Comparison of Regression Curves Using Quasi Residuals

K.B. Kulasekera*
Department of Mathematical Sciences
Clemson University
Clemson, SC 29634

Abstract

We consider testing the equality of two regression functions using two independent samples. Three tests are proposed that are free of the restriction of having the same covariate values or sample sizes for both samples. Asymptotic distributions are given and results from a simulation study are presented that show the superior power properties of these tests over a competing test in a variety of cases, including the testing of hypotheses involving high frequency curves when the design points for the two samples differ. It is also observed that the tests have good level properties when proper smoothing parameters are selected.

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

1. INTRODUCTION

In scientific and social studies, it is very common to compare groups of individuals (or items) in terms of some parameter. For example, a medical researcher may wish to compare the mean reaction time for a drug in male and female patients. If the experimenter has auxiliary information such as the age of each patient, then he or she may use regression to model the mean reaction time as a function of age in each population and compare the two regressions. Incorporating the extra

*K.B. Kulasekera is Associate Professor, Department of Mathematical Sciences, Clemson University, Clemson, SC 29634. The author wishes to thank an associate editor and an anonymous referee for helpful comments which improved the presentation of the article.

information, such as age in the preceding example, may improve the accuracy of the experimenter's procedures.

The comparison of regression functions for two or more populations is a widely discussed issue. The classical methods employ parametric models for the regression function and compare the resulting model parameters. A disadvantage of this approach is that it requires the specification of a parametric model, which is often difficult.

In this paper we discuss testing the equality of two regression functions that are specified in terms of only some smoothness conditions and, as a result, the proposed tests can be used for a wide class of functions. The specific setting can be described as follows. We observe data in the form $\{(Y_{1i}, x_{1i}), i = 1, ..., n\}$ and $\{(Y_{2j}, x_{2j}), j = 1, ..., m\}$ with

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

where $\epsilon_i, i = 1, ..., n$, 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. We are interested in testing the hypothesis

$$H_0 : f = g \quad vs \quad H_1 : f \neq g \tag{2}$$

over the domain of the covariate (x) measurements, which is taken as $[0, 1]$ in this article. The covariate values (design points) for the two samples may be different, and the sample sizes can be unequal. Let $Var(\epsilon_1) = \sigma^2$ and $Var(\eta_1) = \tau^2$.

Testing hypotheses of the above type in a nonparametric setting has been discussed by Hall and Hart (1990), Härdle and Marron (1990), and King, Hart and Wehrly (1991). Hall and Hart (1990) discuss a bootstrap procedure for testing hypotheses (2) under the assumptions that $m = n$ and that there are common design points (i.e., $x_{1i} = x_{2i}$ for all i). They extend their test to samples of equal size but with unequal design points using an assumption of *closeness* of the two designs. King et al. (1991) propose two tests, both with common design points, one for normal errors and the other for nonnormal errors, that are shown to have good power properties. The semi-parametric test proposed by Härdle and Marron (1990) also assumes equal designs. In addition to these, Cox, Koh, Wahba and Yandell (1988), Eubank and Spiegelman (1990), and Eubank and Hart (1992) have discussed closely related issues.

The equality of designs and/or sample sizes for the two populations may be difficult to achieve in practical studies. Equality of designs also fails to hold when the design variable is random. In this paper, we discuss three tests for testing hypotheses (2) using independent samples from the two populations, where the tests do not require *common design points* or *equal sample sizes*. Although we consider non-stochastic designs in our descriptions, the methods are equally applicable for random designs with suitable modifications.

The first two tests, which we will denote as $\mathcal{W}_1$and $\mathcal{W}_2$, use weighted averages of squares of partial sums of *quasi residuals*, which are defined as

$$e_{1i} = Y_{1i} - \hat{g}(x_{1i}),\ i = 1, ..., n, \tag{3}$$

and

$$e_{2j} = Y_{2j} - \hat{f}(x_{2j}),\ j = 1, ..., m. \tag{4}$$

Here $\hat{f}$ and $\hat{g}$ are suitable estimators, such as kernel estimators, of f and g, respectively, each based only on the observations from the corresponding population.

The tests are motivated by the fact that, for sufficiently large m and n and under the null hypothesis, the e_{1i}'s and e_{2j}'s have distributions close to those of the ϵ's and η's, respectively, and the distributions of these quasi residuals depend on the difference $f - g$, under the alternative hypothesis. The first two tests are based on statistics constructed using arguments similar to those in Hall and Hart (1990), that is, we form suitable averages of squares of partial sums of the quasi residuals. These averages tend to be large for departures from the null hypothesis.

The third test, denoted as $\mathcal{W}_3$, is a von Neumann-type test based on the estimators of the variances of the error distributions. In particular, we compare an estimator of the variance of the ϵ's based on the e_{1i}, $i = 1, ..., n$, to an estimator based on difference methods (Hall, Kay, and Titterington 1990). The approach is similar to the approaches of Munson and Jernigan (1989) and Eubank and Hart (1993).

An empirical study of the level and the power of the three proposed tests, which includes high frequency curves (Eubank and Hart 1992), shows the power superiority of these tests over the test of Hall and Hart (1990) for unequal designs and they all seem to have comparable levels. We show that the first two tests have "*root-* n asymptotics". It is also observed that test $\mathcal{W}_2$, after suitable modification, can be used in one-sided situations. Test $\mathcal{W}_3$ detects local alternatives that converge at a somewhat slower rate. However, our empirical study reveals that $\mathcal{W}_3$ outperforms $\mathcal{W}_1$ and $\mathcal{W}_2$ in certain cases. Finite sample performances of this type are not surprising in view of the results by Ghosh and Huang (1991), Eubank and LaRiccia (1992), and Eubank, Hart, and LaRiccia (1993). The same asymptotic rates apply when the sample sizes are unequal.

In the special case of equal sample sizes and equal designs, the first two tests we propose are similar to the procedures used by Hall and Hart (1990) and King et al. (1991). Also, as will be seen from our empirical study, the Hall and Hart procedure continues to have levels close to the nominal level in spite of the differences in the designs, although there is a slight drop in accuracy.

The rest of this article is organized as follows. In Section 2, we describe each test and discuss its asymptotic properties in detail. Section 3 gives results of a empirical study and examines the finite sample performances of each test. The data from an engine exhaust study at different settings (Ethanol data, Cleveland 1993) are also analysed in Section 3 using the proposed methodology. A discussion regarding the usage of these tests is given in Section 4. Finally, the proofs of the theoretical results are provided in Section 5.

2. TESTS AND PROPERTIES

Let the quasi residuals (hereafter, residuals for brevity) be defined as in (3) and (4) where $\hat{f}$ and $\hat{g}$ are suitable estimators of f and g. In this article we use a version of kernel estimators of f and g (Eubank 1988; Müller 1987), which are defined in the following fashion.

Let K be a known symmetric probability density function on $[-1, 1]$ that has finite second moment. For the covariate values x_{1i}, $i = 1, ..., n$, and x_{2j}, $j = 1, ..., m$, all falling in $[0, 1]$, define

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{Y_{1i} w_{1i}}{b} \tag{5}$$

and

$$\hat{g}(x) = \sum_{j=1}^{m} \frac{Y_{2j} w_{2j}}{b} \tag{6}$$

where $w_{1i} = \int_{s_{1,i-1}}^{s_{1,i}} K\left((x-u)/b\right) du$, $w_{2j} = \int_{s_{2,j-1}}^{s_{2,j}} K\left((x-u)/b\right) du$, $s_{1,0} = 0, s_{1,i} = (x_{1i} + x_{1\,i+1})/2, s_{1,n} = 1, s_{2,0} = 0, s_{2,j} = (x_{2j} + x_{2\,j+1})/2$, and $s_{2,m} = 1$. Here, b is the bandwidth of the kernel estimators. Although in practice the bandwidth is typically a function of the sample size and the sample itself, we omit this dependence in the notation.

Using the estimators (5) and (6) of f and g we define the partial sums

$$\mu_n(t) = \sum_{i=1}^{[nt]} \frac{e_{1i}}{\sqrt{n}} \tag{7}$$

and

$$\mu_m(t) = \sum_{j=1}^{[mt]} \frac{e_{2j}}{\sqrt{m}}, \tag{8}$$

for $0 < t < 1$. These partial sums, which under the null hypothesis act as partial sums of mean zero random errors, are the primary building blocks of the tests $\mathcal{W}_1$ and $\mathcal{W}_2$ to be defined shortly.

2.1 Tests $\mathcal{W}_1$, $\mathcal{W}_2$ and $\mathcal{W}_3$

Let s_n^2 and s_m^2 be consistent estimators of σ^2 and τ^2 (e.g., difference type estimators of Hall et al. 1990 or Rice 1984) and define

$$T_1 = \sum_{k=1}^{n} \frac{\mu_n^2(k/n)}{ns_n^2}, \quad T_1' = \sum_{k=1}^{m} \frac{\mu_m^2(k/m)}{ms_m^2}. \tag{9}$$

Under the alternative hypothesis, T_1 and T_1' tend to be *large*. Thus, one test for H_0 can be obtained by rejecting H_0 if $\mathcal{W}_1 = \Delta(T_1, T_1')$ is too large for some suitable function Δ, such as $\Delta(u, v) = min(u, v)$ or $\Delta(u, v) = u$. In general, we can define a class of test statistics,

$$\{\mathcal{W}_1 = \Delta(T_1, T_1') : \Delta \in \mathcal{S}\} \tag{10}$$

where $\mathcal{S}$ is an approprite subclass of the class of continuous functions on the two-dimensional Euclidean space equipped with the natural topology. In fact, we can expand the class of functions $\mathcal{S}$ to be the class of measurable functions, but in applications it is not difficult to choose an appropriate Δ which is continuous.

Note that the residuals are normalized with the estimated standard deviations for each population without adjusting for the variance of the difference. For example, the variance of $Y_{1i} - \hat{g}(x_{1i})$ is $\sigma^2 + O(1/mb)\tau^2$, which is almost the variance of Y_{1i} for large m. Thus any adjustments will play a minimal role in the asymptotics. In a small sample one can substitute a consistent estimator of the variance of the e_{1i} or e_{2j} in normalizing, with no effect on the large sample critical points.

The statistics T_1 and T_1' are similar to T_n in Eubank and Hart (1992) who note possible deficiencies for such statistics in testing hypotheses of special natures. However, von Mises type statistics are effective for detecting departures from the null hypothesis in many directions, particularly low frequency deviations from H_0 (Eubank et al. 1993), and may provide adequate power for such alternatives.

As another test of H_0 we propose a weighted version of $\mathcal{W}_1$ that is designed to enhance its power properties against special alternatives, such as $f > g$. Define statistics T_2 and T_2' as

$$T_2 = \int_0^1 \frac{\mu_n^2(t-)}{s_n^3} d\mu_n(t), \quad T_2' = \int_0^1 \frac{\mu_m^2(t-)}{s_m^3} d\mu_m(t) \tag{11}$$

where $\mu_n(t-)$ and $\mu_m(t-)$ are left hand-limits at t. Note that we can write T_2 as

$$T_2 = \sum_{i=1}^{n-1} \frac{\mu_n^2(i/n) e_{1i+1}}{\sqrt{n} s_n^3}$$

with a similar representation for T_2'. These statistics differ from the statistics in $\mathcal{W}_1$ by the weight given to the square of each partial sum. The motivation here is that if there is a trend in the residuals, it can be magnified by an appropriate weight in the same direction. The magnitudes of T_2 and T_2' tend to be large under the alternative hypothesis, and therefore, one would reject the null hypothesis if $\mathcal{W}_2 = \Delta(T_2, T_2')$ is too large for a suitably chosen function Δ. The function Δ can be chosen to provide a one-sided test. For example, for an alternative of the type $f > g$ one can choose $\Delta(u, v) = u - v$.

Our final test is based on a comparison of variance estimators assuming that the ratio of error variances $q = \sigma^2/\tau^2$ is known. Let d_j, $j = 0, ..., r$ be one of the optimal difference sequences for variance estimation in nonparametric regression given in Hall et al. (1990). Now, define $T_3 = \sum_{i=1}^n e_{1i}^2/n$ and let

$$\mathcal{W}_3 = \frac{\sqrt{n+m}\,(T_3 - q\hat{\tau}^2)}{\hat{\sigma_1}^2}. \tag{12}$$

where $\hat{\tau}^2 = (m-r)^{-1} \sum_{i=1}^{m-r} \left[\sum_{j=0}^r d_j Y_{2\ i+j}\right]^2$ is a difference estimator of τ^2 based on the Y_{2j}'s, $\hat{\sigma_1}^2 = \left[\omega^{-2}\widehat{Var}(\epsilon^2) + (1-\omega^2)^{-1}(aq)^2\right]^{1/2}$, $\omega^2 = n/(n+m)$, $a^2 = \hat{\kappa}\hat{\tau}^4 + 2\hat{\tau}^4 \sum_{i=1}^{m-r} \left[\sum_{j=0}^r d_j d_{i+j}\right]^2$, $\hat{\kappa} = \widehat{E(\eta^4)}\hat{\tau}^{-4} q^{-2} - 3$, and $\widehat{E(\eta^4)}$ and $\widehat{Var}(\epsilon^2)$ are consistent estimators of $E(\eta^4)$ and $Var(\epsilon^2)$, respectively. Note that under the alternative hypothesis T_3 tends to depart from σ^2 so that a properly normalized version of T_3, (i.e., $\mathcal{W}_3$), will detect depatures from the null hypothesis. Thus one would reject the null hypothesis for large values of $\mathcal{W}_3$.

There are many practical situations in which it is reasonable to assume $q = 1$ and we shall use $q = 1$ in the sequel. Also, the terms involving $\hat{\kappa}$, a consistent estimator of the kurtosis of the error distribution, do not have to be calculated under the assumption of normal errors. The test becomes much simpler to implement in such cases.

2.2 Asymptotic properties

We list the asymptotic properties of T_i, T_i' , $i = 1, 2$ and $\mathcal{W}_3$ in the form of three theorems. If the function Δ used in forming the test statistics is continuous, the independence of the samples can be used along with Theorems 2.1 and 2.2, to obtain the asymptotic distribution of $\mathcal{W}_i = \Delta(T_i, T_i'), i = 1, 2$. For some choices of Δ, e.g., $\Delta(u, v) = min(u, v)$, this asymptotic distribution can be written in a near closed form. In addition to the conditions imposed on the kernel function K, we shall need the following assumptions at various points.

$\mathcal{A}_1$**:** The functions f and g are twice continuously differentiable.

$\mathcal{A}_2$**:** The bandwidth sequence b used in estimating f and g satisfies $min(m, n)b \to \infty$ and $min(m, n)b^4 \to 0$ as $min(m, n) \to \infty$.

$\mathcal{A}_3$**:** 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 b used in estimating f and g satisfies $min(m, n)b \to \infty$ and $min(m, n)b^5 \to 0$ as $min(m, n) \to \infty$.

Theorem 2.1 *Let assumptions $\mathcal{A}_1 - \mathcal{A}_3$ be satisfied. Then, under H_o, T_1 converges in distribution to a random variable*

$$\xi = \int_0^1 B^2(t)dt,$$

as $min(m, n) \to \infty$, where B is the Weiner process on $[0, 1]$. The distribution of ξ is the same as that of the random variable

$$\sum_{j=0}^{\infty} \frac{4Z_j^2}{[(2j+1)\pi]^2}, \tag{13}$$

where the Z_j's are i.i.d. standard normal random variables. The statistic T_1' also converges in distribution to a random variable ξ', which has the same representation as ξ under the same conditions and is independent of ξ.

Theorem 2.2 *Under the null hypothesis and under the conditions of Theorem 2.1, T_2 converges in distribution to a random variable*

$$\zeta = \frac{B^3(1)}{3} - \int_0^1 B(t)dt,$$

where B is a Weiner process on $[0, 1]$. Moreover, we can express $\zeta = (N_1^3/3) - (N_2/\sqrt{3})$, where $N_i, i = 1, 2$ are joint normal random variables with mean zero, *marginal variances 1 and* $Cov(N_1, N_2) = \sqrt{3}/2$*. The statistic* T_2' *also converges in distribution to a random variable* ζ'*, with a representation like* ζ *and independent of* ζ*.*

Theorem 2.3 *Assume that conditions* $\mathcal{A}_1, \mathcal{A}_3$ *and* $\mathcal{A}_5$ *hold and that* $m/(m+n) \to \theta, 0 < \theta < 1$*, as* $min(n, m) \to \infty$*. Then, under* H_0*, the test statistic* $\mathcal{W}_3$ *converges in distribution to a standard normal random variable.*

Proofs of these Theorems are given in Section 5.

Remark 2.1 Using arguments similar to those in Hall and Hart (1990), T_1, T_1', T_2 and T_2' can be shown to diverge to ∞ at a rate of n under the alternative hypothesis. This enables these statistics to detect differences of order $1/\sqrt{n}$. Thus, when they are suitably combined to form a test statistic, such as $min(T_1, T_1')$, they can detect local alternatives comparable to those detectable with parametric tests. The statistic $\mathcal{W}_3$ diverge to ∞ at a much slower rate $(\sqrt{n})$ and, therefore, may not be very suitable for detecting slight depatures from the null hypothesis. However, its performance in our simulation was comparable to the other procedures, especially when high frequency functions were being tested.

Remark 2.2 In $\mathcal{W}_3$, only residuals from one sample are considered since the other sample is used to estimate the variance τ^2. An alternative is to combine the residuals from the both samples to form a statistic such as $\sqrt{n+m}\left[\sum_{i=1}^{n} e_{1i}^2 - \sum_{j=1}^{m} e_{2j}^2\right]$ and normalize it properly to obtain a test. However, the power of the test will suffer since it will depend on $\int_0^1 \left[f^2(x) - g^2(x)\right] dx$, which can be small.

Remark 2.3 For a given continuous function Δ on the product space, one has under the null hypothesis, the result that $\Delta(T_1, T_1') \Rightarrow \Delta(\xi, \xi')$, where $\Rightarrow$ indicate weak convergence. A similar result holds for T_2 and T_2'. Thus, the asymptotic distribution of the test statistic $\Delta(T_1, T_1')$ (or $\Delta(T_2, T_2')$) can be found.

Remark 2.4 It is clear that all the proposed tests are heavily dependent upon the bandwidth of the kernel estimators of f and g. The bandwidth sequence for all tests have been taken to be deterministic and are different from the asymptotic optimal bandwidths in terms of the mean squared error in regression curve estimation with positive kernel functions (Müller 1987). The conditions imposed on the bandwidths also rule out various data-based procedures for selecting a stochastic bandwidth for estimating f and g. The simulations described in the sequel showed that both the power and size of each test were very sensitive to the choice of the fixed bandwidth. The choice of the most suitable kernel order and bandwidths for optimal power and size for each type of alternative as well as the error structure are crucial issues that must be investigated theoretically and empirically. Perhaps one may be able to determine classes of alternatives for which some recommendations can be made regarding the bandwidth. In some of the simulations, we used simple cross validation in the selection of the bandwidth, while in the others we considered a large number of fixed bandwidths and examined the properties of the tests under each bandwidth with the same set of data. Although cross validation for large samples with positive kernel functions conflicts with the asymptotic requirement of $b = o((min(m,n))^{-1/4})$ in the first two tests and the requirement $b = o((min(m,n))^{-1/5})$ in the third, we believe, that for sample sizes such as $n, m \leq 50$, the simulated results obtained by selecting the bandwidth using a data-based method are realiable enough to get a feel for the performance of the proposed methodology.

3. EMPIRICAL RESULTS AND AN EXAMPLE

In this section we give the results of a simulation study of the level and power of the proposed tests and an example involving a real data set.

3.1 Simulations

The simulation study of level and power for the proposed procedures was conducted for a variety of regression functions f and g. For tests $\mathcal{W}_1$ and $\mathcal{W}_2$ we used $\Delta_1(u,v) = |u|$ and $\Delta_2(u,v) = min(|u|,|v|)$. This choice was motivated by ease of computing the limiting null distribution for such statistics. We chose the sample sizes to be equal so that we could also calculate the level and power for the Hall and Hart test in each case. For the Hall and Hart procedure we used their smoothing parameter $p = 1/2$ (Hall and Hart 1990) with the suggested modification for unequal designs. The functions chosen for the study of the level were mainly high frequency type functions, which can present greater difficulties in testing. Also, in the level study we chose three error distribution pairs as in Hall and Hart (1990), except that all errors were rescaled for smaller variability. The error distribution pairs used were

$$
\begin{aligned}
&(i)\{0.1N_1, 0.1N_2\},\\
&(ii)\left\{0.1\left(|N_1| - (2/\pi)^{1/2}\right), 0.1\left(|N_2| - (2/\pi)^{1/2}\right)\right\},\\
&(iii)\left\{0.1\left(|N_1| - (2/\pi)^{1/2}\right), 0.1\left((2/\pi)^{1/2} - |N_2|\right)\right\},
\end{aligned}
$$

where $N_i, i = 1, 2$ are standard normal variates. Larger noises, such as error standard deviation of 1 in (i), seem to give levels substantially different from the nominal levels unless very good choices of the bandwidths are made. However, for the power study we used the standard normal distribution in almost every situation. This choice of the error distribution made the data very noisy compared to the signal; power properties should improve with less noise. A few simulations for power with nonnormal errors, similar to the error distributions in Hall and Hart (1990) produced results similar to those with normal errors. In each sample, design points were generated using a uniform distribution, and whenever possible, *IMSL Fortran* subroutines were used in the computations.

To estimate the regression functions we used a boundary-modified rectangular kernel function on $[-1, 1]$ for nonperiodic functions. For periodic functions we extended the observations to $[-1, 2]$ (Eubank 1988; Rice 1984). In cases where edge effects seem to have some influence even after boundary modifications, one can use partial sums of the type $\mu_n^*(t) = \sum_{i:x_{1i} \geq b}^{min([nt],[n(1-b)])} e_{1i}$ and adjust the statistics T_i and T_i', $i = 1, 2$ accordingly, so that a few extreme quasi residuals would be removed from both edges. This has no influence in the asymptotics and improved the level accuracy in some simulations. This may be particularly useful when one has design points distributed according to two densities with skewnesses in opposite directions. Also, one can use local linear smoothers (Fan 1992) to estimate the functions f and g, which reduces boundary effects without modified kernels. In our simulations, we used a set of fixed bandwidths ranging from 0.01 to 0.15 as well as the bandwidths selected by simple cross validation, although the latter method does not quite meet the requirements of the asymptotic results, as was remarked on above. We present only a representative portion of the results in what follows.

The critical points for $\mathcal{W}_1$ were found by simulating the quantiles of ξ. To determine a quantile, a Monte Carlo simulation with 10,000 trials was used, where each trial generated 100 independent standard normal variates that were summed as in the series representation (13) for ξ. The decision to terminate the series at 100 terms was made because additional terms did not seem to produce significant changes in the quantiles of ξ. Critical points for $\mathcal{W}_2$ were determined using appropriate quantiles of ζ which can be obtained by a two variable quadrature.

The independence of ξ and ξ' as well as the independence of ζ and ζ' make the calculations of critical points very simple when using the min(.,.) function. We found that the critical points for Δ_2 are smaller than those for Δ_1.

Tables 1-3 give the empirical level for the tests $\mathcal{W}_1$, $\mathcal{W}_2$, $\mathcal{W}_3$ and for the Hall and Hart test (denoted as HH hereafter) conducted at a nominal level $\alpha = 0.05$.

The tables we provide here are for sample sizes $n = m = 25$. Tables 4-6 give the empirical power results for $\mathcal{W}_1$, $\mathcal{W}_2$, $\mathcal{W}_3$ and HH for sample sizes $20, 30$ and 50. Due to the large variety of the functions compared for the power study, we only give the tables corresponding to the case where the bandwidths were selected by cross validation. However, the power of the tests is very sensitive to the choice of the bandwidth, as was seen when different fixed bandwidths were used. This was especially true when the two functions were close, for example $f(x) = \sqrt{x}$ and $g(x) = f(x) + x$.

Tables 1-3 suggest that with asymptotic critical points the empirical level is close to the nominal level 0.05 in most cases except for high frequency curves. Even for high frequency curves, the departures from the nominal level for most cases are not alarmingly high when an appropriate smoothing parameter has been selected. Such departures seem to be minimal with normal errors and largest for the error distribution pair (iii). The sensitivity to bandwidth is clearly evident from all the results. Cross validation has provided somewhat conservative results in a majority of the cases we considered. It is also noteworthy that $\mathcal{W}_3$ seems to give much smaller levels in most situations. The HH procedure is more stable with respect to the level for all cases, which is expected due to the resampling scheme in that procedure.

An examination of the tables for empirical power indicate that the tests, including HH, are all comparable when very smooth functions are involved in testing. Table 4 indicates that tests $\mathcal{W}_1$, $\mathcal{W}_2$, and HH tend to be slightly better than test $\mathcal{W}_3$. However, when the functions f and g are oscillating functions or high frequency curves, the power of HH is considerably less than that of the tests we have proposed.

Although the HH test performs well for equal design points, the modification for unequal designs has a substantial drawback. The matching of responses with respect to closest design points scheme does not perform well when the noise is large and the mean curves are heavily oscillating. In such cases the HH test becomes insensitive to the difference in the mean functions.

The power of tests using Δ_2 tends to be slightly higher in almost all cases than the power using Δ_1. This is mostly due to the fact that both statistics T_i and $T_i', i = 1, 2$, tend to be close in magnitude in almost all the cases we have considered because of the uniform designs, but the critical points for Δ_2 are smaller than the critical points for Δ_1. The situation could be different with entirely different design distributions for the two samples.

In all cases we examined, $\mathcal{W}_3$, in spite of its low rate of divergence under the alternative hypothesis, performed remarkably well. Its power is well above all other procedures in some cases, confirming the discussions of statistics of this type by Ghosh and Huang (1991), Eubank and Hart (1992), and Eubank et al. (1993). All in all we can say that all three proposed tests have good power, even under nasty alternatives.

3.2 Ethanol Data Example

We now apply the proposed methodology to analyze a real data set. This data set pertains to an exhaust study using ethanol as the fuel in an experimental engine (Cleveland 1993). In particular, the reponse 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 different equivalence ratios (x) and compression ratios (c). These oxides of nitrogen are major air pollutants and the experiment is to determine the dependency of the concentration of oxides of nitrogen on various engine settings. The compression ratio takes on only a few values, which we categorize as *Low*: $c < 10$ and *High*: $c \geq 10$. We model Y with x as in (1) for the two types of compression ratios and compare the mean concentration functions. The *Low* group had 39 observations, while the *High* group had 49 observations. Boundary-modified rectangular kernels were used in the estimation of f and g. An examination of the residuals showed nonnormality and a mild correlation for the sample with *Low* compression ratios. However, we proceeded with the three tests for equality of the mean curves ignoring possible correlations in the error structure in the *Low* group.

In this example, we tried a large number of fixed bandwiths, ranging from 0.01 to 0.25, and all the values gave the same conclusions. The figure and the results we present are for bandwidths selected by simple cross validation which happened to fall in the range of the fixed bandwidth values we considered. The mean regression curves show a high nonlinearity with a mode around $x = 0.9$ for both populations. Also, it appears that lower compression ratios tend to give lower mean concentrations of the oxides. The calculated test statistics (with corresponding 0.05 critical values in parentheses) were

$$\mathcal{W}_1 : \Delta_1 = 76.4(1.673),\ \ \Delta_2 = 76.4(0.698),$$
$$\mathcal{W}_2 : \Delta_1 = 26.3(1.570),\ \ \Delta_2 = 24.1(0.521),$$

and

$$\mathcal{W}_3 = 2.27(1.645).$$

This shows strong evidence against equality at a 0.05 level. When the bandwidths were changed in the range 0.01 to 0.25, the values of the test statistics changed without an apparent pattern, although none of the values led to a different decision.

4. CONCLUSION

We have discussed the testing of the equality of regression functions under various settings in this article. The methods proposed have the flexibility to be used with different designs and were found to be significantly more powerful than a competing method while maintaining good level accuracy in a simulation study. The test $\mathcal{W}_2$ can be adjusted to be used for testing one-sided alternatives. Although we had exclusively used method of kernels in the estimation of f and g, one can use asymptotically equivalent estimators, such as local linear smoothers (Fan 1992). In addition, appropriate modifications in the assumptions will allow the use of these tests even under random designs, although we have not provided any details.

With equal designs, however, the Hall and Hart procedure has superior level accuracy because under the null hypothesis resampling is done from the empirical distribution, which is possible even when the reponses are matched. In our tests, we always have the problem of estimating the regression functions f and g, which will have the attendant problem of smoothing parameter selection. This is the main drawback of the tests $\mathcal{W}_1$, $\mathcal{W}_2$, and $\mathcal{W}_3$, making the Hall and Hart (1990) procedure preferred for equal designs. In the proposed tests, asymptotics require smaller bandwidths than those optimal in the sense of the mean squared error. This is because we are using the residuals to construct the tests, which requires smaller bias. Although the small samples used in the empirical study did not seem to produce much different power and size results for fixed versus stochastic bandwidths, usage of data-based smoothing parameters is not reliable for large samples. This is a problem which needs to be addressed in detail.

The tests we have proposed here and those in the literature all deal with univariate concommitant variables. When the covariate is multivariate, it is very unlikely that the common design assumption will be met. Thus the Hall and Hart procedures have to be modified accordingly, and such modifications may result in a substantial loss of power due to "matching" of close covariates. On the other hand, extensions of the tests $\mathcal{W}_1$, $\mathcal{W}_2$, and $\mathcal{W}_3$ to the multiple covariate case, will require bandwidths to be chosen small in order to match the slow convergence rates. This may cause deterioration of the small sample properties. Hence it may be of interest to develop further tests for multiple covariate cases.

5. PROOFS OF THEOREMS

Proof of Theorem 2.1

We first rewrite the residuals e_{1i} under the null hypothesis as

$$e_{1i} = \epsilon_i + g(x_{1i}) - E\left[\hat{g}(x_{1i})\right] + E\left[\hat{g}(x_{1i})\right] - \hat{g}(x_{1i}).$$

Then, $g(x_{1i}) - E\left[\hat{g}(x_{1i})\right] = O(b^2)$, and under assumptions $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$, if $min(m, n) \to \infty$, we have that

$$n^{-1/2} \sum_{i=1}^{j} \left\{g(x_{1i}) - E\left[\hat{g}(x_{1i})\right]\right\} \to 0 \tag{14}$$

for all $j \leq n$. Thus, the contribution to the test statistics T_1 from partial sums constructed out of terms of the type $g(x_{1i}) - E\left[\hat{g}(x_{1i})\right]$ tend to 0 as $min(m, n) \to \infty$. Now, we write

$$\varphi_i = E\left[\hat{g}(x_{1i})\right] - \hat{g}(x_{1i}),$$

and note that $E[\varphi_i] = 0$ and $var[\varphi_i] = O((mb)^{-1})$. Letting λ to be a sequence of real numbers such that $0 < \lambda < 1$ and $\lambda \to 0$ at a slow rate as $min(m, n) \to \infty$, e.g., $\lambda = 1/(\ln n + \ln m)$, and using the compactness of the support of the kernel function K, we have $Cov(\varphi_i, \varphi_j) = O((mb)^{-1})$ for $0 < |x_{1i} - x_{1j}| \leq 2b\lambda$, $Cov(\varphi_i, \varphi_j) = o((mb)^{-1})$ for $2b\lambda < |x_{1i} - x_{1j}| \leq 2b$ and $Cov(\varphi_i, \varphi_j) = 0$ otherwise. Now, using Chebyshev's inequality, it is seen that

$$n^{-1/2} \sum_{i=1}^{j} \varphi_i \xrightarrow{p} 0 \tag{15}$$

as $min(m, n) \to \infty$ for all $j \leq n$. Hence, by Donsker's invariance principle and the continuous mapping theorem (Billingsley 1968, p. 31), we obtain

$$T_1 \xrightarrow{\mathcal{D}} \xi = \int_0^1 B^2(t)dt, \tag{16}$$

where B is the regular *Weiner* process on $[0, 1]$ and $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

To obtain the exact distribution of ξ one proceeds as in Durbin (1973, sec. 4). First expand the covariance function of ξ in a series of orthonormal trigonometric functions and then use the orthonormality to get the i.i.d. standard normal variates in the statement of the theorem. We omit the details.

Proof of Theorem 2.2

Note that by (14) and (15) we have, under the null hypothesis, that the asymptotic distribution of T_2 is the same as the asymptotic distribution of the statistic $\int_0^1 \beta^2(s-)d\beta(s)$, where $\beta(t) = n^{-1/2}\sum_{i=1}^{[nt]} \phi_i$, $0 < t < 1$, with the ϕ_i i.i.d. random variables having mean 0 and variance 1. Now, using the property that β is a continuous time step martingale with respect to the natural filtration generated by the ϕ's, we invoke *Itô*'s formula for functions of bounded variation (Jacod and Shiryaev 1987, p. 57) to get

$$3\int_0^t \beta^2(s-)d\beta(s) = \beta^3(t) - \beta^3(0) - \sum_{0<s\leq t}\left[\beta^3(s) - \beta^3(s-) - 3\beta^2(s-)\delta(\beta(s))\right],$$

where $\delta(\beta(s)) = \beta(s) - \beta(s-)$. Thus,

$$3\int_0^t \beta^2(s-)d\beta(s) = \beta^3(t) - \sum_{0<s\leq t}\left\{[\delta(\beta(s))]^3 + 3\beta(s-)\left[\delta(\beta(s))\right]^2\right\}. \tag{17}$$

It can be seen that $\sum_{0<s\leq t}[\delta(\beta(s))]^3 \xrightarrow{p} 0$ as $n \to \infty$. To obtain the limit of the second term inside the sum, we proceed as follows. Let $\sum_{0<s\leq t}[\delta(\beta(s))]^2 = e_n(t)$ and $e(t) = t$. Then $e_n(t) \xrightarrow{p} e(t)$. Since $\beta \Rightarrow B$, where B is the regular *Wiener* process, on $D[0,1]$, we have that $(e_n, \beta) \Rightarrow (e, B)$.

We need to obtain the weak limit of

$$\sum_{0<s\leq t}\beta(s-)\left[\delta(\beta(s))\right]^2 = \int_0^t \beta(s-)de_n(s)$$

so that the asymptotic distribution of the test statistic can be calculated. It follows that $\int_0^t \beta(s-)de_n(s)$ converges in distribution to $\int_0^t B(s)de(s)$. Now, by (17) and the continuous mapping theorem, we get

$$3\int_0^t \beta^2(s-)d\beta(s) \xrightarrow{\mathcal{D}} B^3(t) - 3\int_0^t B(s)ds \tag{18}$$

for all $0 < t \leq 1$. When $t = 1$, the second term above is a normal variable with mean 0 and variance 1/3 (Ross 1983, p. 194). Also, $B(1)$ is a standard normal variate. This completes the proof of the theorem. Incidentally, the right hand side of (18) is the stochastic integral $3\int_0^t B^2(s)dB(s)$, which follows from *Itô*'s formula for continuous martingales (Chung and Williams 1990, p. 94).

Proof of Theorem 2.3

We will prove the theorem for $\sigma^2/\tau^2 = 1$. The proof for other cases is similar. Using the same decomposition of the residuals as in proof of Theorem 2.1, one can write

$$T_3 = n^{-1}\left[\sum_{i=1}^{n}\epsilon_i^2 + \sum_{i=1}^{n} 2\epsilon_i \left\{g(x_{1i}) - \hat{g}(x_{1i})\right\} + \sum_{i=1}^{n}\left\{g(x_{1i}) - \hat{g}(x_{1i})\right\}^2\right].$$

The last two terms of the above sum, when multiplied by $\sqrt{n}$, converge in probability to zero under the assumptions on b and K. The first sum, properly normalized, satisfies

$$n^{-1/2}\frac{\sum_{i=1}^{n}\left[\epsilon_i^2 - \sigma^2\right]}{\sqrt{var(\epsilon_1^2)}} \xrightarrow{\mathcal{D}} N(0,1), \tag{19}$$

where $N(0,1)$ denotes a standard normal random variable. It is known that (Hall et al. 1990)

$$\sqrt{m}\left[\hat{\tau}^2 - \tau^2\right]/c \xrightarrow{\mathcal{D}} N(0,1), \tag{20}$$

where $c^2 = \kappa\tau^4 + 2\tau^4\sum_{i=1}^{m-r}\left[\sum_{j=0}^{r} d_j d_{i+j}\right]^2$ and $\kappa = E(\eta^4)\tau^{-4} - 3$. Since the estimators of the moments of η are selected to be consistent, $a^2 \xrightarrow{P} c^2$. Now, since the ratio $m/(m+n)$ converges to θ and the two samples are independent, we combine (19) and (20) to get the desired result.

REFERENCES

Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.

Chung, K.L., and Williams, R. (1990), *Introduction to Stochastic Integration*, Birkhauser, Boston.

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

Cox, D., Koh, E., Wahba, G. and Yandell, B.S. (1988), " Testing the (Parametric) Null Model Hypothesis in (Semiparametric) Partial and Generalized Spline Models," *The Annals of Statistics,* 16, 113-119.

Durbin, J. (1973), *Distribution Theory for Tests Based on the Sample Distribution Function*, SIAM, Philadelphia.

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.

__________(1993), "Commonality of Cusum, von Neumann and Smoothing Based Goodness -of-fit Tests," *Biometrika,* 80, 89-98.

Eubank, R.L., Hart, J.D., and LaRiccia, V. (1993), "Testing Goodness of Fit via Nonparametric Function Estimation Techniques," *Communications in Statistics-Theory and Methods,* 22, 3327-3354.

Eubank, R.L., and LaRiccia, V. (1992), "Asymptotic Comparison of Cramèr-von Mises and Nonparametric Function Estimation Techniques for Testing Goodness-of-Fit," *Annals of Statistics,* 20, 2071-2086.

Eubank, R.L., and Spiegelman, C.H. (1990), "Testing Goodness of Fit of a Linear Model via Nonparametric Regression Techniques," *Journal of the American Statistical Association,* 85, 387-392.

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

Ghosh, B.K., and Huang, W. (1991), "The Power and Optimal Kernel of the Bickel-Rosenblatt Test for Goodness-of-Fit," *The Annals of Statistics,* 19, 999-1009.

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.

Jacod, J. and Shiryaev, A.N. (1987), *Limit Theorems for Stochastic Processes,* Springer-Verlag, Berlin.

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.

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

Munson, P.J., and Jernigan, R.W. (1989), "A Cubic Spline Extension of the Durbin-Watson Test," *Biometrika,* 76, 39-47.

Rice, J. (1984), "Bandwidth choice for Nonparametric Regression," *The Annals of Statistics,* 12, 1215-1230.

Ross, S.M. (1983), *Stochastic Processes,* John Wiley, New York.