were equispaced. Except for a few simulations with sample sizes 100, we used sample sizes of 25 and 50 in the simulations to keep the computation time of reasonable length. For the exact test in KHW we used standard normal errors as difference samples and found the critical values $c_\alpha(h_l), l = 1, ..., 50$ for the equispaced design by simulation. Simulation for getting the critical points in KHW under a normal error structure is always possible due to the scale invariance of the test statistic. However this adds some extra variability into the procedure. In our study we used the upper 95th percentile of 1000 test statistics generated under the null hypothesis as the critical point for a given h. Once the corresponding $\hat{h}$ was found by maximizing the function $\hat{\Delta}_1$ over the set of values $h_l, l = 1, ..., 50$, $c_\alpha(\hat{h})$ was used as the critical value in testing. In KHW for non normal errors, the asymptotic critical points were used.

Critical points for KU were obtained by simulation (with 1000 repetitions) with the assumption $f = g = 0$, since it appeared that the asymptotic critical points were somewhat lower when compared with simulated values for small samples. The main reason for such differences is the high noise levels in the error distributions. In spite of the convergence of the null distribution of the test statistics in KU to appropriate limiting distributions, the rate of convergence can be largely affected by the possible correaltions within the quasi residuals. Even with samples of size $n = m = 100$, we have observed that there were some differences between the simulated and the asymptotic critical points. Adding to this, is the effect of the assumption of $f = g = 0$. When the two functions are different from zero, a bias of order h^2, which may be somewhat large for small sample sizes, is ignored in all calculations. We beleive that the variability due to simulation is much smaller compared with the errors due to ignoring bias terms and the correlations in the quasi residuals.

For HH, we maximized $\hat{\Delta}_2$ over $p_l, l = 1, ..., 50$, where p's were equispaced in $P = [0.2, 0.8]$. Once the maximizing value $\hat{p}$ was found, it was used in the bootstrap procedure to obtain the critical points and to calculate the test statistic $S(\hat{p})$. We used local linear smoothers (Fan, 1992) in estimation of regression functions since they need no boundary modifications and asymptotically they are equivalent to kernel smoothers. In the case of HH, we used a fixed bandwidth 0.2 in all regression curve estimations. For all simulations, IMSL functions were used whenever possible.

From Tables 1-8, we see that when selecting a data based smoothing parameter, the size of the test is inflated in KHW and KU. The amount of inflation was much higher if the asymptotic critical values were used. The size inflation in KHW also seemed to occur when using the exact test. In HH with normal errors, the empirical size tends to be smaller than the nominal level; the test becomes more conservative with increasing sample size. The differences between the empirical and the nominal size seem to diminish with larger sample sizes and smaller error variability. The size does not seem to be significantly affected by the type of the error distribution, except for a few cases in KHW.

The behavior of the power with respect to the smoothing parameters shows that it is advantageous to select a data based smoothing parameter. The tables present empirical power results for a few selected difference functions used in

KHW, HH and KU along with some other difference functions.

In KHW, the behavior of power with various smoothing parameters for several difference functions were given. Although it is not meaningful to compare the simulation results given in KHW to the results here due to differences in the random number generation in getting critical points and test statistics and due to different number of simulations, with $n = 50$, it can be seen that the proposed empirical smoothing parameters for the same difference functions provide powers that are very close to the maximum power given in KHW. Also given in the tables is the power π^* (under the heading Δ_1), which uses the smoothing parameter value that maximizes the exact Δ_1 function in each case. The power with empirical smoothing parameters seem to be higher than π^* in some cases although the difference becomes negligible when the sample size increases. This is most likely due to high variability that is associated with $\hat{\Delta}_1$. It is also noteworthy that the power is somewhat lower with the second pair of error distributions than with the normal errors.

In HH, the power was examined only for three values of the smoothing parameter p. The results of the power simulations given in HH indicate that the empirical power is monotone in p. This is not in general true and using the method proposed here, we observe that one can enhance the power of the bootstrap test by selecting an optimal p. The value of p was estimated using the data and that value was used in implementing the bootstrap procedure. This makes the bootstrap procedure meaningful, since the smoothing parameter is fixed conditional on the sample. The powers using the p that maximized $\hat{\Delta}_2$ and Δ_2 are also given in Tables 1 and 2. The powers resulting a maximization of $\hat{\Delta}_2$ are close to the powers resulting from a maximization of Δ_2 in most cases. In this case also, the former is larger than the latter in some cases indicating the effect of the variability of $\hat{\Delta}_2$ on the choice of p.

The behavior of power in KU with the empirical bandwidths shows that if the difference $f - g$ is monotonic, then the test $\mathcal{W}_2$ performs better than $\mathcal{W}_1$ with the proposed selection criteria. For differences that are non-monotonic, $\mathcal{W}_1$ seem to perform better. Also, for alternatives with a larger frequency, the powers of $\mathcal{W}_1$ and $\mathcal{W}_2$ are somewhat lower than the power reported in Kulasekera (1995). This is due to the use of simulated critical points in this paper, which are larger than the asymptotic critical points that were used in Kulasekera (1995). The power with unequal designs was not much different than the equal designs case in KU (see Tables 8 and 9).

Although most simulated results corresponding to $\hat{\Delta}_i$,$i = 1, 2, 3$, agree with the results corresponding to Δ_i, $i = 1, 2, 3$, there were a few descrepencies. These are mainly due to the variability and the bias of the sample versions of the Δ's. In spite of the asymptotic statements, one may not see close agreement of the two results for some alternatives due to slow convergence. Also, the power corresponding to the smoothing parameter that maximizes the appropriate Δ may not be as large as the optimal power unless the sample size is very large. In this study, we actually examine the relationship between the powers with the sample version and the true parameter Δ that infuences the power in each test. Optimality of choosing the smoothing parameters by maximizing the $\hat{\Delta}_i$, $i = 1, 2, 3$ in each case is justified using the result that for large samples, the optimal power is achieved with a smoothing parameter that is very closed to the smoothing parameter that maximizes the appropriate Δ_i. A few repeated simulation sets with different seeds of random number generation showed no evidence of serious changes in the power results, suggesting non significant variability due to the pseudo randomness.

3.2 Drilling Data Example

This data was collected to see whether drilling holes in rock is faster using *dry* or *wet* drilling (Hand et al. 1994). In an area of about 10 feet by 20 feet, six holes, three *dry* holes and three *wet* holes were drilled and the depths in 5 feet increments and drill times (in 1/100 minutes) were recorded. There were a few unusually large drilling times in the collection of data. For our analysis, we removed such observations. The reduced data set had 40 observations in each group, and the average of the three drilling times in *dry* and *wet* groups were taken as the response variable for the two populations. Here, the two averaged Y values for the two groups are actually matched and do not satisfy the independence assumptions for KU. The fitted curves using local linear smoothing is given in Figure 1. An examination of the residuals in each group did not indicate serious departures from the assumption of independence within the group. The tests statistic values and the data based optimal smoothing parameters that were used in calculating the test statistics in KHW and HH (with the 0.05 critical points in parentheses) are

$$\text{KHW: } T = 11.45(0.287) \quad \hat{h} = 20;$$
$$\text{HH: } S = 116.09(1.309) \quad \hat{p} = 0.7.$$

3.3 Ethanol Data Example

Since KU is designed for unequal sample sizes and designs, we used the ethanol data (Kulasekera 1995, Cleveland 1993) to illustrate the use of Δ_3 and $\hat{\Delta}_3$. The response variable Y is the concentration of nitric oxide plus the concentration of nitrogen dioxide in the exhaust of an experimental engine when the engine is set at various equivalence ratios (x) and compression ratios (c). The (x, Y) pairs are divided into two groups according to *Low*: $c < 10$ and *High*: $c \geq 10$, and we model Y with x as in (2) for the two groups. The *Low* group had 39 observations while the other had 49 observations. The residuals show non normality in both samples and a mild correlation for the *Low* compression ratios. We proceeded, ignoring the possible correlations, to fit the curves and calculated the value of $\mathbf{h}^* = (h_1^*, h_2^*)$ and $\hat{\mathbf{h}} = (\hat{h}_1, \hat{h}_2)$ for these two samples as well the test statistics $\mathcal{W}_1 = \min(T_1, T_1')$ and $\mathcal{W}_2 = \min(|T_2|, |T_2'|)$. The fitted curves are given in Figure 2. The optimal bandwidths and the values of the test statistics (with 0.05 asymptotic critical values in parentheses) were

$$\mathcal{W}_1 = 45.6(0.698), \;\; \mathcal{W}_2 = 195.4(0.521)$$

and

$$\mathbf{h}^* = (0.04, 0.033).$$

This shows strong evidence against equality at a 0.05 level. In KU several bandwidths including the bandwidth by simple cross validation were used to calculate the tests statistics. None of the values gave contradictory results.

4. CONCLUSION

We have examined methods to obtain optimal power in a test of equality of regression curves in this article. The proposed methodology can be applied in a variety of settings, such as normal and nonnormal error structures, different covariate designs for the two samples etc. It appears that one can achieve significantly better power with a suitable optimality criterion function Δ. The power improvement is clear in all tests we have considered here, some providing more evidence depending on the alternative.

The Δ functions we have introduced here may not be the only choices in optimizing the power. Any parameter that could be related to power which takes on a known value (which can be assumed as 0) under H_0, is suitable in obtaining the appropriate data based smoothing parameter. In particular, this is true for KU. The disadvantage in all these procedures is, however, that we need to specify the parameter $\Delta(h)$, which frequently requires the asymptotic properties of the test and involves f and g. We then have to obtain a good nonparametric estimator of Δ, which has to be maximized with respect to h. It would be desirable to develop a method that involves an estimate of the power itself for a given test. One possibility is to define a function $\Delta(h)$, which is in some sense the empirical power, that can be maximized with respect to h. The authors are currently examining a procedure that divides the samples into several subsamples, which can be used to test the equality in certain parts of the domain of the covariate. This is somewhat analogous to the minimization of the cross validation function $(CV(h))$ in selecting the bandwidth to minimize the squared error.

The main draw back in all empirical smoothing parameter selection procedures is the effect on the size of the test as discussed in the previous sections. Although for large sample sizes the departure of the actual size from the intended size becomes small, it is difficult to asses the inflation of size for each test. One could use a size $r\alpha$, $r < 1$ to obtain $\hat{h}$ and then use it in a test with size $r\alpha$ to keep the level accurate. The choice of r can be made subjectively in each case. One could perhaps implement the following bootstrap procedure to obtain a preliminary value of r.

1. Take two samples of size m and n (two bootstrap samples) with replacement from one of the original samples and select the appropriate smoothing parameter based on the bootstrap samples.

2. Conduct the test of equality using these two samples with the selected smoothing parameter.

3. Count a 1 if this leads to a rejection.

This gives a crude estimate of the size, which could be used to select a value r. We have not pursued this approach here since none of the simulated sizes warranted it.

5. Appendix

We sketch the proof of Theorem 2.1 and Theorem 2.2. The proofs of Theorem 2.3 and Theorem 2.4 are very similar to the proof of Theorem 2.1 and hence omitted.

Proof of Theorem 2.1

$$\begin{aligned}
\hat{\Delta}_1(h) &= \sqrt{n^2h}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_i^2 - \frac{\kappa_2 s^2}{nh}\right] \\
&= \sqrt{n^2h}\left[\sum_{i=1}^{n}\frac{1}{n}\left[\sum_{j=1}^{n} D_j w_j(x_{1i})\right]^2 - \frac{\kappa_2 s^2}{nh}\right] \\
&= \sqrt{n^2h}\left[\sum_{i=1}^{n}\frac{1}{n}\left[\sum_{j=1}^{n} (\mu(x_{1j}) + (\epsilon_j - \eta_j))w_j(x_{1i})\right]^2 - \frac{\kappa_2 s^2}{nh}\right]
\end{aligned}$$

Now, expanding the summation above and using arguments similar to those for getting the mean and variance in regression estimates (Eubank 1988, chap. 4) and observing that s^2 is a consistent estimator of $\sigma^2 + \tau^2$, we see that

$$(n^{2-\delta})^{-1/2}\hat{\Delta}_1(h) = (n^{2-\delta})^{-1/2}\Delta_1(h) + R_n(h) \qquad \text{(A1)}$$

where $R_n(h) \to 0$ in probability when $n \to \infty, nh \to \infty$ and $nh^4 \to 0$. Since $\hat{h}$ and h^* are both in H_n and the two functions Δ_1 and $\hat{\Delta}_1$ are both continuous in h, we use arguments similar to those of Hall (1983) to obtain

$$\frac{\hat{\Delta}_1(\hat{h})}{\Delta_1(h^*)} \to 1$$

in probability and

$$\frac{\hat{h}}{h^*} \to 1$$

in probability.

Proof of Theorem 2.2

It is clear that

$$\pi(h^{**}) \geq \pi(h^*) \tag{A2}$$

Now, consider

$$\begin{aligned}
\pi(h) &= P\left[\frac{T - E(T)}{\sigma(T)} \geq \frac{c_\alpha(h) - E(T)}{\sigma(T)}\right] \\
&= P\left[Z \geq z_\alpha - \frac{\Delta_1(h)}{\sqrt{n^{2-\delta}B}}\right] + a_n(h)
\end{aligned}$$

for n sufficiently large. Here Z is a standard normal variate, z_α is the upper α percentage point of Z and, $a_n(h)$ is the error of normal approximation of the distribution of T with a bandwidth $h \in H_n$, where $a_n(h) \to 0$ as $n \to \infty$. Then,

$$\begin{aligned}
P\left[Z \geq z_\alpha - \frac{\Delta_1(h)}{\sqrt{n^{2-\delta}B}}\right] &\leq P\left[Z \geq z_\alpha - \frac{\Delta_1(h^*)}{\sqrt{n^{2-\delta}B}}\right] \\
&= \pi(h^*) + a_n(h^*)
\end{aligned} \tag{A3}$$

for large n where $a_n(h^*) \to 0$ as $n \to \infty$.

Thus, we have for large n,

$$\pi(h^{**}) \leq \pi(h^*) + a_n(h^*) + a_n(h^{**}) \tag{A4}$$

where $a_n(h^*) + a_n(h^{**}) \to 0$ as $n \to \infty$. Combining (A2) and (A4) we get the desired result.

REFERENCES

Cleveland, W.S. (1993), *Visualizing Data*, Hobart Press, Summit, New Jersey.

Eubank, R.L. (1988), *Spline Smoothing and Nonparametric Regression,* Marcel Dekker, Inc., New York.

Eubank, R.L. and Hart, J.D. (1992), "Testing Goodness-of-Fit in Regression Via Order Selection Criteria," *The Annals of Statistics* , 20, 1412-1425.

Fan, J. (1992), "Design Adaptive Nonparametric Regression," *Journal of the American Statistical Association,* 87, 998-1004.

Hall, P. (1983),"Large Sample Optimality of Least Squares Cross-Validation in Density Estimation," *The Annals of Statistics,* 11, 1156-1174.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., and Ostrowski, E. (1994) *Small Data Sets*, Chapman & Hall, London.

Hall, P., and Hart, J.D. (1990), "Bootstrap Test for Difference Between Means in Nonparametric Regression," *Journal of American Statistical Association,* 85, 1039-1049.

Hall, P., Kay, J.W., and Titterington, D.M. (1990), "Asymptotically Optimal Difference Based Estimation of Variance in Nonparametric Regression," *Biometrika,* 77, 521-528.

Härdle, W., and Marron, J.S. (1990), "Semiparametric Comparisons of Regression Curves," *The Annals of Statistics,* 18, 63-89.

King, E.C. (1988), "A Test for the Equality of Two Regression Curves Based on Kernel Smoothers," Ph.D. dissertation, Department of Statistics, Texas A&M University.

King, E.C., Hart, J.D., and Wehrly, T.E. (1991), "Testing the Equality of Two Regression Curves using Linear Smoothers," *Statistics & Probability Letters,* 12, 239-247.

Kulasekera, K.B. (1995), "Testing the Equality of Regression Curves using Quasi Residuals," *Journal of the American Statistical Association,* 90, 1085-1093.

Ledwina, T. (1994), "Data-Driven Version of Neyman's Smooth Test of Fit," *Journal of the American Statistical Association,* 89, 1000-1005.

Müller, H-G. (1987), *Nonparametric Regression Analysis of Longitudinal Data*, Springer-Verlag, Berlin.