7-1-1995

Testing Additivity in Generalized Nonparametric Regression Models

Pedro Gozalo

Oliver B. Linton

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1106

TESTING ADDITIVITY IN GENERALIZED NONPARAMETRIC REGRESSION MODELS

Pedro Gozalo and Oliver Linton

July 1995
Revised February 1997

TESTING ADDITIVITY IN GENERALIZED NONPARAMETRIC REGRESSION MODELS

P.L. Gozalo

Department of Economics, Brown University,

Providence, RI 02912, USA

O.B. Linton[1]

Cowles Foundation for Research in Economics, Yale University

New Haven, CT 06520, USA

Summary

We develop kernel-based consistent tests of an hypothesis of additivity in nonparametric regression extending recent work on testing parametric null hypotheses against nonparametric alternatives. The additivity hypothesis is of interest because it delivers interpretability and reasonably fast convergence rates for standard estimators. The asymptotic distributions of the tests under a sequence of local alternatives are found and compared: in fact, we give a ranking of the different tests based on local asymptotic power. The practical performance is investigated via simulations and an application to the German migration data of Linton and Härdle (1996).

Some key words: Additive regression models; Dimensionality reduction; Kernel estimation; Nonparametric regression; Testing.

JEL Classification Number: C12, C13.

February 24, 1997

[1]Address correspondence to: Oliver Linton, Cowles Foundation for Research in Economics, Yale University, 30 Hillhouse Avenue, New Haven, CT 06520-8281, USA. Phone: (203) 432-3699. Fax: (203) 432-6167. http://www.econ.yale.edu/~linton

1 Introduction

According to Luce and Tukey (1964), additivity is basic to science. This simplifying structure is present in many models of economic behavior starting with Leontieff (1947); see Deaton and Muellbauer (1980) for examples. It is certainly hard to think of models that are not additive in some sense. Additivity is also widely used in parametric and semiparametric models of economic data. Our purpose here is to investigate a very general class of statistical models that combine additive separability with an unrestricted functional form for the covariate effects; this general class of structures are generically called additive nonparametric regression models. Specifically, this paper introduces a new test for a specified form of additively separability in a class of nonparametric regression models.

Let (Y, X) be a random variable with X of dimension d and Y a scalar and let the regression function be $m(x) = E(Y \mid X = x)$. We say that $m(x)$ has a generalized additive structure if

$$G\{m(x)\} = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \tag{1}$$

for some known "link function" G, where $x = (x_1, \ldots, x_d)'$ are the d-dimensional predictor variables and m_α are one-dimensional nonparametric functions operating on each element of the vector or predictor variables. It is convenient here to assume that $E\{m_\alpha(X_\alpha)\} = 0$ for identification. This class of models includes additive regression when G is the identity and multiplicative regression when G is the logarithm. For binary data it is appropriate to take G to be the inverse of a cumulative distribution function like the normal or logit [this ensures that the regression function lies between 0 and 1 no matter what values $c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)$ takes]. Compare this specification with the semiparametric single index model considered in Ichimura (1993) in which the index on the right hand side of (1) is linear, but the link function $G(\cdot)$ is unrestricted [apart from the fact that it is the inverse of a c.d.f]. Both models considerably weaken the restrictions imposed by parametric binary choice models, but are non-nested. One advantage of the additive model is that it allows for more general elasticity patterns: specifically,

while in the single index model $\eta_{j:k} = (\partial \ln m / \partial x_j) / (\partial \ln m / \partial x_k)$ is restricted to be constant with respect to x, for the additive model $\eta_{j:k}$ can vary with x_j and x_k [although not with other $x's$]. Note that (1) is a partial model specification and we have not restricted in any way the variance or other aspects of the conditional distribution $\mathcal{L}(Y|X)$ of Y given X. A full model specification, widely used in this context, is to assume that $\mathcal{L}(Y|X)$ belongs to an exponential family with known link function G and mean m. This class of models was called *Generalized Additive* by Hastie and Tibshirani (1991). In some respects, econometricians would prefer the partial model specification in which we keep (1), but do not restrict ourselves to the exponential family. This flexibility is a relevant consideration for many datasets where there is overdispersion or heterogeneity.

Estimation in these models was first discussed by Stone (1985,1986) who showed that the optimal rate for estimating $m(\cdot)$ is the one-dimensional rate of convergence e.g. $n^{2/5}$ for twice continuously differentiable functions. In the statistical literature the additive regression model has been advanced in the eighties largely by the work of Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1991). Their estimation methods, called generically backfitting, rely on iteratively computing one-dimensional smooths. Unfortunately, there does not exist a central limit theorem for these procedures as yet. Recently, Linton and Nielsen (1995), Tjøstheim and Auestadt (1994), and Newey (1994) have independently proposed an alternative procedure for estimating m_α, when G is the identity, based on integration of a standard kernel estimator. The procedure is explicitly defined and its asymptotic distribution is easily derived: it converges at the one-dimensional rate and satisfies a central limit theorem. This estimation procedure has been extended to a number of other contexts like the generalized additive model [Linton and Härdle (1996)], to dependent variable transformation models [Linton, Chen, Wang, and Härdle (1996)], to econometric time series models [Härdle and Yang (1996)], and to hazard models with time varying covariates and right censoring [Nielsen (1996)].

The additive structure is important in terms of its interpretability and its ability to deliver fast rates of convergence. Unfortunately, especially when G is not the identity, (1) is not terribly

robust. Specifically, if some relevant variables were omitted, the regression function on this new set of variables does not satisfy (1). Similarly, if the wrong link function were used, additivity would not be maintained. More generally, how do we know that the additive structure provides a good approximation? In light of this we think it important to test (1) statistically. In this paper we propose tests of additivity based on the integration method of estimation and derive their asymptotic properties under sequences of local alternatives. Our tests are based on similar principles to other smoothing-based tests for parametric null hypotheses against general nonparametric alternatives, as for example in Eubank and Spiegelman (1990), Gozalo (1993, 1995), Härdle and Mammen (1993), Hjellvik and Tjøstheim (1995), Hong and White (1995), Staniswalis and Severini (1991) and Zheng (1996), and to tests for omitted variables in nonparametric regression such as Hidalgo (1994), Fan and Li (1996), Gozalo (1993, 1995), Aït-Sahalia, Bickel, and Stoker (1994), and Lavergne and Vuong (1996). The contribution we make is to apply these principles to testing the additivity hypothesis which may be important for future applications. We examine four different tests and derive their asymptotic distributions under sequences of local alternatives; this enables us to make a ranking of the tests according to their power. We also provide some simulation evidence on the small sample properties of all tests.

Notation. We use $\|\cdot\|_p$ to denote the L_p norm of a function, e.g. $\|g\|_2 = (\int g^2(x)dx)^{1/2}$ and $\|g\|_\infty = \sup_x |g(x)|$. The convolution between two densities K and L is denoted $K * L(t) = \int K(s)L(t-s)ds$. We use $\overset{AD}{=}$ to denote asymptotic equivalence in distribution.

2 Hypotheses of Interest

Throughout we work with independently sampled data $\{(Y_i, X_i)\}_{i=1}^n$ from a common population, although we allow for considerable heterogeneity in the conditional distribution $\mathcal{L}(Y|X)$. We first provide the machinery to reformulate (1) in a convenient way for application of the

integration methodology. For any direction α, partition $x = (x_\alpha, x_{\underline{\alpha}})$ and $X_i = (X_{\alpha i}, X_{\underline{\alpha} i})$, and let $p(\cdot)$, $p_\alpha(\cdot)$ and $p_{\underline{\alpha}}(\cdot)$ denote the marginal density of X, X_α and $X_{\underline{\alpha}}$, respectively. Let

$$\varphi_\alpha(x_\alpha) = \int G\left\{m(x_\alpha, x_{\underline{\alpha}})\right\} p_{\underline{\alpha}}(x_{\underline{\alpha}}) dx_{\underline{\alpha}},$$

$\overline{c}_\alpha = \int \varphi_\alpha(x_\alpha) p_\alpha(x_\alpha) dx_\alpha$, and let $\overline{c} = d^{-1} \sum_\alpha \overline{c}_\alpha$. Finally, let

$$m^0(x) = F\left\{\sum_{\alpha=1}^{d} \varphi_\alpha(x_\alpha) - (d-1)\overline{c}\right\},$$

where $F = G^{-1}$. When the additive restriction (1) is true,

$$\varphi_\alpha(x_\alpha) = m_\alpha(x_\alpha) + c \quad ; \quad \overline{c}_\alpha = c \quad ; \quad m^0(x) = m(x). \tag{2}$$

Relation (2) is the basis of the so-called integration method of estimating additive nonparametric models as exploited in Linton and Härdle (1996).

We are concerned with testing the validity of the additive specification (1) of the regression function $m(x)$ over a subset of interest $\mathcal{X} \subseteq \mathbb{R}^d$ of the support of X. Thus the null hypothesis to be tested can be reformulated as

$$\mathbf{H}_0 : \; m(x) = m^0(x), \qquad \text{all } x \in \mathcal{X}, \tag{3}$$

against the general alternative that $\mathbf{H}_0$ is false, which we denote by $\mathbf{H}_A$. Both $\mathbf{H}_0$ and $\mathbf{H}_A$ are nested within the following general class of local alternatives:

$$\mathbf{H}_n : \; m(x) = m^0(x) + \delta_n \lambda(x), \quad \text{all } x \in \mathcal{X}, \tag{4}$$

for certain δ_n. The null hypothesis (3) is given by $\delta_n = 0$, while $\delta_n = 1$ yields the global alternative. We shall choose the rate at which δ_n tends to zero to obtain a limiting distribution with noncentrality parameter bounded away from 0 and ∞. Note that the null hypothesis can be equivalently restated as $G\{m^0(x)\} = \sum_{\alpha=1}^{d} \varphi_\alpha(x_\alpha) - (d-1)\overline{c}$, for all $x \in \mathcal{X}$, which suggests specifying the local alternatives as

$$\mathbf{H}_n^*: \; G\{m(x)\} = \sum_{\alpha=1}^{d} \varphi_\alpha(x_\alpha) - (d-1)\overline{c} + \delta_n \lambda^*(x), \quad \text{all } x \in \mathcal{X}. \tag{5}$$

However, for δ_n small: a given $\lambda(x)$ in (4) is equivalent to having $\lambda^*(x) = \lambda(x)G'\{m^0(x)\}$ in (5), and a given $\lambda^*(x)$ in (5) is equivalent to having $\lambda(x) = \lambda^*(x)F'[G\{m^0(x)\}]$ in (4). Therefore, without loss of generality we can restrict our attention to (4).

3 Estimation and Test Statistics

3.1 Estimation

We provide two estimates of $m(x)$: one that is consistent when (1) is assumed and one that is consistent more generally. To estimate $m(x)$ in the general case we will use the multidimensional Nadaraya-Watson product kernel estimator

$$\widehat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^n K_h(x - X_i)Y_i}{n^{-1}\sum_{i=1}^n K_h(x - X_i)} \equiv \frac{\widehat{r}_h(x)}{\widehat{p}_h(x)}, \tag{6}$$

where $K_h(x - X_i) = \prod_{\alpha=1}^d k_h(x_\alpha - X_{\alpha i})$ in which $k_h(\cdot) = h^{-1}k(\cdot/h)$ with $k(\cdot)$ a one-dimensional kernel and $h = h(n)$ a bandwidth sequence. Under our regularity conditions given below, the Nadaraya-Watson estimator satisfies

$$\widehat{m}_h(x) - m(x) \overset{AD}{=} N\left[\frac{h^q}{q!}\mu_q(k)b(x), \frac{1}{nh^d}\nu_0(K)v(x)\right], \tag{7}$$

where $\mu_q(k) = \int u^q k(u)du$ and $\nu_0(K) = \int K^2(u)du$, while $v(x) = \sigma^2(x)/p(x)$, where $\sigma^2(x) = \mathrm{var}(Y|X = x)$ is the conditional variance function, and $b(x)$ is the bias function (when $q = 2$, this is $\mathrm{tr}\left[\partial^2 m(x)/\partial x \partial x' + 2\{\partial \ln p(x)/\partial x\}\partial m(x)/\partial x'\right]$). Here, q is the order of the kernel. Note that the (mean squared error) optimal bandwidth is of order $n^{-1/(2q+d)}$ for which the asymptotic mean squared error is of order $n^{-2q/(2q+d)}$, see Härdle and Linton (1994), which reflects the curse of dimensionality − as d increases, the rate of convergence decreases.

When $m(\cdot)$ satisfies the generalized additive model structure (1), we can estimate $m(x)$ with a better rate of convergence by imposing these restrictions. Following Linton and Härdle (1996) we define empirical versions of $\varphi_\alpha(\cdot)$, $\overline{c}_\alpha$, and c;

$$\widetilde{\varphi}_\alpha(x_\alpha) = n^{-1} \sum_{i=1}^{n} G\left\{\widehat{m}_{h_0}(x_\alpha, X_{\underline{\alpha} i})\right\}, \tag{8}$$

and $\widetilde{c}_\alpha = n^{-1} \sum_{i=1}^{n} \widetilde{\varphi}_\alpha(X_{\alpha i})$ and $\widetilde{c} = d^{-1} \sum_{\alpha=1}^{d} \widetilde{c}_\alpha$. We then reestimate $m(x)$ by

$$\widetilde{m}_{h_0}(x) = F\left\{\sum_{\alpha=1}^{d} \widetilde{\varphi}_\alpha(x_\alpha) - (d-1)\widetilde{c}\right\}. \tag{9}$$

Linton and Härdle (1996) derived the pointwise asymptotic properties of $\widetilde{\varphi}_\alpha(x_\alpha)$ and $\widetilde{m}(x)$: under their regularity conditions,

$$\widetilde{m}_{h_0}(x) - m^0(x) \stackrel{AD}{=} N\left[\frac{h_0^q}{q!}\mu_q(k) b_0(x), \frac{1}{nh_0}\nu_0(k) v_0(x)\right], \tag{10}$$

where $b_0(x) = F'\left[G\left\{m(x)\right\}\right] \sum_\alpha b_{\alpha 0}(x_\alpha)$ and $v_0(x) = F'\left[G\left\{m(x)\right\}\right]^2 \sum_\alpha v_{\alpha 0}(x_\alpha)$ with $b_{\alpha 0}(x_\alpha) = \int G'\left\{m(x)\right\} b(x) p_{\underline{\alpha}}(x_{\underline{\alpha}}) dx_{\underline{\alpha}}$ and $v_{\alpha 0}(x_\alpha) = \int G'\left\{m(x)\right\}^2 v(x) p^2_{\underline{\alpha}}(x_{\underline{\alpha}}) dx_{\underline{\alpha}}$.[2] By choosing $h_0 \propto n^{-1/(2q+1)}$ one can achieve the optimal rate of convergence i.e. mean squared error of order $n^{-2q/2q+1}$, which is independent of the dimensions d. One feature of the Linton and Härdle (1996) analysis was that for large dimensions they used bias reduction in the "directions not of interest". This added flexibility is important (at least in the technical analysis) for large dimensions when the objective is to obtain the optimal one-dimensional rate of convergence. In this paper we are concerned with the properties of the test and so will take a slightly different approach: indeed we shall chose the same kernel for both procedures, but allow k to be higher order throughout.

[2]This result does not assume (1), but when (1) is true $m^0(x) = m(x)$ and the target is actually $m(x)$.

3.2 Test Statistics

The first proposed test statistic is, rather like a Hausman test, based on the mean squared difference between the unrestricted nonparametric estimate and the restricted estimate over the sample points

$$\widehat{\tau} = n^{-1} \sum_{j=1}^{n} \left\{ \widehat{m}_h(X_j) - \widetilde{m}_{h_0}(X_j) \right\}^2 \pi(X_j), \tag{11}$$

where $\pi(\cdot)$ is a prespecified nonnegative weighting function used, for example, to eliminate or ameliorate boundary problems. This test relies on the fact that $\widehat{\tau}$ consistently estimates $\tau_0 = \int \left\{ m(x) - m^0(x) \right\}^2 \pi(x) p(x) dx$, which is zero if and only if the null hypothesis is satisfied. We also consider testing in the G-scale using the statistic

$$\widehat{\tau}' = n^{-1} \sum_{j=1}^{n} \left[G\left\{ \widehat{m}_h(X_j) \right\} - G\left\{ \widetilde{m}_{h_0}(X_j) \right\} \right]^2 \pi(X_j) \tag{12}$$

and to accommodate both scales at once we shall write

$$\widehat{\omega}_{0Q} = n^{-1} \sum_{j=1}^{n} \left[Q\left\{ \widehat{m}_h(X_j) \right\} - Q\left\{ \widetilde{m}_{h_0}(X_j) \right\} \right]^2 \pi(X_j) \tag{13}$$

for some known function Q [we shall on occasion drop the additional Q subscript in $\widehat{\omega}_{0Q}$].

There are a number of alternative paradigms for testing additivity that have been used in other nonparametric contexts. Two tests that have some analogy with a Lagrange Multiplier test are

$$\widehat{\omega}_1 = \frac{1}{n} \sum_{i=1}^{n} \widetilde{u}_i \left\{ \widehat{m}_h(X_i) - \widetilde{m}_{h_0}(X_i) \right\} \pi(X_i) \tag{14}$$

considered, among others, in Hong (1993) (with $\pi(X_i) \propto \widehat{p}_h(X_i)$), and the quadratic form

$$\widehat{\omega}_2 = \frac{1}{n^2 h^d} \sum\sum_{i \neq j} K\left(\frac{X_i - X_j}{h} \right) \widetilde{u}_i \widetilde{u}_j \pi(X_i) \pi(X_j), \tag{15}$$

where $\widetilde{u}_i = Y_i - \widetilde{m}_{h_0}(X_i)$ are the additive (restricted) residuals. Both these tests look for correlations between the restricted residuals and suitable functions of X. The latter type of test has been proposed in Zheng (1996) in the context of testing a parametric null against a nonparametric alternative, and in Fan and Li (1996) and Lavergne and Vuong (1996) to test for nonparametric exclusion of variables. Another paradigm is motivated by the likelihood ratio method of parametric statistics. If the errors were homoskedastic Gaussian, we might consider the difference in the sum of squared residuals

$$\widehat{\omega}_3 = n^{-1} \sum_{i=1}^{n} \left(\widetilde{u}_i^2 - \widehat{u}_i^2 \right) \pi(X_i), \tag{16}$$

where $\widehat{u}_i = Y_i - \widehat{m}_h(X_i)$. In situations where Y is subject to support restrictions it may be appropriate to use other criteria. Our application below is to binary data for which the following criterion suggests itself

$$\widehat{\omega}_{3b} = n^{-1} \sum_{i=1}^{n} \left[Y_i \ln \left\{ \frac{\widehat{m}_h(X_i)}{\widetilde{m}_{h_0}(X_i)} \right\} + (1 - Y_i) \ln \left\{ \frac{1 - \widehat{m}_h(X_i)}{1 - \widetilde{m}_{h_0}(X_i)} \right\} \right] \pi(X_i). \tag{17}$$

A version of this statistic has been proposed in Azzalini, Bowman, and Härdle (1989) as a device for checking parametric models. See also Staniswalis and Severini (1991) for some theoretical analysis.

Finally, Härdle and Mammen (1993) consider a modification to (11) that replaces the null model estimate with a kernel smoothed version of it. The purpose of this is to eliminate smoothing bias terms associated with the unrestricted estimation (which are present even under the null hypothesis). This is really a version of the trend removal procedure whose study was initiated in Stone (1977). It can be readily adapted to the current context. Specifically, consider in place of $\widehat{m}_h(X_j) \equiv \sum_\ell w_{j\ell} Y_\ell$ (with $w_{j\ell} = n^{-1} K_h(X_j - X_\ell)/\widehat{p}_h(X_j)$, see (6)) the estimator

$$\widehat{m}_h^{tra}(X_j) = \widetilde{m}_{h_0}(X_j) + \sum_\ell w_{j\ell} \widetilde{u}_\ell.$$

This has bias essentially like $\widetilde{m}_{h_0}(X_j)$ under the null hypothesis [so you get a cancellation of bias terms]. One can also make the adjustment in a multiplicative fashion, i.e.

$$\widehat{m}_h^{trb}(X_j) = \widetilde{m}_{h_0}(X_j) \times \sum_\ell w_{j\ell} \frac{Y_\ell}{\widetilde{m}_{h_0}(X_\ell)},$$

see Jones, Linton, and Nielsen (1995) for discussion. Rather than analyze versions of $\widehat{\omega}_0$ and $\widehat{\omega}_1$ with $\widehat{m}$ replaced by $\widehat{m}_h^{tra}$ or $\widehat{m}_h^{trb}$ we point out that they behave statistically rather like $\widehat{\omega}_2$ (after interchanging summations), which also has no bias from the unrestricted estimation.

4 Asymptotic Properties

The test statistics $\widehat{\omega}_j$, $j = 0, \ldots, 3$ have very similar statistical properties; namely, after location and scale adjustment they are asymptotically standard normal. Let

$$T_{nj} = \frac{nh^{d/2}\widehat{\omega}_j - \mu_{nj}}{V_{nj}^{1/2}}, \quad j = 0, \ldots, 3, \tag{18}$$

where $\mu_{nj} = E\left\{nh^{d/2}\widehat{\omega}_j \,|\mathbf{H}_0\right\}$ and $V_{nj} = \text{var}\left\{nh^{d/2}\widehat{\omega}_j \,|\mathbf{H}_0\right\}$, $j = 0, \ldots, 3$, or asymptotic approximations thereof. We will show that

$$T_{nj} \Rightarrow N(0,1), \quad j = 0, \ldots, 3, \tag{19}$$

under the null hypothesis $\mathbf{H}_0$, while under fixed alternatives $\mathbf{H}_A$, $T_{nj} \to_p \infty$. Therefore, a consistent test is provided by T_{nj} (or suitable approximations thereof). A natural rejection rule here is then

$$\text{reject at level } \alpha \text{ if } T_{nj} > z_\alpha, \tag{20}$$

where $\Phi(z_\alpha) = 1-\alpha$, that is, one-sided.[3] Given (20), $\Pr(\text{reject} \,|\mathbf{H}_0) \to \alpha$, while $\Pr(\text{reject} \,|\mathbf{H}_A) \to 1$.

4.1 A Central Limit Theorem

Our strategy is as follows. We first state and discuss our regularity conditions. We then give the relevant location and scale quantities for $\widehat{\omega}_j$, $j = 0, 1, 2, 3$ and provide a theorem that states the asymptotic properties of $\widehat{\omega}_j$ under the sequence of local alternatives. We provide two methods for obtaining critical values by either (a) explicitly estimating the unknown quantities or by (b) applying the bootstrap. Finally, we argue that the main result holds for the feasible tests.

We work throughout with a common kernel k but allow for two bandwidths h and h_0. We shall assume that:

ASSUMPTION A: (a) *The random sample* $\{Z_i \equiv (Y_i, X_i')', \ Y_i \in \mathbb{R}, \ X_i \in \mathbb{R}^d\}_{i=1}^n$, *is independent and identically distributed.* (b) *On the compact set* $\mathcal{X}$, *the variance function* $\sigma^2(X_i) = \text{var}(u_i|X_i)$ *is Lipschitz continuous and* $E(\exp(tu_i)|X_i) < \infty$ *almost surely for some* t *in a neighborhood of* 0, *where* $u_i = Y_i - m(X_i)$. (c) *The regression function* $m(\cdot)$ *and the marginal density* $p(\cdot)$ [*with respect to Lebesgue measure*] *of* X *are both* q *times continuously differentiable on* $\mathcal{X}$. (d) *The design density* $p(\cdot)$ *is bounded away from zero on* $\mathcal{X}$. (e) *The weighting function* $\pi : \mathcal{X} \to \mathbb{R}_+$ *is bounded continuous and positive. (f) The functions* G *and* F *have bounded continuous* q*'th derivatives over* $\mathcal{X}$. *(g) The kernel function* $k(\cdot)$ *with* $\int k(u)du = 1$ *is of bounded support, symmetric about zero, Lipschitz continuous, and of order* q *– that is* $\int u^i k(u)du = 0$, $i = 1, \ldots, q-1$. (h) *The alternative function* $\lambda : \ \mathcal{X} \to \mathbb{R}$ *is continuous.*

[3]We do not consider two-sided rejection. Although in the Hausman test, for example, one can achieve power against some root-n alternatives by this method, one loses power against other alternatives and in a minimax sense one is better off with the one-sided rejection. See Gourieroux and Tenreiro (1994) for a comparison of these two rules.

These assumptions are slightly stronger than those given in Linton and Härdle (1996). They provide sufficient conditions for the uniform convergence of $\widehat{p}_h(x)$, $\widehat{r}_h(x)$, and $\widetilde{m}_{h_0}$, so that

$$\left\|\widehat{p}_h - p\right\|_\infty, \left\|\widehat{r}_h - r\right\|_\infty = O_p\left(\frac{\log n}{\sqrt{nh^d}}\right) + O_p(h^q), \tag{21}$$

where $r(x) = m(x)p(x)$, see Härdle and Mammen (1993).[4] Let $\varrho_n = \max\left\{\left\|\widehat{p}_h - p\right\|_\infty, \left\|\widehat{r}_h - r\right\|_\infty\right\}$. In our proof given in the appendix we make a number of approximations: specifically, in the expansion of (13) we drop terms that are cubic in $\widehat{m} - m$; also, there is an error term in (10) of order ϱ_n^2 in probability. We end up with an approximation of the form

$$\widehat{\omega}_j = \widehat{\omega}_j^* + O_p(\varrho_n^3),$$

where the random variable $\widehat{\omega}_j^*$ is much more tractable: its moments can readily be found and a degenerate U-statistic central limit theorem can also be applied to it. A sufficient condition enabling us to drop the $O_p(\varrho_n^3)$ term from the analysis is that for some $\epsilon > 0$,

$$n^\epsilon nh^{d/2}\frac{1}{(nh^d)^{3/2}} \to 0 \quad \text{and} \quad nh^{d/2}h^{3q} \to 0.$$

The first condition requires that $n^{1-2\epsilon}h^{2d} \to \infty$ and the second condition requires that $3q + d/2 > 2d$, i.e. $q > d/2$. Under these restrictions it is possible to provide formulae for μ_{nj} and V_{nj} and to carry out the test. However, some of the terms [in μ_{n0} particularly] depend on the bias function of both the unrestricted and the restricted estimates, both of which are rather complicated quantities to estimate. We therefore impose a further restriction that $nh^{d/2}h^{2q} \to 0$ which makes these terms of smaller order. This requires that $q > 3d/4$.

Provided $h_0 \leq h$, the restricted estimator has bias magnitude less than or equal to that of the unrestricted estimator but variance considerably less [this is the result of the additional

[4]See also the recent paper by Masry (1996) which gives the sharp rate of uniform convergence for local polynomial estimation with dependent data under weaker conditions.

averaging]. In fact, the bias terms may tend to offset each other and in any case, by virtue of the higher order kernels, the bias terms do not appear in the limiting distribution of the test. So although the restricted and unrestricted estimators might nominally have the same rate of convergence [when $h = h_0$], the test statistic works because of the difference in the stochastic terms. Note that the larger the dimensionality, the greater the difference in the stochastic terms between the two estimators and perhaps the easier it is to discriminate between them statistically. This is apparently in conflict with the usual belief that high dimensions lead to poorer performance.

Before stating the main result we give the location and scale adjustments for the various test statistics. These are

$$\begin{aligned}
\mu_{n0} &= h^{d/2}\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ji}^2\sigma^2(X_i)\overline{\pi}(X_j) \quad ; \quad V_{n0} = 2h^d\sum_{i\neq j}\sum \rho_{ji}^2\sigma^2(X_i)\sigma^2(X_j)\overline{\pi}(X_i)\overline{\pi}(X_j) \quad ; \\
\mu_{n1} &= h^{d/2}\sum_{i=1}^{n} w_{ii}\sigma^2(X_i)\pi(X_i) \quad ; \quad V_{n1} = 2h^d\sum_{i\neq j}\sum w_{ji}^2\sigma^2(X_i)\sigma^2(X_j)\pi(X_i)\pi(X_j) \quad ; \\
V_{n2} &= \frac{2}{n^2h^d}\sum_{i\neq j}\sum K\left(\frac{X_j - X_i}{h}\right)^2\sigma^2(X_i)\sigma^2(X_j)\pi(X_i)^2\pi(X_j)^2 \quad ; \\
V_{n3} &= 2h^d\sum_{i\neq j}\sum \gamma_{ji}^2\sigma^2(X_i)\sigma^2(X_j)\pi(X_i)\pi(X_j),
\end{aligned}$$

where $\overline{\pi}(x) = \pi(x)Q'[m(x)]^2$, $w_{ji} = n^{-1}K_h(X_j - X_i)/\widehat{p}_h(X_j)$, $\rho_{ji} = \sum_\ell w_{j\ell}w_{i\ell}$, and $\gamma_{ji} = -(\rho_{ji} - 2w_{ji})$. Note also that $\mu_{n2} = 0$, while since $\widehat{\omega}_3 = 2\widehat{\omega}_1 - \widehat{\omega}_0$ (with $Q = 1$), we have $\mu_{n3} = 2\mu_{n1} - \mu_{n0}$.

THEOREM 1. *Suppose that Assumption A is satisfied and that the sequence of local alternatives (4) holds with $\delta_n = (nh^{d/2})^{-1/2}$. Suppose that for some $\epsilon > 0$, $n^{1-\epsilon}h^{2d} \to \infty$, where $d > 2$ and $q > 3d/4$. Finally, suppose that for some $\epsilon > 0$, $h^{(d/2)-\epsilon} \leq h_0 \leq h$. Then*

$$\sup_{-\infty<z<\infty}\left|\Pr(T_{nj} - \Delta_{nj}(\lambda) \leq z) - \Phi(z)\right| \to 0, \quad j = 0, 1, 2, 3, \tag{22}$$

where $\Phi(\cdot)$ is the standard normal distribution function and the noncentrality parameters are

$$\begin{aligned}\Delta_{n0}(\lambda) &= \int \lambda^2(x)\overline{\pi}(x)p(x)dx/V_{n0}^{1/2} \quad ; \quad \Delta_{n1}(\lambda) = \int \lambda^2(x)\pi(x)p(x)dx/V_{n1}^{1/2} \quad ; \\ \Delta_{n2}(\lambda) &= \int \lambda^2(x)\pi^2(x)p^2(x)dx/V_{n2}^{1/2} \quad ; \quad \Delta_{n3}(\lambda) = \int \lambda^2(x)\pi(x)p(x)dx/V_{n3}^{1/2}.\end{aligned}$$

REMARK. When $d = 2$, there is insufficient difference between the magnitudes of the variances of the restricted and unrestricted estimators for the above theorem to be valid as stated. In this case, one must also include additional terms in μ_n coming from the stochastic part of the restricted estimator.

The tests have some power against all alternatives of magnitude $(nh^{d/2})^{-1/2}$. Furthermore, their local asymptotic powers are constant with respect to the data generation process given the value of the quantity $\int \lambda^2 d\mu$ [for a certain measure μ that varies somewhat from test to test]. This is in accord with the common sense requirement that power should increases with the magnitude [in this case a certain L_2-norm] of departure. We now compare the tests according to their power. By appropriate choice of π, the differences between $\Delta_{n0}(\lambda)$ [for $G = I$] and $\Delta_{nj}(\lambda)$, $j = 1, 2, 3$, only arise form the differences in the kernel constants in the variances.[5] We have essentially

$$V_{nj} \to_p 2 \left\| K_j \right\|_2^2 \times \int \sigma^4(x)\pi^2(x)dx, \quad j = 0, 1, 2, 3,$$

where $K_0 = K * K$, $K_1, K_2 = K$, and $K_3 = K_0 - 2K_1$. In fact,

[5]The local power of the test based on the transformed differences $\widehat{\omega}_{0Q}$ varies with Q. We are ignoring the difference between π and $\overline{\pi}$ and the fact that V_{n2} should have an additional $\pi^2(x)p^2(x)$ inside the integral, which counters the additional $\pi(x)p(x)$ in the numerator of the noncentrality $\Delta_{n2}(\lambda)$.

$$\|K_0\|_2^2 \leq \|K\|_2^2 \leq \|K_3\|_2^2$$

which suggests that $\widehat{\omega}_0$ is the most powerful test, at least in large samples.

4.2 Implementation of the Tests

There are two alternative ways of carrying out the test: one based on the asymptotic critical values and the other on the bootstrap. In the first approach we must replace μ_n and V_n by consistent estimators $\widehat{\mu}_n$ and $\widehat{V}_n$. Define the relevant quantities

$$\begin{aligned}
\widehat{\mu}_{n0} &= h^{d/2}\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ji}^2 \widehat{u}_i^2 \widehat{\overline{\pi}}(X_j) \quad ; \quad \widehat{V}_{n0} = 2h^d \sum\sum_{i\neq j} \rho_{ji}^2 \widehat{u}_i^2 \widehat{u}_j^2 \widehat{\overline{\pi}}(X_i)\widehat{\overline{\pi}}(X_j) \quad ; \\
\widehat{\mu}_{n1} &= h^{d/2}\sum_{i=1}^{n} w_{ii} \widehat{u}_i^2 \pi(X_i) \quad ; \quad \widehat{V}_{n1} = 2h^d \sum\sum_{i\neq j} w_{ji}^2 \widehat{u}_i^2 \widehat{u}_j^2 \pi(X_i)\pi(X_j) \quad ; \\
\widehat{V}_{n2} &= \frac{2}{n^2 h^d}\sum\sum_{i\neq j} K\left(\frac{X_j - X_i}{h}\right)^2 \widetilde{u}_i^2 \widetilde{u}_j^2 \pi(X_i)^2 \pi(X_j)^2 \quad ; \\
\widehat{V}_{n3} &= 2h^d \sum\sum_{i\neq j} \gamma_{ji}^2 \widehat{u}_i^2 \widehat{u}_j^2 \pi(X_i)\pi(X_j),
\end{aligned}$$

where $\widehat{\overline{\pi}}(x) = \pi(x)Q'[\widehat{m}(x)]^2$; for the test $\widehat{\omega}_3$, take $\widehat{\mu}_{n3} = 2\widehat{\mu}_{n1} - \widehat{\mu}_{n0}$. Now let

$$\widehat{T}_{nj} = \frac{nh^{d/2}\widehat{\omega}_j - \widehat{\mu}_{nj}}{\widehat{V}_{nj}^{1/2}}, \quad j = 0, 1, 2, 3. \tag{23}$$

We then apply the rejection rule (20) with $\widehat{T}_{nj}$ replacing T_{nj}. The results of Theorem 1 apply to $\widehat{T}_{nj}$ provided $\widehat{V}_{nj}$ is consistent and $\widehat{\mu}_{nj}$ are consistent at a rate better than $h^{-d/2}$. In fact, since μ_{nj} and V_{nj} are averages and do not involve higher derivatives of m, one can estimate these quantities root-n consistently with bias reduction and under sufficient smoothness conditions.

The asymptotic approximations here can work poorly because there are many terms of slightly smaller magnitude that have been omitted, as pointed out in Hjellvik and Tjøstheim (1995). To mitigate this problem, our second method relies on the bootstrap to compute critical values. This can be done for the unadjusted statistics $\widehat{\omega}_j$, which is much simpler to implement, and for the asymptotically pivotal statistics $\widehat{T}_{nj}$, which ought to have better size according to standard bootstrap theory, see Horowitz (1995). There are two general approaches to the bootstrap here depending on whether the support of the dependent variable is restricted or not. When it is not subject to restrictions, we might proceed as in other nonparametric regression problems allowing for heteroskedasticity by using the wild bootstrap. This consists of the following steps:

1. Construct residuals $\widetilde{u}_i = Y_i - \widetilde{m}_{h_0}(X_i)$, $i = 1, \ldots, n$.

2. For each index i, randomly draw (with replacement) the bootstrap residual u_i^* from an arbitrary distribution $\widehat{F}_i^W$ such that for $Z \sim \widehat{F}_i^W$,

$$E_{\widehat{F}_i^W} Z = 0, \tag{24}$$

$$E_{\widehat{F}_i^W} Z^2 = (\widetilde{u}_i)^2, \tag{25}$$

$$E_{\widehat{F}_i^W} Z^3 = (\widetilde{u}_i)^3. \tag{26}$$

3. Generate the bootstrap sample $Y_i^* = \widetilde{m}_{h_0^*}(X_i) + u_i^*$, for $i = 1, \ldots, n$ for some bandwidth h_0^*.

4. Use the bootstrap sample $\{(Y_i^*, X_i)\}_{i=1}^n$ to calculate the quantities $\widehat{m}_h^*$ and $\widetilde{m}_{h_0}^*$ and thus the bootstrap test statistics, $\widehat{\omega}^*$ and $\widehat{T}_n^*$, in identical fashion to the way $\widehat{m}_h$, $\widetilde{m}_{h_0}$, $\widehat{\omega}$ and $\widehat{T}_n$ were computed from the original sample.

5. Repeat steps 2-4 B times and use the B values of $\widehat{\omega}^*$, and $\widehat{T}_n^*$, to construct the empirical bootstrap distribution functions, e.g. $F_n^*(\tau) = B^{-1}\sum_{b=1}^{B} 1\{\widehat{\omega}_b^* \leq \tau\}$. Use the empirical bootstrap distribution function to calculate empirical critical values or p-values.

This general proposal was first made in Wu (1986) in a parametric context who noticed how the naive bootstrap can fail to approximate the distribution of the linear model least squares estimators in the presence of heteroskedasticity. The wild bootstrap for nonparametric kernel regression was used by Härdle and Marron (1991) to construct simultaneous error bars. Note that it is necessary to take a larger bandwidth in the resample, as was suggested in Härdle and Marron (1991), even though we have no requirement to approximate the bias of $\widehat{m}$ or $\widetilde{m}$ well because of our bias reduction.[6] In the appendix we establish the following result:

THEOREM 2. *Let the same assumptions as in Theorem 1 hold with $\delta_n = 0$ and suppose that $h_0 \leq h_0^* \leq h$ and that for some $\epsilon > 0$, $h^{(2q+d/2)/(2q+1)-\epsilon} \leq h_0^*$. Then, we have along almost all sample sequences,*

$$\sup_{-\infty<z<\infty} |\Pr^*(\hat{T}_n^* < z) - \Phi(z)| \to 0,$$

where $\Pr^*$ *denotes the bootstrap distribution of* $Y_1^*, \ldots, Y_n^* | \{(X_i, Y_i)\}_{i=1}^n$.

When the dependent variable is subject to support constraints, e.g. it is binary, the above approach will not work well since one can obtain data inconsistent with the support constraints. Our second approach is appropriate when the dependent variable is binary, i.e. $\Pr(Y_i = 1|X_i = x) = m(x)$, although the main idea can be extended to other types of limited dependent variables. In this case, we recommend generating the bootstrap sample $\{Y_i^*\}_{i=1}^n$ by sampling

[6]The relevance of this issue in small samples is briefly addressed in our Monte Carlo simulations.

from a Bernoulli distribution with $\Pr(Y_i^* = 1|X_i = x_i) = \widetilde{m}_{h_0}(x_i)$. The bootstrap steps are therefore:

1. Calculate $\widetilde{m}_{h_0^*}(X_i)$, $i = 1, \ldots, n$, for some bandwidth h_0^*.
2. Construct Y_i^* as a random draw from a Bernoulli$\left\{\widetilde{m}_{h_0^*}(X_i)\right\}$, for $i = 1, \ldots, n$.
3. Use the bootstrap sample $\{(Y_i^*, X_i)\}_{i=1}^n$ to calculate the bootstrap test statistics, $\widehat{\omega}^*$ and $\widehat{T}_n^*$, in identical fashion to the way $\widehat{\omega}$ and $\widehat{T}_n$ were computed from the original sample.
4. Repeat steps 2 and 3 B times and use the B values of $\widehat{\omega}^*$, and $\widehat{T}_n^*$, to construct the empirical bootstrap distribution functions, e.g. $F_n^*(\tau) = B^{-1} \sum_{b=1}^B 1\{\widehat{\omega}_b^* \leq \tau\}$. Use the empirical bootstrap distribution function to calculate empirical critical values or p-values.

The same result can be obtained for this bootstrap procedure as was given in Theorem 2.

5 Numerical Results

5.1 Empirical Illustration

We first applied our procedures to the study of migration between East and West Germany using data from the 1991 Social and Economic Panel survey conducted by the Deutsche Institut Wirtschaftsforschung. This data were used in Linton and Härdle (1996). The binary dependent variable to be explained is whether the individual intended to migrate from East to West Germany at this time. The four continuous explanatory variables are: age, household income, rent, and a subjective measure of personal satisfaction on a scale of 1 to 10. Our sample consists

of 315 individuals who all had at least Abitur education[7] and had some friends in the West. Of these, 172 had the intention of leaving the East. Linton and Härdle (1996) presented estimates of the components $m_\alpha(\cdot)$; we now provide a test of whether their estimates can be interpreted as being additive in the logit-scale.

We used $B = 200$ iterations to compute the bootstrap critical values. The unrestricted bandwidth was chosen by cross validation to be $h = 0.8$, while in the implementation of the restricted estimate we adopted precisely the Linton and Härdle (1996) method.[8] Table 1 shows the empirical p-values based on both asymptotic and bootstrap approximations for a number of the tests. We fail to reject the null hypothesis of additivity at 10% significance levels using any of the tests.

*** Table 1 Here ***

5.2 Simulations

The simulation study compares the performance of the different test statistics in a setup determined by the Linton and Härdle (1996) dataset. The model used to generate the data under the null was

$$Y_i \sim \text{Bernoulli}\left\{\widetilde{m}_{h_0}\left(X_i\right)\right\}, \tag{27}$$

where $\widetilde{m}_{h_0}\left(X_i\right) = \Lambda\left\{\widetilde{c} + \sum_{\alpha=1}^{4} \widetilde{m}_\alpha\left(X_i\right)\right\}$ is the estimate obtained in the previous section imposing additivity with $\Lambda\left(\cdot\right)$ the logistic distribution function.

The model under the alternative was created by adding an interaction term to the additive null model. Linton and Härdle (1996) found the probability of migration to West Germany to

[7] i.e. they graduated from High School.

[8] In fact, they chose as pilot

$$\ddot{m}_{h_0}(x) = \frac{\sum_{j=1}^{n} \ell_{h_0}(x_\alpha - X_{\alpha j}) L_g(x_{\underline{\alpha}} - X_{\underline{\alpha} j}) Y_j}{\sum_{i=1}^{n} \ell_{h_0}(x_\alpha - X_{\alpha i}) L_g(x_{\underline{\alpha}} - X_{\underline{\alpha} i})}$$

where $\ell_{h_0}(\cdot) = h_0^{-1}\ell(h_0^{-1}\cdot)$ and $\ell(\cdot)$ is a one-dimensional Gaussian kernel, while $L_g(\cdot) = g^{-(d-1)}L(g^{-1}\cdot)$ and $L(\cdot)$ is a $d-1$-dimensional Gaussian kernel; $h_0 = 0.5$ and $g = 1.0$ are scalar bandwidths.

decrease with X_1 = age and X_4 = personal satisfaction, and increase with X_2 = income and X_3 = rent. The interaction term was constructed to enhance these individual variable effects. Specifically,

$$Y_i \sim \text{Bernoulli}\left[\Lambda\left\{\widetilde{c} + \sum_{\alpha=1}^{4} \widetilde{m}_\alpha\left(X_i\right) + 2\sin(2I_i)\right\}\right], \tag{28}$$

where I_i denotes the interaction $\log\{X_{2i}(X_{3i}+5)\}/\{X_{1i}(X_{4i}+0.5)\}$ (after location and scale normalization).

The Monte Carlo distributions of the tests under the null and alternative were obtained from 200 replications of (27) and (28), respectively. The computational requirements of the restricted estimator $\widetilde{m}_{h_0}(\cdot)$ prevented us from using in the simulations the same sample of $n = 315$ observations used in the empirical illustration. Instead, in each iteration we drew random samples of size $n = 50$ and 100 from the "population" of 315 values of X_i and its corresponding $\widetilde{m}_{h_0}(X_i)$, and then used these to generate the binary variable Y_i as in (27) and (28). To satisfy the assumptions of the test, we used a Gaussian (mixture) kernel of order $q = 4$, and bandwidths $h = an^{-(1-\varepsilon)/2d}$ ($\varepsilon = 0.01$) for $\widehat{m}_h(\cdot)$, and $h_0 = an^{-1/5}$ for $\widetilde{m}_{h_0}(\cdot)$, with a=1.6, 2.0, and 2.4. This range of the constant a includes the interval where the cross-validated estimates fell in a separate study. The weighting function $\pi(\cdot)$ was chosen to be zero for the 5% of the observations with the smallest density estimate $\widehat{p}_h(X_i)$ values, and one otherwise.

*** Additional Tables and Figures Here ***

Figures 1 and 2 show the boxplot for the empirical distribution of the statistics $\widehat{\omega}_j$, for $j = 0, 1, 2, 3$ ($\widehat{\omega}_0$ with $Q = 1$), and their normalized counterparts $\widehat{T}_{nj}$, $j = 0, 1, 2, 3$.[9] In each figure, the three boxplots on the left correspond to the distributions under the null for the three bandwidth constants $a = 1.6, 2.0$, and 2.4, respectively. The three boxplots on the right of each figure correspond to the distributions under the alternative for the same values of a.

[9]The tests $\widehat{\omega}_0$ and $\widehat{T}_{n0}$ with $Q = G$ are denoted $\widehat{\omega}_{0G}$ and $\widehat{T}_{n0G}$, respectively. The results for the Likelihood Ratio–type test $\widehat{\omega}_{3b}$ are basically the same as for $\widehat{\omega}_3$ and are omitted.

It is easy to see that the distribution of each test has a different sensitivity to the bandwidth choice. The figures also illustrate the location shift of the distributions under the alternative. Table 2 presents the percentage of rejections by $\widehat{\omega}_j$ and $\widehat{T}_{nj}$ of the alternative hypotheses based on the 5%, 10% and 20% empirical critical values (ECV). The Hausman-type tests, $\widehat{\omega}_0$ and $\widehat{\omega}_{0G}$, and the LM-type test $\widehat{\omega}_2$ (as well as their normalized versions) have power against the alternative even for $n = 50$. The power of these tests increases with the bandwidth size. The highest power among non-normalized tests is achieved by $\widehat{\omega}_2$, and by $\widehat{T}_{n0}$ among normalized tests. In contrast, the LM-type test $\widehat{\omega}_1$ and the error sum of squares type test $\widehat{\omega}_3$ (as well as their normalized versions) have very little power against this alternative at these sample sizes. Their power decreases with the bandwidth size.

Table 3 presents the percentage of rejections by the normalized tests $\widehat{T}_{nj}$ of the null and alternative hypotheses based on the 5%, 10% and 20% asymptotic critical values (ACV). The size of $\widehat{T}_{nj}, j = 0, 1, 2, 3$, based on the ACV is quite poor for all tests and bandwidths at this small sample sizes with the exception of $\widehat{T}_{n0}$ which are reasonable for $n = 100$. The best power corresponds to the two Hausman-type tests $\widehat{T}_{n0}$ and $\widehat{T}_{n0G}$ for $n = 100$.

Given the poor approximation of the asymptotic $N(0, 1)$ distribution to the distribution of most of the normalized tests, we investigated the bootstrap approximation for binary dependent variables described above. To this effect we calculated $\widetilde{m}_{h_0^*}(X_i)$ with $h_0^* = 1.1h_0$ and let Y_i^* be a random draw from a Bernoulli $\left\{\widetilde{m}_{h_0^*}(X_i)\right\}$, for $i = 1, \ldots, n$. The bootstrap sample $\{(Y_i^*, X_i)\}_{i=1}^n$ was then used to calculate the bootstrap statistics $\widehat{\omega}_j^*$ and $\widehat{T}_{nj}^*, j = 0, 1, 2, 3$. In each Monte Carlo iteration, $B = 100$ bootstrap samples were used to compute the bootstrap p–values. Table 3 presents the percentage of rejections of the null and alternative hypothesis for $n = 50$ at 5%, 10% and 20% significance levels. As before, $\widehat{\omega}_0$, $\widehat{\omega}_{0G}$, and $\widehat{\omega}_2$ (as well as their normalized versions) exhibit similar behaviour: the percentage of rejections appears to increase with the bandwidth size both under the null and the alternative. For $\widehat{\omega}_1$ and $\widehat{\omega}_3$ (as well as their normalized versions) the opposite is true. The best size is achieved at the bandwidth constant $a = 2.0$ for $\widehat{\omega}_j$, and $\widehat{T}_{nj}$, $j = 1, 2, 3$, and at $a = 2.4$ for the Hausman-type tests. For these "optimal" bandwidths,

the highest power based on the bootstrap critical values is attained by the Hausman-type tests. In general, though, power is much lower than when using ECV.

6 Conclusion

Perhaps the two most important issues for the implementation of our test are the selection of bandwidths (we have at least three to choose: h, h_0, and h_0^*) and the quality of the asymptotic (or bootstrap) approximations being used. With regard to the latter issue, there is clearly a difference between the results predicted by the asymptotics and those we have experienced in our simulations. Nevertheless, some of the tests performed quite reasonably with samples as small as $n = 100$; in practice, we would have much larger sample sizes and we expect our tests to perform somewhat better. Although it is widely believed that the curse of dimensionality worsens performance in estimation, in our testing problem we think this is not necessarily so because the larger the dimensionality the greater the difference in the null convergence rates of the restricted and unrestricted estimates. Provided bandwidth is chosen appropriately, it should be possible to maintain the same trade-off between size distortion and local power regardless of dimensions, see Fan and Linton (1997) for further discussion. The bandwidth issue is notoriously difficult to resolve even in much simpler situations. Nevertheless, simple rules for h and h_0 (derived from cross-validation or the ocular method) should work tolerably well, while taking h_0^* just a little larger than h_0 seems to work fine.

7 Appendix

Let $E_{|X_j}$ denote expectation conditional on X_j and let $E_{|X}$ denote expectation conditional on $X_1, \ldots, X_n$. We also use the symbol $\simeq$ to denote asymptotic equivalence in probability, thus

$X_n \simeq Y_n + Z_n$ means that $X_n = Y_n \{1 + o_p(1)\} + Z_n \{1 + o_p(1)\}$. Linton, Chen, Wang, and Härdle (1996, Lemma 1) showed that

$$\left\| \widetilde{m}_h - m^0 \right\|_\infty \leq \left\| \widehat{m}_h - m \right\|_\infty + O_p(n^{-1/2}). \tag{29}$$

Proof of Theorem 1. (i) We first examine $\widehat{\omega}_0$. By the Mean Value theorem,

$$\widehat{\omega}_0 = n^{-1} \sum_{j=1}^{n} \left[Q' \{\widetilde{m}(X_j)\} \{\widehat{m}_h(X_j) - \widetilde{m}_{h_0}(X_j)\} + \frac{1}{2} Q'' \{\overline{m}^*(X_j)\} \{\widehat{m}_h(X_j) - \widetilde{m}_{h_0}(X_j)\}^2 \right]^2 \pi(X_j),$$

where $\overline{m}^*(X_j)$ lie between $\widehat{m}_h(X_j)$ and $\widetilde{m}_{h_0}(X_j)$, $j = 1, \ldots, n$. Similarly, expand out $Q' \{\widetilde{m}_{h_0}(X_j)\}$ around $Q' \{m(X_j)\}$. Then, using Cauchy-Schwarz inequality, the boundedness of $Q(\cdot)$, and the uniform convergence results (21) and (29), we obtain

$$\widehat{\omega}_0 = n^{-1} \sum_{j=1}^{n} Q' \{m(X_j)\}^2 \{\widehat{m}_h(X_j) - \widetilde{m}_{h_0}(X_j)\}^2 \pi(X_j) + O_p(\varrho_n^3)$$

as in Linton and Härdle (1996). Write

$$\begin{aligned} \widehat{m}_h(X_j) - \widetilde{m}_{h_0}(X_j) &= \{\widehat{m}_h(X_j) - m(X_j)\} + \left\{m(X_j) - m^0(X_j)\right\} - \left\{\widetilde{m}_{h_0}(X_j) - m^0(X_j)\right\} \\ &\equiv q_1(X_j) + q_2(X_j) - q_3(X_j) \end{aligned}$$

and $\overline{\pi}(X_j) = \pi(X_j) Q' \{m(X_j)\}^2$. Some of the calculations below are best presented after $q_1(X_j)$ and $q_3(X_j)$ have been replaced by their linearized approximations. We have

$$q_1(X_j) \simeq \frac{h^q}{q!} \mu_q(k) b(X_j) + n^{-1} \sum_{i=1}^{n} K_h(X_i - X_j) \frac{u_i}{p(X_j)}, \quad j = 1, \ldots, n, \tag{30}$$

see Härdle (1991). Similarly, $q_3(X_j)$ can be decomposed into a stochastic and a deterministic component: following Linton and Härdle (1996), we have

$$q_3(X_j) \simeq n^{-1} \sum_{i=1}^{n} \overline{k}_{h_0}(X_j, X_i) u_i + \frac{h_0^q}{q!} \mu_q(k) b_0(X_j), \quad j = 1, \ldots, n, \tag{31}$$

where the one-dimensional 'equivalent kernel' is

$$\overline{k}_{h_0}(X_j, X_i) = F' \left[G \left\{ m(X_j) \right\} \right] \sum_{\alpha=1}^{d} k_{h_0}(X_{\alpha j} - X_{\alpha i}) G' \left\{ m(X_{\alpha j}, X_{\underline{\alpha} i}) \right\} \frac{p_{\underline{\alpha}}(X_{\underline{\alpha} i})}{p(X_{\alpha j}, X_{\underline{\alpha} i})}.$$

The approximation errors in (31) and (30) are of small enough order to be ignored; specifically, substituting the right hand sides of (31) and (30) into $\widehat{\omega}_0$ we get

$$\widehat{\omega}_0 \simeq n^{-1} \sum_{j=1}^{n} \left\{ U_1(X_j) + U_2(X_j) + U_3(X_j) - U_4(X_j) - U_5(X_j) \right\}^2 \overline{\pi}(X_j), \tag{32}$$

where

$$\begin{aligned}
U_1(X_j) &= n^{-1} \textstyle\sum_{i=1}^{n} K_h(X_i - X_j) u_i / p(X_j), \\
U_2(X_j) &= \tfrac{h^q}{q!} \mu_q(k) b(X_j) \\
U_3(X_j) &= \delta_n \lambda(X_j) \\
U_4(X_j) &= n^{-1} \textstyle\sum_{i=1}^{n} \overline{k}_{h_0}(X_j, X_i) u_i, \\
U_5(X_j) &= \tfrac{h_0^q}{q!} \mu_q(k) b_0(X_j).
\end{aligned}$$

To establish the result we will show below that:

$$\left\{ n h^{d/2} n^{-1} \sum_{j=1}^{n} U_1^2(X_j) \overline{\pi}(X_j) - \mu_{n0} \right\} \Big/ V_{n0}^{1/2} \quad \Rightarrow \quad N(0,1) \tag{33}$$

$$n h^{d/2} n^{-1} \sum_{j=1}^{n} U_3^2(X_j) \overline{\pi}(X_j) \quad = \quad \Delta_{n0} \cdot V_{n0}^{1/2} + o_p(1), \tag{34}$$

while the other terms are of smaller order, specifically:

$$nh^{d/2}n^{-1}\sum_{j=1}^{n}U_4^2(X_j)\overline{\pi}(X_j) = O_p(h^{d/2}h_0^{-1}) \quad (35)$$

$$nh^{d/2}n^{-1}\sum_{j=1}^{n}U_1(X_j)U_4(X_j)\overline{\pi}(X_j) = O_p(h^{d/2}h^{-1}) \quad (36)$$

$$nh^{d/2}n^{-1}\sum_{j=1}^{n}U_2^2(X_j)\overline{\pi}(X_j) = O_p(nh^{d/2}h^{2q}) \quad (37)$$

$$nh^{d/2}n^{-1}\sum_{j=1}^{n}\{U_1(X_j)-U_4(X_j)\}U_2(X_j)\overline{\pi}(X_j) = O_p(n^{1/2}h^{d/2}h^q) \quad (38)$$

$$nh^{d/2}n^{-1}\sum_{j=1}^{n}2U_2(X_j)U_3(X_j)\overline{\pi}(X_j) = O_p(nh^{d/2}h^q\delta_n). \quad (39)$$

Direct calculations similar to those we will use to prove (33)–(39) show that the other elements in (32) have means, variances and covariances with smaller order of magnitude, and can therefore be ignored. The result for $\widehat{\omega}_0$ then follows from (33)-(39).

Proof of (33): Write

$$\begin{aligned} & nh^{d/2}n^{-1}\sum_{j=1}^{n}U_1^2(X_j)\overline{\pi}(X_j) \\ = \;& n^{-1}\sum_{j=1}^{n}n^{-1}h^{d/2}\sum_{i=1}^{n}K_h^2(X_i-X_j)u_i^2\overline{\pi}(X_j)/p^2(X_j) \\ & +n^{-1}\sum_{j=1}^{n}n^{-1}h^{d/2}\sum\sum\nolimits_{i\neq\ell}K_h(X_i-X_j)K_h(X_\ell-X_j)u_iu_\ell\overline{\pi}(X_j)/p^2(X_j) \\ = \;& A_1+A_2, \end{aligned}$$

say, where $A_1 = n^{-1}h^{-d/2}\sum_{i=1}^{n}u_i^2a_{ni}$ and $A_2 = n^{-1}h^{-d/2}\sum\sum_{i\neq\ell}u_iu_\ell b_{ni\ell}$, with

$$\begin{aligned} a_{ni} &= \frac{1}{nh^d}\sum_{j=1}^{n}K^2\left(\frac{X_i-X_j}{h}\right)\frac{\overline{\pi}(X_j)}{p^2(X_j)} \\ b_{ni\ell} &= \frac{1}{nh^d}\sum_{j=1}^{n}K\left(\frac{X_i-X_j}{h}\right)K\left(\frac{X_\ell-X_j}{h}\right)\frac{\overline{\pi}(X_j)}{p^2(X_j)} \quad (i\neq\ell). \end{aligned}$$

Note that $b_{ni\ell}$ behaves rather like a d-dimensional kernel, say $L((X_i - X_\ell)/h)$ [forgetting the order one quantities $\overline{\pi}(X_j)$ and $p(X_j)$, it is really an empirical convolution of K with itself], in the sense that $E(b_{nil}^t) = O(h^d)$ for all $t \geq 1$, while $E(a_{nil}^t) = O(1)$ for all $t \geq 1$. We have

$$\begin{aligned} E_{|X}(A_1) &= h^{-d/2} \times \frac{1}{n}\sum_{i=1}^{n}\sigma^2(X_i)\left\{\frac{1}{nh^d}\sum_{j=1}^{n}K^2\left(\frac{X_i - X_j}{h}\right)\frac{\overline{\pi}(X_j)}{p^2(X_j)}\right\} \\ &= \mu_{n0} + o_p(1), \end{aligned}$$

$$\begin{aligned} \mathrm{var}_{|X}(A_1) &= \frac{1}{nh^d} \times \frac{1}{n}\sum_{i=1}^{n}E_{|X_i}\left[\left\{u_i^2 - \sigma^2(X_i)\right\}^2\right]a_{ni}^2 \\ &= O_p(n^{-1}h^{-d}). \end{aligned}$$

Consider A_2 now. Let $W_{i\ell n} = n^{-1}h^{-d/2}u_i u_\ell b_{ni\ell}$ $(i \neq \ell)$ and zero else, we write $A_2 = \sum_{i=1}^{n}\sum_{\ell=1}^{n} W_{i\ell n}$ which is a degenerate U-statistic, since $E(W_{i\ell n}|u_i) = E(W_{i\ell n}|u_\ell) = 0$. We can therefore use either Hall's (1984, Theorem 1) or deJong (1987, Theorem 2.1) central limit theorems for i.i.d. and i.n.i.d. degenerate U-statistics, respectively. In particular, following deJong (1987), $\{\mathrm{var}(A_2)\}^{-1/2} A_2 \to N(0,1)$ in distribution if

$$\max_{1\leq i\leq n}\sum_{\ell=1}^{n}\mathrm{var}(W_{i\ell n})/\mathrm{var}(A_2) \to 0, \tag{40}$$

and

$$E(A_2^4)/\{\mathrm{var}(A_2)\}^2 \to 3. \tag{41}$$

The proofs of (40) and (41) follow identical steps to those used by Härdle and Mammen (1993, p. 1943) to prove their conditions (7.7) and (7.8). First, we calculate

$$\mathrm{var}_{|X}(A_2) = E_{|X}\left(\sum_{i=1}^{n}\sum_{\ell=1}^{n}W_{j\ell n}\right)^2 = \sum_{i>\ell}\sum E_{|X}(2W_{i\ell n})^2 = 4\begin{pmatrix} n \\ 2 \end{pmatrix}E_{|X}(W_{12n}^2).$$

Straightforward calculations yield $E_{|X}(W_{12n}^2) = \sigma^2(X_1)\sigma^2(X_2)b_{n12}^2 n^{-2}h^{-d}$, where $E(b_{n12}^2) = O(h^d)$. Condition (40) follows directly from the fact that $\mathrm{var}(W_{i\ell n}) \leq \overline{c}n^{-2}h^{-d}$ for some finite

constant $\overline{c}$ due to the boundedness of K and the other conditions on p, π, and Q. For the proof of (41) we refer the reader to Härdle and Mammen (1993, p. 1943).

Proof of (35): The proof for this stochastic element of the restricted estimator follows identical steps to those used above to show (33). Firstly, we have

$$E_{|X}\left\{U_4^2(X_j)\right\} = n^{-2}\sum_{i=1}^{n}\overline{k}_{h_0}^2(X_j, X_i)\sigma^2(X_i).$$

Then note that $\overline{k}_{h_0}(X_j, X_i)$ really does behave just like a one-dimensional kernel, i.e.

$$E_{|X_j}\left\{\overline{k}_{h_0}(X_j, X_i)g(X_i)\right\} = O_p(1) \quad ; \quad E_{|X_j}\left\{\overline{k}_{h_0}^2(X_j, X_i)g(X_i)\right\} = O_p(h_0^{-1})$$

for integrable functions $g(\cdot)$. Therefore, taking expectations conditional on X_j and using identity of distribution, we have $E_{|X}\left\{U_4^2(X_j)\right\} = O_p(n^{-1}h_0^{-1})$ which implies that the mean of $nh^{d/2}n^{-1}\sum_{j=1}^{n}U_4^2(X_j)\overline{\pi}(X_j)$ is of order $h^{d/2}h_0^{-1} = o(1)$ [its variance is of order h^d/h_0].

Proof of (36): We have by identity of distribution:

$$E_{|X_j}\left\{U_1(X_j)U_4(X_j)\right\} = \frac{1}{n}\frac{E_{|X_j}\left\{\sigma^2(X_1)\overline{k}_{h_0}(X_j, X_1)K_h(X_j - X_1)\right\}}{p(X_j)}.$$

We must essentially examine the covariance between $\overline{k}_{h_0}(X_j, X_1)$ and $K_h(X_j - X_1)$. Letting $I = h_0^{-1}h^{-d}\int k[(X_{\alpha j} - X_{\alpha 1})/h_0]K[(X_j - X_1)/h]p(X_1)dX_1$, we have

$$\begin{aligned} I &= h_0^{-1}h^{-d}\int k\left(\frac{X_{\alpha j} - X_{\alpha 1}}{h_0}\right)k\left(\frac{X_{\alpha j} - X_{\alpha 1}}{h}\right)\underline{K}\left(\frac{X_{\underline{\alpha} j} - X_{\underline{\alpha} 1}}{h}\right)p(X_1)dX_1 \\ &= h_0^{-1}h^{-1}\int k\left(\frac{X_{\alpha j} - X_{\alpha 1}}{h_0}\right)k\left(\frac{X_{\alpha j} - X_{\alpha 1}}{h}\right)\underline{K}(u_{\underline{\alpha}})p(X_{\alpha 1}, X_{\underline{\alpha} j} + u_{\underline{\alpha}}h)du_{\underline{\alpha}}dX_{\alpha 1}, \end{aligned}$$

by the change of variables $(X_{\underline{\alpha} j} - X_{\underline{\alpha} 1})/h \to u_{\underline{\alpha}}$. Here, $\underline{K}(t_{\underline{\alpha}}) = \prod_{\beta\neq\alpha}k(t_\beta)$. Thus we can restrict our attention to integrals of the form

$$I = h_0^{-1}h^{-1}\int k\left(\frac{X_{\alpha j} - X_{\alpha 1}}{h_0}\right)k\left(\frac{X_{\alpha j} - X_{\alpha 1}}{h}\right)g(X_{\alpha 1})dX_{\alpha 1}$$

for suitable bounded continuous functions $g(\cdot)$. We make the change of variables $(X_{\alpha j} - X_{\alpha 1})/h_0 \to u_\alpha$ to write

$$\begin{aligned} |I| &= h^{-1}\left|\int k(h_0 u_\alpha/h)k(u_\alpha)g(X_{\alpha j}+u_\alpha h_0)du_\alpha\right| \\ &\leq h^{-1}\sup_u|k(u)|\sup_x|g(x)|\int|k(u)|du \end{aligned}$$

which is $O_p(h^{-1})$ by the boundedness of the suprema and integral. Therefore, the result is established by substituting back into (36).

Proof of (37): Direct calculation using the magnitude of $U_2(X_j)$. A similar result with h_0^q replacing h^q holds for the analogous term involving $U_5(X_j)$.

Proof of (38): Rewrite the left hand side of (38) as

$$nh^{d/2}h^q\frac{\mu_q(k)}{q!}\frac{1}{n}\sum_{i=1}^n u_i\left[\frac{1}{n}\sum_{j=1}^n\left\{\frac{K_h(X_i-X_j)}{p(X_j)}-\overline{k}_{h_0}(X_j,X_i)\right\}b(X_j)\overline{\pi}(X_j)\right]$$

by interchanging the order of summation. By direct calculation of the first two moments [first conditioning on $X_1,\ldots,X_n$],

$$\begin{aligned} \frac{1}{n^{1/2}}\sum_{i=1}^n u_i\left[\frac{1}{n}\sum_{j=1}^n K_h(X_i-X_j)\frac{b(X_j)\overline{\pi}(X_j)}{p(X_j)}\right] &= O_p(1) \\ \frac{1}{n^{1/2}}\sum_{i=1}^n u_i\left[\frac{1}{n}\sum_{j=1}^n \overline{k}_{h_0}(X_j,X_i)b(X_j)\overline{\pi}(X_j)\right] &= O_p(1), \end{aligned}$$

so that (38) is as claimed.

Proof of (39): Substituting for $U_2(X_j)$ and $U_3(X_j)$ we get

$$\begin{aligned} nh^{d/2}n^{-1}\sum_{j=1}^n 2U_2(X_j)U_3(X_j)\overline{\pi}(X_j) &= nh^{d/2}h^q\delta_n\frac{\mu_q(k)}{q!}n^{-1}\sum_{j=1}^n b(X_j)\lambda(X_j)\overline{\pi}(X_j)\{1+o_p(1)\} \\ &= O_p(nh^{d/2}h^q\delta_n) \end{aligned}$$

by the weak law of large numbers for independent random variables.

Proof of (34): Substituting for $U_3(X_j)$ we get

$$\begin{aligned} nh^{d/2}n^{-1}\sum_{j=1}^{n} U_3^2(X_j)\overline{\pi}(X_j) &= nh^{d/2}\delta_n^2 n^{-1}\sum_{j=1}^{n}\lambda^2(X_j)\overline{\pi}(X_j) \\ &= nh^{d/2}\delta_n^2\int \lambda^2(x)\overline{\pi}(x)p(x)dx + o_p(1) \\ &= \Delta_{n0}\cdot V_{n0}^{1/2} + o_p(1) \end{aligned}$$

by the weak law of large numbers for independent random variables. ■

(ii) We now deal with the test $\widehat{\omega}_1 = n^{-1}\sum_{i=1}^{n}\{\widehat{m}_h(X_i) - \widetilde{m}_{h_0}(X_i)\}\,\widetilde{u}_i\pi(X_i)$, where

$$\widetilde{u}_i \simeq u_i + U_3(X_i) - U_4(X_i) - U_5(X_i)$$

$$\widehat{m}_h(X_i) - \widetilde{m}_{h_0}(X_i) \simeq U_1(X_i) + U_2(X_i) + U_3(X_i) - U_4(X_i) - U_5(X_i)$$

as before. Therefore,

$$\begin{aligned} \widehat{\omega}_1 \simeq\ & n^{-1}\sum_{i=1}^{n}\{U_1(X_i) + U_2(X_i) - U_3(X_i) - U_4(X_i) - U_5(X_i)\}\,u_i\pi(X_i) + \\ & n^{-1}\sum_{i=1}^{n}\{U_1(X_i) - U_4(X_i) + U_3(X_i) + U_2(X_i) - U_5(X_i)\}\,\{U_3(X_i) - U_4(X_i) - U_5(X_i)\}\,\pi(X_i). \end{aligned}$$

The leading stochastic term is $n^{-1}\sum_{i=1}^{n} U_1(X_i)u_i\pi(X_i)$ which has mean $n^{-1}\sum_{i=1}^{n} w_{ii}\sigma^2(X_i)\pi(X_i)$ and variance $2n^{-2}\sum\sum_{i\neq j} w_{ji}^2\sigma^2(X_i)\sigma^2(X_j)\pi(X_i)\pi(X_j)$. The other ("new") terms satisfy

$$
\begin{aligned}
n^{-1}\sum_{i=1}^{n}\left\{U_2(X_i)-U_5(X_i)\right\}u_i\pi(X_i) &= O_p(h^q n^{-1/2})+O_p(h_0^q n^{-1/2}) \\
n^{-1}\sum_{i=1}^{n}U_3(X_i)u_i\pi(X_i) &= O_p(\delta_n n^{-1/2}) \\
n^{-1}\sum_{i=1}^{n}U_4(X_i)u_i\pi(X_i) &= O_p(h^{d/2}h_0^{-1}).
\end{aligned}
$$

The other terms have been handled above: specifically,

$$
n^{-1}\sum_{i=1}^{n}U_3^2(X_i)\pi(X_i)=\delta_n^2\int\lambda^2(x)\pi(x)p(x)dx\left\{1+o_p(1)\right\}.
$$

Everything else is smaller order in probability.

■

(iii) We now deal with $\widehat{\omega}_2$. We first substitute for $\widetilde{u}_i$ to obtain

$$
\begin{aligned}
\widehat{\omega}_2 &= \frac{1}{n^2h^d}\sum\sum_{i\neq j}K\left(\frac{X_i-X_j}{h}\right)u_iu_j\pi(X_i)\pi(X_j) \\
&\quad+\frac{2}{n^2h^d}\sum\sum_{i\neq j}K\left(\frac{X_i-X_j}{h}\right)\left\{\widetilde{m}_{h_0}(X_i)-m^0(X_i)\right\}u_j\pi(X_i)\pi(X_j) \\
&\quad+\frac{2}{n^2h^d}\sum\sum_{i\neq j}K\left(\frac{X_i-X_j}{h}\right)\left\{m^0(X_i)-m(X_i)\right\}u_j\pi(X_i)\pi(X_j) \\
&\quad+\frac{1}{n^2h^d}\sum\sum_{i\neq j}K\left(\frac{X_i-X_j}{h}\right)\left\{\widetilde{m}_{h_0}(X_i)-m^0(X_i)\right\}\left\{\widetilde{m}_{h_0}(X_j)-m^0(X_j)\right\}\pi(X_i)\pi(X_j) \\
&\quad+\frac{2}{n^2h^d}\sum\sum_{i\neq j}K\left(\frac{X_i-X_j}{h}\right)\left\{\widetilde{m}_{h_0}(X_i)-m^0(X_i)\right\}\left\{m^0(X_j)-m(X_j)\right\}\pi(X_i)\pi(X_j) \\
&\quad+\frac{1}{n^2h^d}\sum\sum_{i\neq j}K\left(\frac{X_i-X_j}{h}\right)\left\{m^0(X_i)-m(X_i)\right\}\left\{m^0(X_j)-m(X_j)\right\}\pi(X_i)\pi(X_j) \\
&\equiv Z_1+Z_2+Z_3+Z_4+Z_5+Z_6.
\end{aligned}
$$

The leading term is

$$Z_1 = \frac{1}{n^2 h^d} \sum\sum_{i \neq j} K\left(\frac{X_i - X_j}{h}\right) u_i u_j \pi(X_i)\pi(X_j) = O_p(n^{-1}h^{-d/2}).$$

In fact, $nh^{d/2}Z_1 \Rightarrow N(0, V_{n2})$, see Zheng (1996, Theorem 1). The noncentrality term comes from

$$\begin{aligned} Z_6 &= \delta_n^2 \frac{1}{n^2 h^d} \sum\sum_{i \neq j} K\left(\frac{X_i - X_j}{h}\right) \lambda(X_i)\lambda(X_j)\pi(X_i)\pi(X_j) \\ &= \delta_n^2 \left\{ \int \lambda^2(x)\pi^2(x)p^2(x)dx + o_p(1) \right\} \end{aligned}$$

by a U-statistic law of large numbers. Everything else is smaller order in probability.

■

PROOF OF THEOREM 2. We prove the result for $\widehat{\omega}_0^*$ [with $G = I$] and omit the arguments for the other tests since they are almost identical. The proof follows identical steps to those of Theorem 1. Let

$$\widehat{\omega}_0^* = n^{-1} \sum_{j=1}^{n} \left\{ \widehat{m}_h^*(X_j) - \widetilde{m}_{h_0}^*(X_j) \right\}^2 \pi(X_j) \tag{42}$$

and write

$$\widehat{m}_h^*(X_j) - \widetilde{m}_{h_0}^*(X_j) = \widehat{m}_h^*(X_j) - \widetilde{m}_{h_0^*}(X_j) + \widetilde{m}_{h_0^*}(X_j) - \widetilde{m}_{h_0}^*(X_j).$$

Using similar arguments to those used to derive (32), and noting that $\widehat{p}_h^*(x) = \widehat{p}_h(x)$ (since $X_i^* = X_i$ for all $i = 1, \ldots, n$), we obtain

$$\widehat{\omega}_0^* \simeq n^{-1} \sum_{j=1}^{n} \left\{ U_1^*(X_j) + U_2^*(X_j) - U_4^*(X_j) - U_5^*(X_j) \right\}^2 \pi(X_j), \tag{43}$$

where

$$
\begin{aligned}
U_1^*(X_j) &= n^{-1}\textstyle\sum_{i=1}^n K_h(X_i - X_j)u_i^*/p(X_j), \\
U_2^*(X_j) &= \tfrac{h^q}{q!}\mu_q(k) b_{\widetilde{m}_{h_0^*}}(X_j) \\
U_4^*(X_j) &= n^{-1}\textstyle\sum_{i=1}^n \overline{k}_{h_0}(X_j, X_i)u_i^*, \\
U_5^*(X_j) &= \tfrac{h_0^q}{q!}\mu_q(k) b_{0\widetilde{m}_{h_0^*}}(X_j),
\end{aligned}
$$

where $b_{\widetilde{m}_{h_0^*}}(\cdot)$ is the (unrestricted) bias function $b(\cdot)$ with $\widetilde{m}_{h_0^*}$ replacing m and $b_{0\widetilde{m}_{h_0^*}}$ is the restricted bias function $b_0(\cdot)$ again with $\widetilde{m}_{h_0^*}$ replacing m. We do not have a $U_3^*(X_j)$ element corresponding to $U_3(X_j)$ in the proof of Theorem 1 because Y_i^* is constructed so that $E^*(Y_i^*|X_i) = \widetilde{m}_{h_0^*}(X_i)$. This in turn implies that the bootstrap will approximate the distribution of the test under the null only.

Firstly, because u_i^* are (given the data) mean zero, mutually independent and have variance $\mathrm{var}(u_i^*) = \widetilde{u}_i^2$, we have

$$
\left\{ nh^{d/2}n^{-1}\sum_{j=1}^n U_1^*(X_j)^2\pi(X_j) - \mu_{n0}^* \right\} \Big/ (V_{n0}^*)^{1/2} \Rightarrow N(0,1), \tag{44}
$$

where

$$
\begin{aligned}
\mu_{n0}^* &= h^{d/2}\sum_{j=1}^n\sum_{i=1}^n w_{ji}^2\widetilde{u}_i^2\pi(X_j) = \mu_{n0} + o_p(1) \\
V_{n0}^* &= 2h^d\sum\sum_{i\neq j}\rho_{ji}^2\widetilde{u}_i^2\widetilde{u}_j^2\pi(X_i)\pi(X_j) = V_{n0} + o_p(1),
\end{aligned}
$$

by the uniform consistency of $\widetilde{m}_{h_0}$. Similar arguments show that $U_4^*(X_j)$ contributes terms to $\widehat{\omega}_0^*$ like $U_4^*(X_j)$ does to $\widehat{\omega}_0$ of order $h^{d/2}/h_0$ in probability.

We next calculate the contributions of the two terms $U_2^*(X_j)$ and $U_5^*(X_j)$ using the approach of Härdle and Marron (1991). Note that $b_{\widetilde{m}_{h_0^*}}(X_j)$ is a linear combination of the partial derivatives of $\widetilde{m}_{h_0^*}(X_j)$ up to and including the q'th order, and is of the form $\sum_i \widetilde{w}_{ji}^{(q)}(h_0^*)Y_i$ for some weighting sequence $\left\{\widetilde{w}_{ji}^{(q)}(h_0^*)\right\}$; likewise, $b_{0\widetilde{m}_{h_0^*}}(X_j) = \sum_i \widetilde{w}_{0ji}^{(q)}(h_0^*)Y_i$ for weights $\left\{\widetilde{w}_{0ji}^{(q)}(h_0^*)\right\}$. As for $\widehat{m}$ and $\widetilde{m}$, both these quantities can be divided into an $O(1)$ mean and a stochastic term which is $O_p(n^{-1/2}h_0^{*-(2q+1)/2})$ as in Severance-Lossin and Sperlich (1995).[10] These magnitudes carry over to $\widehat{\omega}_0^*$ using the same methods of Theorem 1: we get terms (in $\widehat{\omega}_0^*$) of order $h^{2q}h^{d/2}\Big/h_0^{*(2q+1)}$ and $h_0^{2q}h^{d/2}\Big/h_0^{*(2q+1)}$ from the stochastic terms. ■

Acknowledgements

We thank Joel Horowitz, Stefan Sperlich, and Dag Tjøstheim for helpful discussions, and the National Science Foundation for financial support. A Fortran program is available (from the first author) which carries out the procedures described in this paper.

References

Aït-Sahalia, Y., P.J. Bickel, and T.M. Stoker (1994): "Goodness-of-fit for regression using kernel methods," MIT working paper #3747.

Azzalini, A., A.W. Bowman, and W. Härdle (1989): "On the use of nonparametric regression for model checking," *Biometrika* **76**, 1-11.

Bierens, H.J., and W. Ploberger (1996): "Asymptotic theory of integrated conditional moment tests," Southern Methodist University working paper. Forthcoming in *Econometrica.*

[10] Note that because the bias terms play no part in the asymptotic properties of $\widehat{\omega}_0$ we do not need $U_2^*(X_j)$ to consistently estimate the bias of $\widehat{m}_h(X_j)$, i.e. $U_2(X_j)$, nor $U_5^*(X_j)$ to consistently estimate the bias of $\widetilde{m}_{h_0}(X_j)$, which is why we have not had to use additional smoothness properties.

Buja, A., T. Hastie and R. Tibshirani. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics* **17**, 453-555.

Deaton, A. and J. Muellbauer (1980): *Economics and Consumer Behavior*. Cambridge University Press: Cambridge.

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

Fan, Y., and Q. Li (1996): "Consistent model specification tests: Omitted variables and semiparametric functional forms," *Econometrica* **64**, 865-890.

Fan, Y., and O. Linton (1997): "Some higher order theory for a consistent model specification test," *Cowles Foundation Discussion Paper* no. 1148.

Gourieroux, C., and C., Tenreiro (1994): "A comparison of kernel based goodness of fit tests," CREST working paper.

Gozalo, P.L. (1993) "A Consistent Model Specification Test for Nonparametric Estimation of Regression Function Models," *Econometric Theory*, **9**, 451-477.

Gozalo, P.L. (1995) "Nonparametric Specification Testing with $\sqrt{n}$-Local Power and Bootstrap Critical Values," Brown University Working Paper 95-21.

Hall, P. (1984): "Central limit theorem for integrated square error of multivariate nonparametric density estimators," *Journal of Multivariate Analysis* **14**, 1-16.

Härdle, W. and E. Mammen (1993) "Comparing Nonparametric versus Parametric Regression Fits," *Annals of Statistics*, **21**, 1926-1947.

Härdle, W. and E. Marron (1991) "Bootstrap Simultaneous Error Bars for Nonparametric Regression," *Annals of Statistics*, **19**, 778-796.

Härdle, W. and L. Yiang (1996) "Nonparametric Autoregression with Multiplicative Volatility and Additive Mean," Forthcoming in *Journal of Time Series Analysis*.

Hastie, T. and R. Tibshirani (1991). Generalized Additive Models. *Chapman and Hall, London*.

Hidalgo, J. (1994): "A nonparametric test for nested and non-nested models," Manuscript, LSE.

Hjellvik, V and D. Tjøstheim (1995): "Nonparametric tests of linearity for time series. *Biometrika* **82**, 351-68.

Hong, Y., (1993): "Consistent specification testing using optimal nonparametric kernel estimation." Unpublished Manuscript, Cornell University.

Hong, Y., and H. White (1995): "Consistent specification testing via nonparametric series regression, *Econometrica* **63**, 1133-1159.

Horowitz, J., (1995): "Bootstrap methods in econometrics: Theory and numerical performance," in *Advances in Economics and Econometrics: 7th World Congress*, D. Kreps and K.W. Wallis, eds., Cambridge: Cambridge University Press, Forthcoming.

Ichimura, H., (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics* **58**, 71-120.

Jones, M.C., O.B. Linton, and J.P. Nielsen (1995): "A simple bias reduction method for density estimation," *Biometrika* **82**, 327-338.

Lavergne, P. and Q. Vuong, (1996): "Nonparametric Specification Testing," Unpublished Manuscript, INRA, University of Toulouse.

Leontieff, W. (1947). "Introduction to a theory of an internal structure of functional relationships," *Econometrica*, **15**, 361-373.

Linton, O.B., R. Cheng, N. Wang, and W. Härdle. (1996). An analysis of transformations for additive nonparametric regression," Forthcoming in *Journal of the American Statistical Association.*

Linton, O.B. and W. Härdle. (1996). "Estimating additive regression models with known links," *Biometrika*, **83**, 529-540.

Linton, O.B. and J.P. Nielsen. (1995). "Estimating structured nonparametric regression by the kernel method," *Biometrika*, **82**, 93-101.

Luce, R.D., and J.W. Tukey (1964): "Simultaneous conjoint measurement: a new type of fundamental measurement." *Journal of Mathematical Psychology* **1**, 1-27.

Masry, E., (1996): "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis* **17**, 571-599.

McCullagh, P., and J.A. Nelder (1989) *Generalized Linear Models*, 2nd edition. Chapman and Hall.

Newey, W.K. (1994). Kernel estimation of partial means," *Econometric Theory*, **10**, 233-253.

Nielsen, J.P. (1996): "Estimating Multiplicative Hazard Functions," Manuscript, PFA Pension.

Severance-Lossin, E., and S. Sperlich (1995): "Estimation of derivatives for additively separable models," SFB 373 Discussion Paper no. 60.

Staniswalis, J. G. and T.A. Severini (1991) "Diagnostics for Assessing Regression Models," *Journal of the American Statistical Association*, **86**, 684-692.

Stone, C.J. (1985). Additive regression and other nonparametric models," *Annals of Statistics*, **13**, 685-705.

STONE, C.J. (1986). The dimensionality reduction principle for Generalized additive models. *Annals of Statistics,* **14**, 592-606.

TJØSTHEIM, D., AND B. AUESTAD (1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association*, **89**, 1398-1409.

WU, C.F.J. (1986): "Jacknife, Bootstrap and other resampling methods in regression analysis," (with discussion) *The Annals of Statistics* **14**, 1261-1295.

ZHENG, J.X. (1996): "A consistent test of functional form via nonparametric estimation techniques," *Journal of Econometrics*, **75**, 263-289.