Statistica Sinica **15**(2005), 135-152

TESTING LACK-OF-FIT OF PARAMETRIC REGRESSION MODELS USING NONPARAMETRIC REGRESSION TECHNIQUES

Randall L. Eubank, Chin-Shang Li and Suojin Wang

Texas A&M University, St. Jude Children's Research Hospital and Texas A&M University

Abstract: Data-driven lack-of-fit tests are derived for parametric regression models using fit comparison statistics that are based on nonparametric linear smoothers. The tests are applicable to settings where the usual bandwidth/smoothing parameter asymptotics apply to the null model, which includes testing for nonlinear models and some linear models. Large sample distribution theory is established for tests constructed from both kernel and series type estimators. Both types of smoothers are shown to give consistent tests that are asymptotically normal under the null model after appropriate centering and scaling. However, the projection nature of series smoothers results in a simplified scaling factor that produces computational savings for the associated tests.

Key words and phrases: Bandwidth selection, fit comparison test, kernel smoother, least squares, series smoother.

1. Introduction

An area of recent research interest concerns the use of nonparametric regression techniques for testing the lack-of-fit of parametric regression models. See, e.g., Aerts, Claeskens and Hart (2000), Baraud, Huet and Laurent (2003), Barry and Hartigan (1990), Cox, Koh, Wahba and Yandell (1988), Dette (1999), Eubank and Hart (1992), Fan (1996), Fan and Linton (2003), Fan, Zhang and Zhang (2001), Härdle and Mammen (1993), Hart (1997), Lee and Hart (2000) and Zheng (1998). Most of this work (with the notable exception of Müller (1992)) concerns the use of kernel or series type smoothers to assess the lack-of-fit of parametric null models for which the usual smoothing parameter asymptotics do not apply. In this paper we focus on situations where this is not the case, and smoothing parameters can be expected to exhibit standard asymptotic behavior. The large sample properties of fit comparison type tests are found to be substantially different in this setting from others that have been treated in the literature.

Assume that responses $y_1, \ldots, y_n$ are observed at nonstochastic design points $0 \le t_{1n} < \cdots < t_{nn} \le 1$. The responses and design points are related by

$$y_i = \mu(t_{in}) + e_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where μ is an unknown regression function and $e_1, \ldots, e_n$ are independent and identically distributed (i.i.d.) random errors with $E(e_1) = 0$ and $\text{var}(e_1) = \sigma^2 < \infty$. We are interested in the case where a parametric model $\mu(\cdot;\theta)$ has been postulated for μ in (1.1) and, accordingly, want to test the composite null hypothesis

$$H_0 : \mu(\cdot) = \mu(\cdot;\theta), \quad \theta \in \Theta. \tag{1.2}$$

Here θ is an unknown vector of p parameters that must be estimated and Θ is some subset of $\mathbb{R}^p$.

Let $\hat{\theta}$ be an estimator of θ_0, the true value of θ under the null model, and for $\mu_n^T(\theta) = (\mu(t_{1n};\theta), \ldots, \mu(t_{nn};\theta))$ define $\hat{\mu}_{0n} = \mu_n(\hat{\theta})$ as a parametric estimator of $\mu_{0n} = \mu_n(\theta_0)$. Then, a number of proposed tests for H_0 have been based on fit comparison statistics of the form

$$\|\hat{\mu}_n - \hat{\mu}_{0n}\|^2 = \sum_{i=1}^{n} \left\{\hat{\mu}(t_{in}) - \mu(t_{in};\hat{\theta})\right\}^2,$$

with $\hat{\mu}_n^T = (\hat{\mu}(t_{1n}), \ldots, \hat{\mu}(t_{nn}))$ representing a nonparametric fit to the data. Statistics of this nature have the intuitive appeal of providing omnibus, across the design, comparisons of a fit under the null model with a more flexible fit (i.e., $\hat{\mu}_n$) that should be closer to the true mean function when H_0 is false.

Typically $\hat{\mu}_n$ is from a linear smoother. This entails that there is an associated smoother matrix $\mathbf{S}_\lambda$, depending on a smoothing parameter λ, with

$$\hat{\mu}_n = \mathbf{S}_\lambda Y_n \tag{1.3}$$

for $Y_n^T = (y_1, \ldots, y_n)$ the response vector. The fit comparison statistic can then be expressed as

$$T_{n\lambda} = \|\mathbf{S}_\lambda Y_n - \hat{\mu}_{0n}\|^2. \tag{1.4}$$

Examples of linear smoothers that have been used in this context include kernel and orthogonal series regression estimators for which the parameter λ corresponds to the bandwidth and number of terms in the estimator, respectively.

A question of interest concerns practical methods for selecting a value for λ in (1.4). In this paper we focus on a case where it is possible to choose λ from the data in a way that produces globally consistent tests that are asymptotically distribution-free under H_0. Specifically, we treat the case where, under H_0,

$$M_n = \inf_\lambda E\|\mathbf{S}_\lambda Y_n - \mu_{0n}\|^2 \to \infty \quad \text{as} \quad n \to \infty.$$

This condition means that the selected smoother can only estimate the null mean function at a nonparametric rate. It will generally be satisfied when the null regression function is nonlinear and, in particular, not a polynomial of order less than or equal to that of the smoother. It can also be satisfied for a polynomial null model if the order of the smoother is less than the order of the polynomial. (See, e.g., Sections 3 and Example 1.2 below.)

When the condition $M_n \to \infty$ holds, it is often possible to asymptotically characterize the null mean squared error optimal choice λ_{0n} of the smoothing parameter λ to an extent that it can be estimated directly from the data with $\sqrt{n}$-consistency. This produces a data-driven choice for the level of smoothing $\hat{\lambda}_{0n}$ and our proposed test will be based on (re-centered and re-scaled versions of)

$$T_n = \|\mathbf{S}_{\hat{\lambda}_{0n}} Y_n - \hat{\mu}_{0n}\|^2. \tag{1.5}$$

To clarify the above discussion and illustrate the basic idea behind our proposed methodology it will be useful to consider the following two specific examples that will be revisited throughout the paper.

Example 1.1. Suppose that the design is uniform over $[0, 1]$ and consider testing the composite null hypothesis $H_0 : \mu(t) = 5e^{-\theta t}$, for all $t \in [0, 1]$ and some unknown $\theta > 0$. If $\mathbf{S}_\lambda$ corresponds to a boundary corrected kernel estimator with bandwidth λ and quadratic kernel $K(u) = 0.75(1 - u^2)I_{[-1,1]}(u)$ for $I_{[-1,1]}(u)$ the indicator function for the interval $[-1, 1]$, then it is known (e.g., Müller (1988) and Eubank (1999)) that, under H_0 and certain other restrictions,

$$n^{-1}E\|\mathbf{S}_\lambda Y_n - \mu_{0n}\|^2 \sim \frac{3\sigma^2}{5n\lambda} + \lambda^4 J(\mu(\cdot;\theta_0))$$

for $J(\mu(\cdot;\theta_0)) = 0.125\theta_0^3(1 - e^{-2\theta_0})$ and θ_0 the true, null value of θ. Thus, the asymptotically optimal choice of the bandwidth is $\lambda_{0n} = (3\sigma^2/20J(\mu(\cdot;\theta_0))n)^{1/5}$. The only unknowns in this bandwidth are σ and θ_0. There are numerous $\sqrt{n}$-consistent variance estimators (e.g., Rice (1984), Gasser, Sroka and Jennen-Steinmetz (1986), Hall, Kay and Titterington (1990) and Hall and Marron (1990)) and θ_0 can be estimated with $\sqrt{n}$-consistency under H_0 via nonlinear least-squares, for example. By plugging such estimators into the formula for λ_{0n} we obtain a data-driven bandwidth that can be used to compute the test statistic (1.5).

Example 1.2. Now suppose that we wish to test for a linear regression function, i.e., $H_0 : \mu(t) = \theta_1 + \theta_2(t - 0.5)$ for all $t \in [0, 1]$ and θ_1, θ_2 unknown intercept and slope coefficients. If we used a second (or higher) order smoother, such as the second order kernel estimator of Example 1.1, then under H_0, $n^{-1}E\|\mathbf{S}_\lambda Y_n - \mu_{0n}\|^2 \sim n^{-1}E\|\mathbf{S}_\lambda \varepsilon_n\|^2$ for $\varepsilon_n = (e_1, \ldots, e_n)^T$ the vector of random errors. In the case of the Example 1.1 kernel estimator, $n^{-1}E\|\mathbf{S}_\lambda \varepsilon_n\|^2 \sim 3\sigma^2/5n\lambda$, while $n^{-1}E\|\mathbf{S}_\lambda \varepsilon_n\|^2 \sim \sigma^2\lambda/n$ for a second order regression series smoother with λ terms. Consequently, the optimal bandwidth and number of terms are ∞ and 0, for second order kernel and series smoothers respectively, so $M_n \to \infty$ does not hold.

In order to have $M_n \to \infty$ one can use a first order kernel (see, e.g., Müller (1992)) or series estimator. One example of a first order series estimator is provided by cosine series regression where one regresses on the constant function and cosine functions of increasing frequency. If the design is uniform and θ_{02} is the true slope under H_0, then (cf., Section 3) one finds that, under the null model, the cosine regression smoother satisfies

$$n^{-1}E\|\mathbf{S}_\lambda Y_n - \mu_{0n}\|^2 \sim \frac{\sigma^2\lambda}{n} + \frac{2^3\theta_{02}^2}{\pi^4}\sum_{j=[(\lambda/2)+1]}^{\infty}(2j-1)^{-4},$$

and an asymptotically optimal choice for the number of terms is $\lambda_{0n} = (4n\theta_{02}^2/\pi^4\sigma^2)^{1/4}$. As in the previous example, this can be estimated by replacing the unknown parameters with $\sqrt{n}$-consistent estimators with the slope being estimated by ordinary least-squares, for example. The resulting data-driven choice for the number of terms can then be used to compute (1.5) in the case of a cosine series smoother.

The condition that $M_n \to \infty$ represents a departure from much of the literature on lack-of-fit testing using nonparametric function estimation techniques. Much of this work, such as Eubank and Spiegelman (1990) and Härdle and Mammen (1993, Proposition 1) can be viewed as smoothing residuals from parametric fits. Thus, one considers a statistic of the form

$$\tilde{T}_{n\lambda} = \|\mathbf{S}_\lambda(Y_n - \hat{\mu}_{0n})\|^2. \tag{1.6}$$

Note that $T_{n\lambda} = \tilde{T}_{n\lambda} + 2 < \mathbf{S}_\lambda(Y_n - \hat{\mu}_{0n}),\ \mathbf{S}_\lambda\hat{\mu}_{0n} - \mu_{0n} > + \|\mathbf{S}_\lambda\hat{\mu}_{0n} - \mu_{0n}\|^2$ with $< \cdot, \cdot >$ representing the Euclidean vector inner product. This expression makes $\tilde{T}_{n\lambda}$ seem somewhat simpler since it avoids the "bias" term $\|\mathbf{S}_\lambda\hat{\mu}_{0n} - \mu_{0n}\|^2$ that appears in $T_{n\lambda}$. However, under standard regularity conditions, $\tilde{T}_{n\lambda}$ behaves like $\|\mathbf{S}_\lambda\varepsilon_n\|^2$ under H_0 which means that one is essentially smoothing the zero function in the null case. The optimal level of smoothing is therefore fixed under H_0 (e.g., $\lambda = \infty$ and $\lambda = 0$ for kernel and series type smoothers as in Examples 1.1 and 1.2) so that the usual smoothing parameter asymptotics (e.g., $\lambda \to 0$ as $n \to \infty$ for kernel smoothers and $\lambda \to \infty$ as $n \to \infty$ for series estimators) will not obtain in the null case. This makes conditions which require decay of bandwidths for kernel estimators (e.g., condition (K2) of Härdle and Mammen (1993) or Theorem 2 of Eubank and Spiegelman (1990)) or growth in the number of terms of series smoothers (e.g., Theorem 1 of Eubank and Spiegelman (1990)) hard to implement from a practical viewpoint. In any case, such conditions will not be met by smoothing parameter estimators that derive from applying cross-validation type methodology to $\mathbf{S}_\lambda(Y_n - \hat{\mu}_{0n})$. Instead such stochastic choices of λ will tend to have non-degenerate limiting distributions which cause data-driven versions of $\tilde{T}_{n\lambda}$ to have non-standard and, in particular, non-normal, limiting null distributions. Examples of this can be found in Eubank and Hart (1992), Hart ((1997), Section 8.2) and Härdle and Kneip (1999), all of whom deal with the simpler situation where the parameters enter the model linearly which allows tabulation of percentage points for the tests. However, in more complicated scenarios where the null model is nonlinear in the parameters, the limiting null distribution can depend on the unknown parameters (Kuchibhatla and Hart (1996) and Hart (1997), Section 8.3.1) so that the tests are not even asymptotically distribution free in the null case. This then requires bootstrapping or some other method to obtain approximate critical values for the tests on a case-by-case basis.

In contrast to smoothing residuals or, equivalently, comparing a smoothed fit of the data to a smoothed fit of the null mean function estimator, our statistic T_n in (1.5) compares the nonparametric fit to the data with the (non-smoothed) mean function estimator under the null hypothesis. We will show that this statistic is asymptotically distribution-free in that it can be re-centered and scaled in a fashion that gives it a standard normal limiting null distribution. While some computational effort is necessary to obtain the requisite re-centering and scaling factors, this may be preferable to bootstrapping or other measures needed to obtain critical values for practical, and hence, data-driven versions of statistics like (1.6).

The remainder of the paper is organized as follows. In Section 2, we study the behavior of the data-driven statistic T_n for kernel type smoothers, while in Section 3 we develop parallel results for series type smoothers. The limiting distribution of T_n is shown to be asymptotically standard normal under H_0, after appropriate centering and scaling, for both types of smoothers. However, the projection nature of series smoothers produces a useful simplification in the re-scaling factor for T_n. Proofs and technical conditions are collected in Section 4. Empirical studies involving kernel smoother based tests in a related setting can be found in Li (2001).

2. Kernel Type Smoothers

In this section we study the case where $\mathbf{S}_\lambda$ in (1.3) corresponds to a kernel type smoother. By this we mean that there is a value $m \geq 1$ and a constant C such that if $\lambda \asymp n^{-1/(2m+1)}$ (i.e., λ is exactly of order $n^{-1/(2m+1)}$), then under H_0,

$$E\|\mathbf{S}_\lambda Y_n - \mu_{0n}\|^2 = \frac{C}{\lambda} + n\lambda^{2m} J_1(\mu(\cdot;\theta_0)) + o(n^{1/(2m+1)}) \tag{2.1}$$

for $J_1(\cdot)$ a known functional. We will be concerned with the case where $J_1(\mu(\cdot;\theta)) \neq 0$ for all θ in a neighborhood of θ_0, which entails that $\lambda_{0n} =$

$[C/(2mnJ_1(\mu(\cdot;\theta_0)))]^{1/(2m+1)}$ provides an asymptotically optimal choice for the "bandwidth" parameter λ in (2.1).

A canonical example of where (2.1) applies is for $\mathbf{S}_\lambda$ corresponding to an mth order, boundary corrected Gasser-Müller kernel smoother with $m \geq 2$ now an integer. If $K(\cdot)$ denotes the kernel for the estimator, supported on $[-1,1]$, then the estimator is of order m when the kernel satisfies $\int_{-1}^1 K(u)du = 1$, $\int_{-1}^1 u^j K(u)du = 0$, $j = 1,\ldots,m-1$, $B_m = \int_{-1}^1 u^m K(u)du \neq 0$ and $V = \int_{-1}^1 K^2(u)du < \infty$. For a design point t_{in} in $[\lambda, 1-\lambda]$ Gasser-Müller kernel weights are defined by $s_{ij\lambda} = \lambda^{-1}\int_{u_{j-1}}^{u_j} K(\lambda^{-1}(t_{in}-u))du$ with $u_j = (t_{(j+1)n}+t_{jn})/2$ and $\lambda > 0$ the kernel bandwidth. For estimation in the boundary region where $t_{in} \in [0,\lambda)$ or $t_{in} \in (1-\lambda, 1]$ these weights are modified by replacing $K((t_{in}-\cdot)/\lambda)$ with, respectively, functions $K_+(t_{in}/\lambda, (t_{in}-\cdot)/\lambda)$ and $K_-((1-t_{in})/\lambda, (t_{in}-\cdot)/\lambda)$. The boundary kernel functions K_+ and K_- satisfy the same moment conditions as the interior kernel $K(\cdot)$ except over the boundary region. More details on boundary kernels can be found in Müller ((1988), Section 5.8) and Müller (1991).

If we now take $\mathbf{S}_\lambda = \{s_{ij\lambda}\}$ for $s_{ij\lambda}$ Gasser-Müller weights with the above indicated boundary modifications, the null regression function has m continuous derivatives and $p(\cdot)$ is the design density, then it is known (e.g., Müller (1988) or Eubank (1999)) that (2.1) holds with $C = \sigma^2 V$ and $J_1(\mu(\cdot;\theta_0)) = B_m^2(\int_0^1 \mu^{(m)}(t;\theta_0)^2 p(t)dt/\ (m!)^2)$, where $\mu^{(m)}(t;\theta) = \partial^m \mu(t;\theta)/\partial t^m$. If $p(\cdot) > 0$ on $[0,1]$, then the condition $J_1(\mu(\cdot;\theta)) > 0$ is equivalent to the null regression function not being a polynomial of order m. Similar expressions hold for local polynomial regression smoothers as a result of, e.g., Müller ((1988), Section 4.6).

Another instance where (2.1) can be verified is for cubic smoothing splines with a uniform design. In that case Theorems 1-2 of Rice and Rosenblatt (1983) have the consequence that if, for example, $\mu(\cdot;\theta)$ has four continuous derivatives, and either $\mu^{(2)}(0;\theta_0) \neq 0$ or $\mu^{(2)}(1;\theta_0) \neq 0$, then (2.1) holds with $m = 2.5$, $C = 3\sigma^2/2^{7/2}$ and $J_1(\mu(\cdot;\theta_0)) = [\mu^{(2)}(0;\theta_0)^2 + \mu^{(2)}(1;\theta_0)^2]/2^{3/2}$.

We now wish to develop a test statistic for H_0 in the case where (2.1) applies. For this purpose we assume that we have available a $\sqrt{n}$-consistent estimator $\hat\theta$ for θ_0 under the null model, such that $\hat\mu_{0n} = \mu_n(\hat\theta)$ satisfies

$$\hat\mu_{0n} - \mu_{0n} = \mathbf{P}\varepsilon_n + r_n, \tag{2.2a}$$

with $\varepsilon_n^T = (e_1,\ldots,e_n), \mathbf{P}$ an $n \times n$ projection matrix that may depend on θ_0 and $r_n^T = (r_{1n},\ldots,r_{nn})$ a random vector satisfying

$$\max_{1\leq i\leq n}|r_{in}| = o_p\left(n^{-1/2}\right). \tag{2.2b}$$

Important examples of when (2.2) holds are linear and nonlinear regression where $\hat\theta = \arg\min_\theta \sum_{i=1}^n\ (y_i - \mu(t_{in};\theta))^2$. Under conditions in Seber and Wild

(1988), (2.2) is satisfied under this choice of $\hat{\theta}$ with $\mathbf{P} = X(X^TX)^{-1}X^T$ for $X = \{x_j(t_{in})\}$ and

$$x_j(t) = \frac{\partial \mu(t;\theta)}{\partial \theta_j}|_{\theta=\theta_0}. \tag{2.3}$$

In the particular case of Example 1.1 this gives $\mathbf{P} = xx^T/x^Tx$, with $x = (x_1, \ldots, x_n)^T$ for $x_i = -5t_ie^{-\theta_0 t_i}$.

We also assume here that there are estimators $\hat{\mathbf{P}}$ and $\hat{J}_1$ for $\mathbf{P}$ and $J_1(\mu(\cdot;\theta_0))$. For nonlinear least-squares an estimator of $\mathbf{P}$ can be obtained by replacing θ_0 with $\hat{\theta}$ in the formula for the projection matrix. In the case of mth order kernel smoothing we can use $\hat{J}_1 = B_m^2(\int_0^1 \mu^{(m)}(t;\hat{\theta})^2p(t)dt/(m!)^2)$. For Example 1.1, with $\hat{\theta}$ obtained from nonlinear least-squares, this produces the estimators $\hat{J}_1 = 0.125\hat{\theta}^3(1-e^{-2\hat{\theta}})$, $\hat{\mathbf{P}} = \hat{x}\hat{x}^T/\hat{x}^T\hat{x}$ with $\hat{x} = -5(t_1e^{-\hat{\theta}t_1}, \ldots, t_ne^{-\hat{\theta}t_n})^T$.

For the smoothing parameter λ associated with the nonparametric estimator we require that there is an estimator $\hat{\lambda}_{0n}$ of λ_{0n} with the property

$$(\hat{\lambda}_{0n} - \lambda_{0n})/\lambda_{0n} = O_p(n^{-1/2}). \tag{2.4}$$

One estimator which satisfies (2.4) in certain cases is $\hat{\lambda}_{0n} = [\hat{C}/(2mn\hat{J}_1)]^{1/(2m+1)}$ with $\hat{C}$ a $\sqrt{n}$-consistent estimator of C in (2.1). In particular, for kernel smoothing we can use $\hat{\lambda}_{0n} = [\hat{\sigma}^2V/2mn\hat{J}_1]^{1/(2m+1)}$ for the choice of $\hat{J}_1$ described above and $\hat{\sigma}$, for example, the Gasser, Sroka and Jennen-Steinmetz (1986) estimator of σ.

The specific test statistic we now propose for H_0 is

$$T_n = \|\hat{\mathbf{S}}Y_n - \hat{\mu}_{0n}\|^2 \tag{2.5}$$

with $\hat{\mathbf{S}} = \mathbf{S}_{\hat{\lambda}_{0n}}$ and $\hat{\lambda}_{0n}$ as in (2.4). The asymptotic properties of T_n follow from results in Section 4 where we show that $T_n - \sigma^2\mathrm{tr}(\hat{\mathbf{S}} - \hat{\mathbf{P}})^T(\hat{\mathbf{S}} - \hat{\mathbf{P}}) - n\hat{\lambda}_{0n}^{2m}\hat{J}_1$ can be approximated by

$$\|(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\varepsilon_n\|^2 - \sigma^2\mathrm{tr}(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{S}_{\lambda_{0n}} - \mathbf{P}) - 2\langle(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\varepsilon_n, (\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mu_{0n}\rangle. \tag{2.6}$$

Expression (2.6) is a sum of a linear and (re-centered) quadratic form in ε_n which has asymptotic variance $2\sigma^4\mathrm{tr}[(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})]^2 + 4\sigma^2\|(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mu_{0n}\|^2$. Similar to condition (2.1), we assume that there is a functional $J_2(\cdot)$ such that if $\lambda \asymp n^{-1/(2m+1)}$ we have

$$\|(\mathbf{S}_\lambda - \mathbf{P})^T(\mathbf{I} - \mathbf{S}_\lambda)\mu_{0n}\|^2 = n\lambda^{2m}J_2(\mu(\cdot;\theta_0)) + o(n\lambda^{2m}) \tag{2.7}$$

under H_0, and that there is an estimator $\hat{J}_2$ available for $J_2(\mu(\cdot;\theta_0))$.

In the case of kernel regression with Gasser-Müller weights and nonlinear least-squares estimation of θ one finds that the functional J_2 in (2.7) has the form

$$J_2(g) = \frac{B_m^2}{(m!)^2} \int_0^1 \left\{ g^{(m)}(t) - \tilde{g}^{(m)}(t) \right\}^2 p(t)dt$$

for

$$\tilde{g}^{(m)}(\cdot) = \left\{ \int_0^1 g^{(m)}(z)x(z)p(z)dz \right\}^T \left\{ \int_0^1 x(z)x^T(z)p(z)dz \right\}^{-1} x(\cdot)$$

with $x(\cdot)^T = (x_1(\cdot), \ldots, x_p(\cdot))$ as in (2.3), and we can take $\hat{J}_2 = J_2(\mu(\cdot;\hat{\theta}))$. In the particular case of Example 1.1 this produces the estimator

$$\begin{aligned}\hat{J}_2 &= 0.0025 \int_0^1 \{5\hat{\theta}^2 e^{-\hat{\theta}t} - (\int_0^1 25\hat{\theta}^2 s e^{-2\hat{\theta}s}\, ds)(\int_0^1 25 s^2 e^{-2\hat{\theta}s}\, ds)^{-1} 5te^{-\hat{\theta}t}\}^2\, dt \\ &= \hat{J}_1 - 0.0625\hat{\theta}^3 \{1 - (1+2\hat{\theta})e^{-2\hat{\theta}}\}^2 / \{1 - (1 + 2\hat{\theta} + 2\hat{\theta}^2)e^{-2\hat{\theta}}\}.\end{aligned}$$

Our principle asymptotic result for T_n can now be stated as follows.

Theorem 1. *Let $\hat{\sigma}^2$ be a $\sqrt{n}$-consistent estimator of σ^2 and assume that* (2.2), (2.4), (2.7) *and conditions* (A1)−(A6) *in Section* 4.1 *are satisfied. Define*

$$V_n = \frac{T_n - \hat{\sigma}^2 tr(\hat{\mathbf{S}} - \hat{\mathbf{P}})^T(\hat{\mathbf{S}} - \hat{\mathbf{P}}) - n\hat{\lambda}_{0n}^{2m}\hat{J}_1}{\left[2\hat{\sigma}^4 tr\left\{ (\hat{\mathbf{S}} - \hat{\mathbf{P}})^T(\hat{\mathbf{S}} - \hat{\mathbf{P}}) \right\}^2 + 4\hat{\sigma}^2 n\hat{\lambda}_{0n}^{2m}\hat{J}_2 \right]^{1/2}}. \tag{2.8}$$

Then, $V_n \xrightarrow{d} N(0,1)$ under H_0 and, for any $\alpha \in (0,1)$, the test obtained by rejecting H_0 if $V_n > z_\alpha$, for z_α the $100(1-\alpha)$th percentile of the standard normal distribution, has asymptotic significance level α.

Under some additional regularity restrictions one may show that the test based on V_n in (2.8) is consistent against fixed alternatives. For this we need a parallel of condition (2.2) to hold for the alternative model wherein the parameter vector θ_0 now represents a value that gives an approximation to the true regression function from the functions in $\{\mu(\cdot;\theta); \theta \in \Theta\}$. For example, if $\hat{\theta}$ is a nonlinear least-squares estimator we can take θ_0 to be the minimizer over Θ of $\int_0^1 (\mu(t) - \mu(t;\theta))^2 p(t)dt$. One then finds that for a specific (fixed) alternative vector μ_n, V_n will behave asymptotically like a constant multiple of $n^{-1/(4m+2)}\|\mu_n - \mu_n(\theta_0)\|^2$, which diverges at the rate $n^{(4m+1)/(4m+2)}$. This indicates an ability to detect local alternatives that converge only at nonparametric rates (e.g., $n^{-9/20}$ for the second order kernel case of Example 1.1) and could be perceived as a drawback when these tests are compared to others capable of detecting local alternatives of orders $1/\sqrt{n}$. However, this viewpoint appears to be incorrect in general since, in certain cases, tests which detect local alternatives converging at nonparametric rates have been shown to be more efficient asymptotically than others that can detect alternatives of order $1/\sqrt{n}$. See, e.g, Inglot and Ledwina (1996) and Eubank (2000).

The following corollary states that the conclusions of Theorem 1 apply to the test statistic V_n in (2.8) in the specific case of kernel regression smoothing and nonlinear least-squares estimation of θ_0.

Corollary 1. *Assume that $\hat{\theta}$ satisfies* (2.2) *for* $\mathbf{P}$*, the projection matrix obtained from* (2.3)*, and that conditions* (B1)$-$(B8) *of Section* 4.2 *are satisfied. Then the results in Theorem* 1 *apply to V_n in the case where the nonparametric smoother is obtained from boundary corrected kernel estimation.*

Test statistics similar in form to V_n in the kernel case have been considered by Härdle and Mammen (1993) in the case of random designs and nonstochastic (i.e., not data-driven) choices for the bandwidth. Somewhat more closely related to our work is that of Müller (1992) who uses a data estimated bandwidth like $\widehat{\lambda}_{0n}$ to develop pointwise diagnostic tests with kernel estimators. In contrast to Müller's procedure our test is global being, in effect, an average across the design of Müller's pointwise tests with a global, rather than local, bandwidth. We note however that Müller's results do not imply ours, and conversely.

3. Series Type Estimators

In this section we study the properties of T_n in (1.5) for cases where the smoother matrix $\mathbf{S}_\lambda$ is from a series type regression estimator. Thus, we take λ to be a positive integer and define

$$\mathbf{S}_\lambda = n^{-1} \sum_{j=1}^{\lambda} \Phi_{jn}\Phi_{jn}^T, \qquad 1 \le \lambda \le n, \tag{3.1}$$

for $\Phi_{jn}^T = (\phi_{jn}(t_{1n}), \ldots, \phi_{jn}(t_{nn}))$, and $\phi_{jn}(\cdot), j = 1, \ldots, n$, functions that satisfy

$$\langle \Phi_{jn}, \Phi_{kn} \rangle = n\delta_{jk} \tag{3.2}$$

for $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ if $j \ne k$. Examples of smoothers which can be treated in this fashion include polynomial and trigonometric regression estimators. The orthogonality conditions imposed here are for theoretical developments, and the test statistic below can be computed using standard least-squares methodology regardless of whether or not (3.2) holds.

In Example 1.2 we discussed the use of cosine series regression under a uniform design $t_{in} = (2i - 1)/2n$, $i = 1, \ldots, n$. In that instance we take $\phi_{1n}(t) = 1$ and $\phi_{jn}(t) = \sqrt{2}\cos\{\pi(j - 1)t\}$, $j = 2, \ldots, n$, and condition (3.2) is satisfied due to Lemma 3.4 in Eubank (1999).

For series estimators we can no longer expect condition (2.1) to hold since the rate of decay for the null model mean squared error is linked to the rate of decay of the null Fourier coefficients $n^{-1}\langle\Phi_{jn},\mu_{0n}\rangle$ (Cox (1988) and Eubank and Jayasuriya (1993)). Instead we assume that there exists a sequence of integers λ_{0n} which satisfies

$$\lambda_{0n} \asymp \langle\mu_{0n},(\mathbf{I}-\mathbf{S}_{\lambda_{0n}})\mu_{0n}\rangle, \tag{3.3a}$$

$$\lambda_{0n}\to\infty \text{ and } \lambda_{0n}/n^{\delta}\to 0 \text{ for some } \delta<1. \tag{3.3b}$$

Conditions of this type can be shown to hold under various assumptions about the rate of decay of the null Fourier coefficients.

In the specific case of cosine series regression with a uniform design one finds that (see, e.g., Eubank (1999), Section 3.4) if $\mu(\cdot;\theta_0)\in C^1[0,1]$, then

$$n^{-1}\langle\Phi_{jn},\mu_{0n}\rangle = a_j(\theta_0)+O(n^{-1}) \tag{3.4a}$$

uniformly in $j=1,\ldots,n$ for $a_1(\theta_0)=\int_0^1\mu(t;\theta_0)\,dt$ and $a_j(\theta_0)=\sqrt{2}\int_0^1\mu(t;\theta_0)$ $\cos\{(j-1)\pi t\}dt$, $j=2,\ldots,n$. Condition (3.3) can then be established by computing integers λ_{0n} that asymptotically minimize the null mean squared error. A simple but important special case occurs when the $a_j(\theta_0)$ exhibit algebraic decay in the sense that

$$a_j(\theta_0)=C(\theta_0)j^{-\tau}+o(j^{-\tau}) \tag{3.4b}$$

uniformly in j for some $\tau\geq 1$ and $C(\cdot)$ is a function with $C(\theta_0)\neq 0$. When (3.4) holds, we can take $\lambda_{0n}=[\{nC^2(\theta_0)/\sigma^2\}^{1/2\tau}]$ and this choice satisfies (3.3).

For the linear regression function $\mu(t;\theta_0)=\theta_{01}+\theta_{02}(t-0.5)$ of Example 1.2, the even frequency null cosine Fourier coefficients vanish. The remaining coefficients exhibit algebraic decay, given explicitly by $a_{2j-1}(\theta_0)=-(2^{3/2}\theta_{02})/(\pi^2(2j-1)^2)$ for θ_{02}, the null model slope coefficient. Combining this with the previous discussion on algebraic decay produces the asymptotically mean squared error optimal number of terms for fitting the linear model that was given in Section 1.

We now wish to develop a version of the V_n test statistic in (2.8) for the series estimator setting. As in Section 2 we assume that (2.2) holds and, since $\mathbf{P}$ is a projection operator, we write $\mathbf{P}=n^{-1}\sum_{j=1}^{p}x_{jn}x_{jn}^T$ for n-vectors x_{jn} satisfying $\langle x_{jn},\, x_{kn}\rangle=n\delta_{jk}$, $j,k=1,\ldots,p$. The x_{jn} are required to satisfy certain mild smoothness conditions discussed in Section 4.3. For example, in the particular case of cosine series regression with a uniform design and $\hat\theta$ from nonlinear least-squares, it is enough to assume that the functions in (2.3) satisfy $x_k(\cdot)\in C^1[0,1]$, $k=1,\ldots,p$, and that there exists finite constants $A_1,\ldots,A_p$ such that

$$\left|\int_0^1\cos(j\pi t)x_k(t)dt\right|<A_kj^{-\eta} \tag{3.5}$$

for some $\eta \geq 1$. This condition holds, for example, with $\eta = 2$ when x'_k is absolutely continuous with x''_k square integrable. In the case of the linear regression function of Example 1.2, parameter estimation reduces to ordinary least-squares with $x_1(t) = 1$ and $x_2(t) = (t - 0.5)$. Conditions (3.4)−(3.5) are then satisfied with $\tau = \eta = 2$.

The sequence λ_{0n} in (3.3) will be assumed to satisfy (2.4) for some estimator sequence $\hat{\lambda}_{0n}$. Setting $\hat{\mathbf{S}} = \mathbf{S}_{\hat{\lambda}_{0n}}$, we then require that there is an estimator $\hat{J}$ such that

$$\langle \mu_{0n}, (\mathbf{I} - \hat{\mathbf{S}})\mu_{0n} \rangle - \hat{J} = o_p(\sqrt{\lambda_{0n}}). \tag{3.6}$$

Under some additional restrictions on $\hat{\theta}$ one may show that $\hat{J} = \langle \hat{\mu}_{0n}, (\mathbf{I} - \hat{\mathbf{S}})\hat{\mu}_{0n} \rangle$ satisfies (3.6). However, there are often more direct estimators. For example, in the cosine regression case with algebraically decaying Fourier coefficients, we can use $\hat{J} = nC^2(\hat{\theta}) \sum_{j=\hat{\lambda}_{0n}+1}^{n} j^{-2\tau}$ in estimating λ_{0n}. This then produces the estimator $\hat{\lambda}_{0n} = [\{nC^2(\hat{\theta})/\hat{\sigma}^2\}^{1/2\tau}]$ for the null mean squared error optimal number of terms, with $\hat{\sigma}^2$ some $\sqrt{n}$-consistent estimator of σ^2.

Our principle result concerning $T_n = \|\hat{\mathbf{S}}Y_n - \hat{\mu}_{0n}\|^2$ can be stated as follows.

Theorem 2. *Assume that conditions* (2.2), (2.4), (3.3), (3.6) *and* (S1)−(S4) *of Section* 4.3 *hold, and that $\hat{\sigma}^2$ is a $\sqrt{n}$-consistent estimator of σ^2. Then,*

$$V_n = \frac{T_n - \hat{\sigma}^2 \hat{\lambda}_{0n} - \hat{J}}{\hat{\sigma}^2 \sqrt{2\hat{\lambda}_{0n}}} \xrightarrow{d} N(0, 1). \tag{3.7}$$

An asymptotic α level test for H_0 is then obtained by rejecting the null hypothesis if $V_n > z_\alpha$.

Comparison of Theorems 1 and 2 reveals that the centering and scaling factors for T_n are simpler in the series estimator setting. The most important difference is that the series estimator statistic does not involve the J_2 functional in (2.7). This occurs because the linear term in (2.6) now satisfies $\langle (\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\varepsilon_n, (\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mu_{0n} \rangle = -\langle (\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mathbf{P}\varepsilon_n, (\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mu_{0n} \rangle$ and, as a result, the projection property of the series smoother allows it to simultaneously exploit smoothness in both $\mathbf{P}$ and μ_{0n} to make this term negligible.

Arguments similar to those for the kernel case can be used to establish consistency for tests with series estimators. For a fixed alternative V_n behaves like a constant multiple of $n\lambda_{0n}^{-1/2}$, which diverges as a result of (3.3b).

As an application of Theorem 2 we state the following result for cosine series regression with a uniform design and nonlinear least-squares estimation of θ_0.

Corollary 2. *Assume that $\mu(\cdot;\theta_0)$, $x_1, \ldots, x_p$ are all in $C^1[0,1]$ and that* (3.4)−(3.5) *hold with $(\eta - 1/4)/\tau > 1/2$. Then if the design is uniform, $C(\cdot)$ is differen-*

tiable in a neighborhood of θ_0 *and* $Ee_1^4 < \infty$, *the conclusions of Theorem* 2 *apply to the cosine series smoother.*

4. Proofs

The basic approach to proving both Theorems 1 and 2 begins with the use of (2.2) to obtain

$$T_n = \|\hat{\mu}_n - \mu_{0n}\|^2 - 2\langle \hat{\mu}_n - \mu_{0n}, \hat{\mu}_{0n} - \mu_{0n}\rangle + \|\mathbf{P}\varepsilon_n\|^2 + O_p(1). \tag{4.1}$$

The proofs of Theorems 1-2 then contain analyses of the terms in (4.1) that specialize to the specific type of smoother under consideration.

4.1. Proof of Theorem 1.

To establish Theorem 1 we require the following assumptions.

(A1) There is a value $m \geq 1$ and constants $C_i > 0$, $i = 1, 2, 3$, such that if $\lambda \asymp n^{-1/(2m+1)}$, then

(a) $\mathrm{tr}\mathbf{S}_\lambda^T\mathbf{S}_\lambda \sim C_1/\lambda$;

(b) $\mathrm{tr}\{(\mathbf{S}_\lambda - \mathbf{P})^T(\mathbf{S}_\lambda - \mathbf{P})\}^j \sim C_{j+1}/\lambda$, $j = 1, 2$;

(c) $\|(\mathbf{I} - \mathbf{S}_\lambda)\mu_{0n}\|^2 = n\lambda^{2m}J_1(\mu(\cdot;\theta_0)) + o(n^{1/(2m+1)})$.

(A2) For λ_{0n} and $\hat{\lambda}_{0n}$ satisfying (2.4) and $\hat{\mathbf{S}} = \mathbf{S}_{\hat{\lambda}_{0n}}$:

(a) $\|(\hat{\mathbf{S}} - \mathbf{P})\varepsilon_n\|^2 - \sigma^2 \text{ tr } (\hat{\mathbf{S}} - \mathbf{P})^T(\hat{\mathbf{S}} - \mathbf{P}) - \|(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\varepsilon_n\|^2 + \sigma^2 \text{ tr } (\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{S}_{\lambda_{0n}} - \mathbf{P}) = o_p\left(n^{1/(4m+2)}\right)$;

(b) $\langle(\hat{\mathbf{S}} - \mathbf{P})\varepsilon_n, (\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}\rangle - \langle(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\varepsilon_n, (\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mu_{0n}\rangle = o_p\left(n^{1/(4m+2)}\right)$;

(c) $\mathrm{tr}\Big\{(\hat{\mathbf{S}} - \mathbf{P})^T(\hat{\mathbf{S}} - \mathbf{P})\Big\}^j - \mathrm{tr}\,\Big\{(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\Big\}^j = o_p\left(n^{1/(4m+2)}\right)$, $j = 1, 2$;

(d) $\|(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mu_{0n}\|^2 = n\lambda_{0n}^{2m}J_2(\mu(\cdot;\theta_0)) + o_p\left(n^{1/(4m+2)}\right)$;

(e) $\|(\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}\|^2 = n\hat{\lambda}_{0n}^{2m}J_1(\mu(\cdot;\theta_0)) + o_p\left(n^{1/(4m+2)}\right)$;

(f) parallels of conditions (a)$-$(d) continue to hold if $\mathbf{P} = \mathbf{0}$.

(A3) There exists $\hat{\mathbf{P}}$ such that

(a) $\mathrm{tr}(\hat{\mathbf{S}} - \hat{\mathbf{P}})^T(\hat{\mathbf{S}} - \hat{\mathbf{P}}) - \mathrm{tr}(\hat{\mathbf{S}} - \mathbf{P})^T(\hat{\mathbf{S}} - \mathbf{P}) = o_p\left(n^{1/(4m+2)}\right)$;

(b) $\mathrm{tr}\Big\{(\hat{\mathbf{S}} - \hat{\mathbf{P}})^T(\hat{\mathbf{S}} - \hat{\mathbf{P}})\Big\}^2 - \mathrm{tr}\Big\{(\hat{\mathbf{S}} - \mathbf{P})^T(\hat{\mathbf{S}} - \mathbf{P})\Big\}^2 = o_p\left(n^{1/(2m+1)}\right)$.

(A4) There exist $\hat{J}_1$ and $\hat{J}_2$ such that $\hat{J}_1 - J_1(\mu(\cdot;\theta_0)) = o_p\left(n^{-1/(4m+2)}\right)$ and $\hat{J}_2 - J_2(\mu(\cdot;\theta_0)) = o_p(1)$.

(A5) Set $\mathbf{W} = \{w_{ij}\} = (\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})$. Then, $\sum_{j=1}^n \left(e_j^2 - \sigma^2\right) w_{jj} = o_p\left(n^{1/(4m+2)}\right)$.

(A6) Define $L_n = (\ell_1, \ldots, \ell_n)^T = (\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T (\mathbf{I} - \mathbf{S}_{\lambda_{0n}}) \mu_{0n}$, $W(n) = 2\sum_{k=1}^{n}\sum_{j=1}^{k-1} w_{kj}e_k e_j$ and $\sigma^2(n) = 2\sigma^4 \operatorname{tr}\{\mathbf{W}^2\} + 4\sigma^2\|L_n\|^2$. Then,

(a) $\max_{1\le k\le n}\sum_{j=1}^{n} w_{kj}^2/\sigma^2(n) \to 0$;

(b) $\sum_{j=1}^{n} \ell_j^4/\sigma^4(n) \to 0$;

(c) $EW(n)^4/[\mathrm{var}W(n)]^2 \to 3$;

(d) $\sum_{j=1}^{n}(\sum_{k=j+1}^{n} w_{kj}\ell_j)^2/\sigma^4(n) \to 0$.

Assumption (A1) is sufficient to imply the risk behavior stated in (2.1). Assumptions (A5)$-$(A6) are required to establish asymptotic normality for the random variable $T(n) = \langle \varepsilon_n, \mathbf{W}\varepsilon_n\rangle - 2L_n^T\varepsilon_n$. Specifically, under (A6) we have the following lemma.

Lemma 1. $(W(n) - 2L_n^T\varepsilon_n)/\sigma(n) \xrightarrow{d} N(0,1)$ *if* (A6) *holds.*

Proof. Define $Q_{kn} = 2\sum_{j=1}^{k-1} w_{kj}e_j/\sigma(n)$, $L_{kn} = -2\ell_k/\sigma(n)$ and $u_{kn} = e_k(Q_{kn} + L_{kn})$. Then, as a result of Heyde and Brown (1970), the Lemma will be verified once we show that $\sum_{k=1}^{n} E|u_{kn}|^4 \to 0$ and $E|\sum_{k=1}^{n} E\left(u_{kn}^2|e_1,\ldots,e_{k-1}\right) - 1|^2 \to 0$. The first of these two conditions follows from (A6b) and de Jong (1987), who shows that (A6a) and (A6c) imply that $\sum_{k=1}^{n}EQ_{kn}^4/\sigma^4(n) \to 0$. The second condition is equivalent to $E\left(\sum_{k=1}^{n}E\left(u_{kn}^2|e_1,\ldots,e_{k-1}\right)\right)^2 = 1 + o(1)$. Thus, if we write $E\left(u_{kn}^2|e_1,\ldots,e_{k-1}\right) = \sigma^2(Q_{kn}^2 + 2L_{kn}Q_{kn} + L_{kn}^2)$ and sum by parts, (A6d) gives $E(\sum_{k=1}^{n} L_{kn}Q_{kn})^2 = o(1)$. Consequently, $E\left(\sum_{k=1}^{n} E\left(u_{kn}^2|e_1,\ldots,e_{k-1}\right)\right)^2 = \sigma^4 E(\sum_{k=1}^{n}\{Q_{kn}^2 + L_{kn}^2\})^2 + o(1)$, and the Lemma now follows from results in de Jong (1987) where it is shown that $\sigma^4(n)E(\sum_{k=1}^{n} Q_{kn}^2)^2/[\mathrm{var}W(n)]^2 \to 1$ when (A6a) and (A6c) hold.

To prove Theorem 1 we begin by using (2.4), (A1), (A2e) and (A2f) to obtain $\|\hat{\mu}_n - \mu_{0n}\|^2 = O_p\left(n^{1/(2m+1)}\right)$. Thus, by the Cauchy-Schwarz inequality and (2.2), $\langle\hat{\mu}_n - \mu_{0n}, \hat{\mu}_{0n} - \mu_{0n}\rangle = \langle\hat{\mu}_n - \mu_{0n}, \mathbf{P}\varepsilon_n\rangle + o_p\left(n^{1/(4m+2)}\right)$. Using (A2e) leads to

$$T_n = \|(\hat{\mathbf{S}} - \mathbf{P})\varepsilon_n\|^2 - 2\langle(\hat{\mathbf{S}} - \mathbf{P})\varepsilon_n, (\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}\rangle + n\hat{\lambda}_{0n}^{2m}J_1(\mu(\cdot,\theta_0)) + o_p\left(n^{1/(4m+2)}\right).$$

Therefore, by (A2a)$-$(A2c), (A3a), (A4) and (A5), we obtain that

$$\begin{aligned}
&T_n - \sigma^2\mathrm{tr}(\hat{\mathbf{S}} - \hat{\mathbf{P}})^T(\hat{\mathbf{S}} - \hat{\mathbf{P}}) - n\hat{\lambda}_{0n}^{2m}\hat{J}_1 \\
&= \|(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\varepsilon_n\|^2 - \sigma^2\mathrm{tr}(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{S}_{\lambda_{0n}} - \mathbf{P}) \\
&\quad -2\langle(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\varepsilon_n, (\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mu_{0n}\rangle + o_p\left(n^{1/(4m+2)}\right) \\
&= \sum_{i\neq j} w_{ij}e_ie_j - 2\sum_{j=1}^{n}\ell_je_j + o_p\left(n^{1/(4m+2)}\right).
\end{aligned}$$

Since by (A1b) and (A2d) $\sigma^2(n) = 2\,\sigma^4\,C_3\,/\,\lambda_{0n} + 4\sigma^2\,n\,\lambda_{0n}^{2m}\,J_2\,(\mu\,(\cdot;\,\theta_0)) + o(n^{1/(2m+1)})$, we now see from Lemma 1 that $\{T_n - \sigma^2\text{tr}(\hat{\mathbf{S}} - \hat{\mathbf{P}})^T(\hat{\mathbf{S}} - \hat{\mathbf{P}}) - n\hat{\lambda}_{0n}^{2m}\hat{J}_1\}/\sigma(n) \xrightarrow{d} N(0,1)$. Finally, $\sigma^2(n)$ may be consistently estimated by $2\sigma^4\text{tr}\{(\hat{\mathbf{S}}-\hat{\mathbf{P}})^T(\hat{\mathbf{S}} - \hat{\mathbf{P}})\}^2 + 4\sigma^2\hat{\lambda}_{0n}^{2m}\hat{J}_2$ due to (A2c), (A2d), (A3b) and (A4).

4.2. Proof of Corollary 1.

The proof of Corollary 1 is quite tedious, but parallels arguments in Eubank and Wang (1994). We therefore provide only a brief sketch of the proof here, along with an explicit set of conditions (B1)−(B8) that are sufficient to imply (A1)−(A6) in this case. A more detailed proof is available on request from the authors. See also Li (1999) for a detailed proof in a related scenario.

(B1) For each $q \in [0,1]$ the support of $K_+(q,\cdot)$ is $[-1,q]$ and that of $K_-(q,\cdot)$ is $[-q,1]$.

(B2) $K_+(q,\cdot) \in \mathcal{C}_m(-1,q)$ and $K_-(q,\cdot) \in \mathcal{C}_m(-q,1)$ where, for finite constants c_1 and c_2, $\mathcal{C}_m(c_1,c_2)$ represents the class of all mth order, Lipschitz continuous kernels with support on $[c_1,c_2]$.

(B3) $\sup_{q\in[0,1]} |K_\pm(q,x_1) - K_\pm(q,x_2)| \le L_1|x_1 - x_2|$ and $\sup_{x\in[-1,1]} |K_\pm(q_1,x) - K_\pm(q_2,x)| \le L_2|q_1 - q_2|$ for some finite constants $L_1,\ L_2 > 0$.

(B4) $K_+(1,\cdot) = K_-(1,\cdot) = K(\cdot)$.

(B5) For each $\theta \in \Theta$, $\mu^{(m)}(t;\theta)$ is Lipschitz continuous in t.

(B6) $(\partial^m/\partial t^m)(\partial/\partial\theta)\mu(t;\theta)$ is continuous on $[0,1] \times \Theta$.

(B7) $E|e_1|^{8m+4+\nu} < \infty$ for some $\nu > 0$.

(B8) The design points satisfy $\int_0^{t_{in}} p(t)dt = i/n$, $i = 1,\ldots,n$, for a continuously differentiable density $p(\cdot)$ on $[0,1]$.

To prove Corollary 1, first observe that under (B1)−(B8) known properties of the bias and variance of kernel smoothers can be used to establish (A1a), (A1b), (A2d)−(A2e) and to see that if $g \in C^r[0,1]$ for $1 \le r < m$, then $\mathbf{S}_\lambda g_n = g_n + o(\lambda^r)$ for $g_n^T = (g(t_{1n}),\ldots,g(t_{nn}))$ and any $\lambda = \lambda_{0n} + o(\lambda_{0n})$. Similarly, one finds the ith element of $\mathbf{S}_\lambda^T g_n$ is $g(t_{in}) + o(\lambda^r)$ for $t_{in} \in [\lambda, 1-\lambda]$ and is $O(1)$ otherwise if $\lambda = \lambda_{0n} + o(\lambda_{0n})$. These facts along with (B6) give (A1c).

For verification of (A2a), it suffices, as in Eubank and Wang (1994), to work on the set $\Lambda_n = \{\lambda : |(\lambda - \lambda_{0n})/\lambda_{0n}| \le n^{-\gamma}\}$ for $\gamma < 1/2$ since $P(\hat{\lambda}_{0n} \in \Lambda_n) \to 1$ as a result of (2.4). For $\lambda \in \Lambda_n$, (B1)−(B4) may be used to check that $\|(\mathbf{S}_\lambda - \mathbf{P})\varepsilon_n\|^2 - \|(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\varepsilon_n\|^2 + \sigma^2$ tr $[(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{S}_{\lambda_{0n}} - \mathbf{P}) - (\mathbf{S}_\lambda - \mathbf{P})^T(\mathbf{S}_\lambda - \mathbf{P})]$ is dominated by $\|\mathbf{S}_\lambda\varepsilon_n\|^2 - \|\mathbf{S}_{\lambda_{0n}}\varepsilon_n\|^2 + \sigma^2$ tr $[\mathbf{S}_{\lambda_{0n}}^T\mathbf{S}_{\lambda_{0n}} - \mathbf{S}_\lambda^T\mathbf{S}_\lambda]$. Then one finds that for $\lambda,\ \lambda' \in \Lambda_n$, $\mathbf{S}_{\lambda'}^T\mathbf{S}_{\lambda'} - \mathbf{S}_\lambda^T\mathbf{S}_\lambda$ is $O(n\lambda)$ banded with nonzero elements that are uniformly $O(|\lambda - \lambda'|/\lambda)$, and the remainder of the argument proceeds as in the proof of Lemma 1 of Eubank and Wang (1994) using a Härdle, Hall and Marron (1988) type partitioning argument for Λ_n along with inequality (8) of Whittle (1960) and the moment condition (B7). Conditions (A2b)−(A2c) are verified similarly.

For conditions (A5)−(A6), (B6) is seen to imply that $\|(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})\varepsilon_n\|^2 - \sigma^2\mathrm{tr}(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})^T(\mathbf{S}_{\lambda_{0n}} - \mathbf{P})$ is dominated by $\|\mathbf{S}_{\lambda_{0n}}\varepsilon_n\|^2 - \sigma^2\mathrm{tr}\mathbf{S}_{\lambda_{0n}}^T\mathbf{S}_{\lambda_{0n}}$ and the trace of $\mathbf{S}_{\lambda_{0n}}^T\mathbf{S}_{\lambda_{0n}}$ is found to be asymptotic to $\lambda_{0n}^{-1}\int_{-2}^{2} K^*(u)^2du$ with $K^*(\cdot)$ the convolution kernel $\int_{-1}^{1} K(z)K(\cdot - z)dz$. Using these results along with the Lemma 1, (A5)−(A6) can now be verified directly. Finally, (A4) is easily seen to hold.

4.3. Proof of Theorem 2

In this section we give a proof of Theorem 2. For this purpose we first specify the conditions that are needed for the result to hold.

(S1) $\|(\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}\|^2 = \|(\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mu_{0n}\|^2 + o_p(\sqrt{\lambda_{0n}})$.

(S2) $\|\hat{\mathbf{S}}\varepsilon_n\|^2 - \|\mathbf{S}_{\lambda_{0n}}\varepsilon_n\|^2 - \sigma^2(\hat{\lambda}_{0n} - \lambda_{0n}) = o_p(\sqrt{\lambda_{0n}})$.

(S3) Let $\mathbf{P} = n^{-1}\mathbf{X}\mathbf{X}^T$ for $\mathbf{X}^T = (x_{1n}, \ldots, x_{pn})$ with $\langle x_{in},\, x_{jn}\rangle = n\delta_{ij},\ i,\ j = 1, \ldots, p$. Then,
(a) $\langle x_{jn},\, (\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}\rangle = \langle x_{jn},\, (\mathbf{I} - \mathbf{S}_{\lambda_{0n}})\mu_{0n}\rangle + o_p(\sqrt{\lambda_{0n}})$;
(b) $\langle x_{jn},\, (\mathbf{I} - \mathbf{S}_{\lambda_{0n}})x_{jn}\rangle = o(n)$ for $j = 1, \ldots, p$.

(S4) Let $\mathbf{S}_{\lambda_{0n}} = \{s_{ij}\}$ and set $Q_n = \sum\sum_{i\neq j} s_{ij}e_ie_j$. Then,
(a) $\sum_{i=1}^{n} s_{ii}^2 = o(\lambda_{0n})$;
(b) $EQ_n^4/(2\sigma^4\lambda_{0n})^2 \to 3$.

Condition (S3b) is a mild smoothness condition on the elements of $\mathbf{P}$. For nonlinear least-squares this reduces essentially to requiring that the functions in (2.3) have convergent series expansions. Condition (S4) is needed to satisfy conditions in de Jong (1987) for asymptotic normality of the quadratic form that dominates T_n. It implicitly requires the existence of a fourth moment for e_1. Condition (S4a) can be established using bounds on the basis functions in (3.1) and is true, for example, if the $\phi_{jn}(\cdot)$ are uniformly bounded.

To prove Theorem 2 we begin again with expansion (4.1). Writing $\hat{\mu}_n = \hat{\mathbf{S}}\varepsilon_n + \hat{\mathbf{S}}\mu_{0n}$ and using $\hat{\mathbf{S}}^2 = \hat{\mathbf{S}}$, we obtain

$$T_n = \|\hat{\mathbf{S}}\varepsilon_n\|^2 + \|(\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}\|^2 - 2\langle\hat{\mathbf{S}}\varepsilon_n, \hat{\mu}_{0n} - \mu_{0n}\rangle + 2\langle(\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}, \hat{\mu}_{0n} - \mu_{0n}\rangle + O_p(1).$$

Using (2.2) we have $\langle(\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}, \hat{\mu}_{0n} - \mu_{0n}\rangle = \langle(\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}, \mathbf{P}\varepsilon_n\rangle + \langle(\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}, r_n\rangle$ and, from (2.2b), (3.3a) and (S1), $\langle(\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}, r_n\rangle = o_p(\sqrt{\lambda_{0n}})$. We can then write $\langle(\mathbf{I} - \hat{\mathbf{S}})\mu_{0n}, \mathbf{P}\varepsilon_n\rangle = \mu_{0n}^T(\mathbf{I} - \hat{\mathbf{S}})\mathbf{X}[n^{-1}\mathbf{X}^T\varepsilon_n]$. This quantity is $o_p(\sqrt{\lambda_{0n}})$ because

$n^{-1}\mathbf{X}^T\varepsilon_n$ is uniformly componentwise $O_p(n^{-1/2})$ and, by (3.3), (S1) and (S3), $\mathbf{X}^T(\mathbf{I}-\hat{\mathbf{S}})\mu_{0n}$ is uniformly componentwise $o_p(\sqrt{n\lambda_{0n}})$.

Now write $\|\mathbf{S}_{\lambda_{0n}}\varepsilon_n\|^2 = Q_n + \sum_{i=1}^n s_{ii}e_i^2$ and observe from (S4a) that $\sum_{i=1}^n s_{ii}e_i^2 - \sigma^2\lambda_{0n} = o_p(\sqrt{\lambda_{0n}})$ and $\mathrm{var}Q_n = 2\sigma^4(\lambda_{0n} - \sum_{i=1}^n s_{ii}^2) = 2\sigma^4\lambda_{0n}\{1+o(1)\}$. We may now apply Theorem 2.1 of de Jong (1987) with conditions (a) and (b) of that theorem following from the inequality $0 \le \max_{1\le i\le n}[s_{ii}-s_{ii}^2] \le 1$ and (S4b), respectively. Consequently,

$$(\|\mathbf{S}_{\lambda_{0n}}\varepsilon_n\|^2 - \sigma^2\lambda_{0n})/\sigma^2\sqrt{2\lambda_{0n}} \xrightarrow{d} N(0,1). \tag{4.2}$$

Using (2.2b), (S2) and (4.2), we see that $\langle\hat{\mathbf{S}}\varepsilon_n, r_n\rangle = o_p(\sqrt{\lambda_{0n}})$. Then, to show that $\langle\hat{\mathbf{S}}\varepsilon_n,\, \mathbf{P}\varepsilon_n\rangle = o_p(\sqrt{\lambda_{0n}})$ it suffices to show that the vector $\mathbf{X}^T\hat{\mathbf{S}}\varepsilon_n$ is componentwise $o_p(\sqrt{n\lambda_{0n}})$. For this purpose we need work only with orders in

$$\Lambda_n = \{\lambda : |(\lambda-\lambda_{0n})/\lambda_{0n}| \le n^{-\gamma}\} \tag{4.3}$$

because $P(\hat{\lambda}_{0n} \in \Lambda_n) \to 1$ for any $0<\gamma<1/2$ due to (2.4). However, for any $\lambda \in \Lambda_n$, we have $\mathrm{var}(x_{in}^T S_\lambda \varepsilon_n) \le \sigma^2 n$, $i = 1,\ldots,p$, and the desired result then follows from Bonferroni's inequality and (3.3b), since the cardinality of Λ_n is $O(n^{-\gamma}\lambda_{0n})$. The remainder of the proof is immediate from (3.3), (S2) and (4.2).

4.4. Proof of Corollary 2

Here we demonstrate that the conditions for Theorem 2 hold under the assumptions made in Corollary 2. We begin by observing that the differentiability assumption on $C(\cdot)$, along with the $\sqrt{n}$-consistency of $\hat{\theta}$, ensures that $\hat{\lambda}_{0n}$ satisfies (2.4). Using this we can see, for example, that if $\hat{\lambda}_{0n} > \lambda_{0n}$ then $\|(\mathbf{I}-\hat{\mathbf{S}})\mu_{0n}\|^2 - \|(\mathbf{I}-\mathbf{S}_{\lambda_{0n}})\mu_{0n}\|^2 = n^{-1}\sum_{j=\lambda_{0n}}^{\hat{\lambda}_{0n}}\langle\Phi_{jn},\, \mu_{0n}\rangle^2 = O_p(n^{-(1-\tau^{-1})/2})$ as a result of (3.4). A similar argument using the condition $(\eta-1/4)/\tau > 1/2$ shows that (S3) holds.

To establish (S2) it again suffices to work on the set Λ_n in (4.3). For $\lambda \in \Lambda_n$ define $D_\lambda = \|\mathbf{S}_\lambda\varepsilon_n\|^2 - \|\mathbf{S}_{\lambda_{0n}}\varepsilon_n\|^2 - \sigma^2(\lambda-\lambda_{0n})$ and note that $ED_\lambda = 0$ and $\mathrm{var}D_\lambda = 2|\lambda-\lambda_{0n}| + o(|\lambda-\lambda_{0n}|^2/n)$ since the elements of $\mathbf{S}_\lambda - \mathbf{S}_{\lambda_{0n}}$ are uniformly bounded by a constant multiple of $|\lambda-\lambda_{0n}|/n$. The result then follows upon observing that $\max_{\lambda\in\Lambda_n}|D_\lambda/\sqrt{\lambda_{0n}}|$ is $o_p(1)$ due to the Bonferroni and Chebychev inequalities and the fact that Λ_n has cardinality $n^{-\gamma}\lambda_{0n}$, for γ arbitrarily close to 1/2.

Condition (S4b) can be verified as in Chen (1994). For (S4a), write $s_{ii} = n^{-1}\{1+2\sum_{j=1}^{\lambda_{0n}}\cos(j\pi t_{in})^2\}$ to see that $\sum_{i=1}^n s_{ii}^2 \le (2\lambda_{0n}+1)^2/n$. Finally, to show (3.6), write $\langle\mu_{0n},\, (\mathbf{I}-\hat{\mathbf{S}})\mu_{0n}\rangle = n\sum_{j=\hat{\lambda}_{0n}+1}^n (n^{-1}\langle\Phi_{jn}, \mu_{0n}\rangle)^2$ and use (3.4), the $\sqrt{n}$-consistency of $\hat{\theta}$, the differentiability of $C^2(\cdot)$ and the bound $\sum_{j=\hat{\lambda}_{0n}+1}^n j^{-2\tau} \le 1/\{(2\tau-1)\hat{\lambda}_{0n}^{(2\tau-1)}\}$.

Acknowledgement

We thank an associate editor and the referees for their helpful comments. The research of R. Eubank was supported by the National Science Foundation. The research of C. S. Li was partially supported by Cancer Center Support Grant CA-21765 from the National Institutes of Health and by the American Lebanese Syrian Associated Charities (ALSAC). The research of S. Wang was partially supported by the National Cancer Institute (CA-57030), the TAMU IRI program, and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30ES09106).

References

Aerts, M., Claeskens, G. and Hart, J. (2000). Testing lack of fit in multiple regression. *Biometrika* **87**, 403-424.

Baraud, Y., Huet, S. and Laurent, B. (2003). Adaptive tests of linear model hypotheses by model selection. *Ann. Statist.* **31**, 225-251.

Barry, D. and Hartigan, J. A. (1990). An omnibus test for departures from constant mean. *Ann. Statist.* **18**, 1340-1357.

Chen, J. C. (1994). Testing for no effect in nonparametric regression via spline smoothing techniques. *Ann. Inst. Statist. Math.* **46**, 251-265.

Cox, D. D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16**, 713-732.

Cox, D., Koh, E., Wahba, G. and Yandell, B. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.* **16**, 113-119.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Statist.* **27**, 1012-1040,

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing.* Second edition, Marcel-Dekker, New York.

Eubank, R. L. (2000). Testing for no effect by cosine series methods. *Scand. J. Statist.* **27**, 747-763.

Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.* **20**, 1412-1425.

Eubank, R. L. and Jayasuriya, B. (1993). The asymptotic average squared error for polynomial regression. *Statistics* **24**, 311-319.

Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Amer. Statist. Assoc.* **85**, 387-392.

Eubank, R. L. and Wang, S. (1994). Confidence regions in non-parametric regression. *Scand. J. Statist.* **21**, 147-158.

Fan, J. (1996). Tests of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.* **91**, 684-688.

Fan, Y. and Linton, O. (2003). Some high-order theory for a consistent non-parametric model specification test. *J. Statist. Plann. Inference* **109**, 125-154.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153-193.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625-633.

Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference based estimation of variance in nonparametric regression. *Biometrika* **77**, 521-528

Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77**, 415-419.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *J. Amer. Statist. Assoc.* **83**, 86-99.

Härdle, W. and Kneip, A. (1999). Testing a regression model when we have a smooth alternative in mind. *Scand. J. Statist.* **26**, 221-238.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926-1947.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York.

Heyde, C. C. and Brown, B. M. (1970). On the departure from normality of a certain class of martingales. *Ann. Math. Statist.* **41**, 2161-2165.

Inglot, T. and Ledwina, T. (1996). Asymptotic optimality of data driven Neyman's test. *Ann. Statist.* **24**, 1982-2019.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probab. Theory Related Fields* **75**, 261-277.

Kuchibhatla, M. and Hart, J. D. (1996). Smoothing-based lack-of-fit tests: variations on a theme. *J. Nonparametr. Statist.* **7**, 1-22.

Lee, G. and Hart, J. D. (2000). Model selection criteria with data dependent penalty, with applications to data-driven Neyman smooth tests. *J. Nonparametr. Statist.* **12**, 683-707.

Li, C. S. (1999). Testing lack of fit of regression models under heteroscedasticity. *Canad. J. Statist.* **27**, 485-496.

Li, C. S. (2001). Testing for linearity in generalized linear models using kernel smoothing. *J. Statist. Comput. Simulation* **69**, 339-351.

Müller, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer, New York.

Müller, H. G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78**, 521-530.

Müller, H. G. (1992). Goodness-of-fit diagnostics for regression models. *Scand. J. Statist.* **19**, 157-172.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

Rice, J. and Rosenblatt, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Ann. Statist.* **11**, 141-156.

Seber, G. A. F. and Wild, C. J. (1988). *Nonlinear Regression*. Wiley, New York.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5**, 302-305.

Zheng, J. X. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econom. Theory* **14**, 123-138

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

E-mail: eubank@stat.tamu.edu

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

E-mail: sjwang@stat.tamu.edu

Department of Biostatistics St. Jude Children's Research Hospital, 332 N. Lauderdale, Memphis, TN 38105-2794, U.S.A.

E-mail: chinshang.li@stjude.org

(Received June 2003; accepted February 2004)