ELSEVIER

Journal of Econometrics 105 (2001) 363–412

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Goodness-of-fit tests for kernel regression with an application to option implied volatilities

Yacine Aït-Sahalia[a,*], Peter J. Bickel[b], Thomas M. Stoker[c]

[a] *Department of Economics, Princeton University, Princeton, NJ 08544-1021, USA*
[b] *Department of Statistics, University of Calfornia, Berkeley, CA 94720-3860, USA*
[c] *Sloan School of Management, MIT, Cambridge, MA 02142-1347, USA*

Received 28 February 2000; revised 13 March 2001; accepted 30 April 2001

Abstract

This paper proposes a test of a restricted specification of regression, based on comparing residual sum of squares from kernel regression. Our main case is where both the restricted specification and the general model are nonparametric, with our test equivalently viewed as a test of dimension reduction. We discuss practical features of implementing the test, and variations applicable to testing parametric models as the null hypothesis, or semiparametric models that depend on a finite parameter vector as well as unknown functions. We apply our testing procedure to option prices; we reject a parametric version of the Black–Scholes formula but fail to reject a semiparametric version against a general nonparametric regression.

JEL classification: C12; C14

Keywords: Goodness-of-fit; Kernel regression; Specification testing; Implied volatility smile

1. Introduction

A primary role of hypothesis testing in empirical work is to justify model simplification. Whether one is testing a restriction implied by economic theory or an interesting behavioral property, the test asks whether imposing the restriction involves a significant departure from the data evidence. A failure

* Corresponding author. Tel.: 1-609-258-4015; fax: 1-609-258-0719.
E-mail address: yacine@princeton.edu (Y. Aït-Sahalia).

0304-4076/01/$ - see front matter
PII: S0304-4076(01)00091-4

to reject implies that the restriction can be imposed without inducing a significant departure, or that the original model can be simplified.[1] Moreover, a simple model is typically easier to understand and use than a complicated model, and therefore can be more valuable for scientific purposes, provided that it is not in conflict with the available evidence. Whether one is testing for the equality of means from two populations, or whether a linear regression coefficient is zero, the aim is to produce a simpler model for subsequent applications.

When the methods of analysis are widened to include nonparametric techniques, the need for model simplification is arguably even more important than with parametric modeling methods. Nonparametric methods permit arbitrarily flexible depictions of data patterns to be estimated. There can be a cost of this flexibility, in terms of data summary and interpretation. If the regression of a response on three or more predictor variables is of interest, then a nonparametric estimator of that regression can be extremely hard to summarize (for instance, graphically). As such, the need for model simplification is paramount—without simplification one may have a hard time even communicating the results of the statistical analysis. But such simplification should be statistically justified and for that hypothesis tests are needed.

In this paper we propose a test of regression model simplification in the nonparametric context. As a base case, we consider the most basic situation of dimension reduction; namely whether certain predictor variables can be omitted from the regression function. Our results apply to the situation where kernel methods are used to estimate the regression under the restricted specification and under the alternative—in our base case, we compare kernel regression estimates that exclude the variables of interest with those that include those variables. Our test statistic measures goodness-of-fit directly, by comparing regression estimates under the null and alternative.

We feel that our testing approach has many attractive aspects for practitioners. We feel it is natural to compare fitted values with and without the restriction of interest, and we indicate how our test statistic is equivalently written as the difference in residual sums-of-squares under the null and alternative hypotheses. As such, we show how the familiar fit comparison that underlies F-tests in linear regression can be applied in very general nonparametric contexts. Interpretation of the results is as straightforward as possible; rejection means that restricted estimates are significantly different from unrestricted estimates; that the sum of squared residuals is significantly larger under the restriction.

We include an analysis of data on option implied volatilities. This illustrates the use of our test statistic to compare various parametric and semi-

[1] This logic applies to nested hypothesis testing—we do not consider non-nested hypothesis tests, or comparisons of substantially different models.

parametric model specifications to general nonparametric regression, showing how our procedure can guide the selection of the most parsimonious, statistically adequate model. A partial linear, semiparametric specification is the only restricted specification that is not rejected against general nonparametric regression. For another illustration, see Blundell and Duncan (1998), who use our test statistic in a similar way to analyze the structure of income effects in British household demand data. A further example is provided by Christoffersen and Hahn (1998), who examine the relevance of ARCH models by using our test statistic to test whether ARCH volatility has additional explanatory power given other relevant variables. Yet another set of applications of our test are contained in Fraga (1999), who tests parametric and semiparametric sample selection models against simpler alternatives to model female labor supply, and in Fernandes (1999), who applies it to test for Markovian dynamics. Labor economics applications often require limited dependent variables; finance ones require time series. Both are allowed in our approach.

The attractive practical aspects of our test statistic have a cost in terms of the technical analysis. Kernel regression estimators are ratios of local averages, which makes for a relatively complicated technical development and proof. Likewise, we need to restrict attention to data areas with sufficient density, or compact support. This latter feature is accommodated by the use of a weighting function, that can be used either to restrict the testing focus to certain subsets of data, or to trim out sparse data areas (low data density).[2]

To describe the technical development, we derive the expansion of the functional for the sum-of-squared departures between the restricted regression and the unrestricted regression, according to the method of von Mises (1947).[3] We carry out the second order expansion, because the distribution of our test statistic has a singularity when the null hypothesis is true.[4] In particular, the first order (influence) terms vanish under the null hypothesis,

[2] Requiring the data density to be bounded away from zero is a familiar requirement for analyzing kernel regression estimators. Analogous conditions are used in much nonparametric work, as we do here. As we discuss later, we believe the only way to avoid this requirement is to use squared-density weighting, but then testing could not easily be performed on subsets of the data. It is also the case that in many empirical settings the density of the predictor variables does in fact have compact support, in which case compact support of the weighting is not a limitation. For instance, in our option pricing application, the option exchanges always trade a finite (and known) range of strike prices and maturities. In the labor economics applications we cited above, the number of hours worked, years of schooling, number of children, etc., have finite ranges. We do allow for limited dependent variables.

[3] See e.g., Filippova (1962), Reeds (1976), Fernholz (1983), and Aït-Sahalia (1996, 1998).

[4] This kind of structure has been noted for other kinds of testing procedures, for instance, see Bickel and Rosenblatt (1973) for density estimation, Hall (1984), Fan (1994), Fan and Li (1996), Hong and White (1995), and Bierens and Ploberger (1997), among others. We are not aware however, of applications of the von Mises expansion to the testing problem that we carry out. Aït-Sahalia (1998) provides an analysis of the first and second order distributions for kernel estimators using functional derivatives.

and our distributional result is based on the next term in the expansion. To derive the distribution of the second order term, we utilize results from the theory of U-statistics that are applicable in situations where the influence terms vanish.

We focus on the base case of dimension reduction because it captures the substantive features of the distributional structure for test statistics based on goodness-of-fit. The variation of the test statistic is primarily related to the unrestricted model—for instance, the rate of convergence of the test statistic is determined by the rates applicable to nonparametric estimation of the general model. We present several corollaries that deal with important practical variations and reinforce the basic structure of the test statistic. In particular, we verify that the distribution of the test statistic is unchanged when the restricted model depends on a (finite dimensional) parameter vector, and the test is performed using an estimate of the parameter. We specialize that case further to the situation where the restricted model is a parametric model, which does not involve any nonparametric estimation at all. Finally, we discuss problems of dependent mixing observations. These corollaries cover many applications involving tests of parametric or semiparametric models against general nonparametric regression, in a wide range of different data situations.

Our procedure allows tests of a nonparametric null hypothesis against a nonparametric alternative. The most closely related articles in the literature are Fan and Li (1996) and Lavergne and Vuong (2000), which also permit such comparisons. Fan and Li (1996) avoid many of the technical difficulties we encounter by using "leave-out" estimators and employing squared-density weighting; namely they choose a specific weighting function, the square of the data density. While much simpler technically, we view weighting by the square of the data density as focusing the testing on a relatively small, high density area of data. Lavergne and Vuong (2000) also propose a procedure with squared density weighting, but with better bias properties than that of Fan and Li (1996). In any case, these papers have an alternative to our procedure that permits a much simpler distributional analysis, but does not permit the testing to be performed on analyst-specified subsets of the data. We discuss squared-density weighting in Section 3.4.3 and show throughout how our assumptions and results should be modified in that case; we just note here that such weighting appears to be the only possibility for test statistics based on a nonparametric regression comparison that avoids our compact support requirement. Other papers that allow tests of a nonparametric null hypothesis against a nonparametric alternative include Gozalo (1993), Yatchew (1992) and Lavergne and Vuong (1996).

If we consider tests of a specific parametric model against nonparametric alternatives, many articles are relevant. As we discuss below, tests involving nested hypotheses require analysis of second order terms in the asymptotic expansion, because the first order terms vanish under the null hypothesis.

Wooldridge (1992) analyzes such first order terms for testing a linear model against a nonnested nonparametric alternative; see also Eubank and Spiegelman (1990), and Doksum and Samarov (1995). Whang and Andrews (1991), Yatchew (1992) and Lavergne and Vuong (1996) propose methods based on "sample splitting", or using different parts of the original data sample for estimation and testing. Also related is the work based on the cross validation function used for choosing smoothing parameters in nonparametric regression: see Zhang (1991) and Vieu (1994). Tests of orthogonality restrictions implied by parametric models are proposed by Bierens (1990), Lewbel (1991), and White and Hong (1993). Lee (1988) proposes tests of nested hypotheses using weighted residuals, Gozalo (1993) examines testing with a fixed grid of evaluation points, and Rodriguez and Stoker (1992) propose a conservative test of nested hypotheses based on estimating first order terms given the value of the smoothing parameters used in nonparametric estimation. A recent analysis of second order terms of tests of parametric models against general nonparametric alternatives was carried out by Bierens and Ploberger (1997), de Jong and Bierens (1994), Chen and Fan (1999) and Li (1999) for general orthogonality restrictions, by Zheng (1996) for tests of parametric models and Ellison and Fisher-Ellison (2000) for tests of linear models. An analysis of other kinds of 'density-weighted' test statistics using kernel estimators is given in Staniswalis and Severini (1991), Hidalgo (1992), White and Hong (1996) and Li (1999). Further work includes the test of Härdle and Mammen (1993) of a parametric model versus a nonparametric alternative, and the test of Horowitz and Härdle (1994) of parametric index models versus semiparametric index models. Finally, Heckman et al., (1998) analyze a test of index sufficiency using local linear regression estimators, and relate their approach to the goodness-of-fit method that we analyze here.

In sum, our work contributes to this literature by considering a general regression testing situation including choice of regressors (dimension reduction) as well as parametric and semiparametric null hypotheses, as well as an analysis of the second order terms that arise with kernel regression estimators. Our results are applicable quite generally (including limited dependent variables and/or time series observations) and cover most testing situations considered in the papers cited above. Further, our test focuses on a goodness-of-fit statistic that is natural and easy to interpret, with no need to choose arbitrary moments, etc., although some regularization parameters need to be specified as is intrinsically the case with every nonparametric procedure.

2. Basic framework

Suppose we study a response Y as a function of a vector of real-valued predictor variables (W, V) where W and V are vectors of dimension p and

q, respectively. The data sample consists of N independent and identically distributed observations (Y_i, W_i, V_i), $i=1,\ldots,N$. Our base case concerns testing whether the predictor variables V can be omitted from the regression of Y on (W,V). For that purpose, we will be comparing the regression of Y on (W,V) to the regression of Y on W alone.

The joint density (resp. cumulative distribution function) of (y,w,v) is denoted by f (resp. F). Below, we need to make reference to several marginal densities from $f(y,w,v)$ which we denote via the list of arguments—for example $f(w,v) \equiv \int f(y,w,v)\,\mathrm{d}y$ and $f(w) \equiv \int\int f(y,w,v)\,\mathrm{d}y\,\mathrm{d}v$ where $\int$ denotes integration on the full range of the argument. While this notation is compact, we hope that it is sufficiently unambiguous. The regression function of Y on (W,V) is defined by

$$m(w,v) \equiv \mathrm{E}(Y|W=w, V=v) = \frac{\int y f(y,w,v)\,\mathrm{d}y}{f(w,v)} \tag{2.1}$$

and the regression function of Y on W by

$$M(w) \equiv \mathrm{E}(Y|W=w) = \frac{\int y f(y,w)\,\mathrm{d}y}{f(w)}. \tag{2.2}$$

for all (w,v) such that $f(w,v) > 0$ and $f(w) > 0$.

We are interested in whether V can be omitted from the regression of Y on (W,V), namely the null hypothesis is

$$\mathrm{H}_0\colon\ \Pr[m(W,V) = M(W)] = 1.$$

The alternative hypothesis is that $m(w,v) \neq M(w)$ over a significant range, or

$$\mathrm{H}_1\colon\ \Pr[m(W,V) = M(W)] < 1.$$

Our testing approach is to assess the significance of squared differences in nonparametric kernel estimates of the functions m and M.

We first introduce the kernel estimators, and then the test statistic of interest. For a kernel function K and bandwidth h, we define

$$K_h(u) \equiv h^{-d} K(u/h),$$

where d is the dimension of the vector u.[5] The standard Nadaraya–Watson kernel regression estimator of $m(w,v)$ is

$$\hat{m}_h(w,v) \equiv \frac{\sum_{i=1}^{N} K_h(w-W_i, v-V_i) Y_i}{\sum_{i=1}^{N} K_h(w-W_i, v-V_i)}, \tag{2.3}$$

[5] To keep the notation simple, we do not explicitly indicate the dependence of the bandwidth parameters h on the sample size N. We also adopt the same notational convention for K as for f, namely to indicate which kernel by the list of arguments.

while

$$\hat{M}_H(w) \equiv \frac{\sum_{i=1}^N K_H(w - W_i)Y_i}{\sum_{i=1}^N K_H(w - W_i)} \tag{2.4}$$

is the standard estimator of $M(w)$ with bandwidth H, and $K_H(u) \equiv H^{-d}K(u/H)$. For simplicity only, we will take the multivariate kernel function to be a product of the univariate kernel functions K.

It is convenient to introduce the two density estimates:

$$\hat{f}_h(y,w,v) = \frac{1}{N}\sum_{i=1}^N K_h(y - Y_i, w - W_i, v - V_i),$$

$$\hat{f}_H(y,w) = \frac{1}{N}\sum_{i=1}^N K_H(y - Y_i, w - W_i).$$

Similarly, we define the estimates $\hat{f}_h(w,v)$ of $f(w,v)$, calculated with bandwidth h, and $\hat{f}_H(w)$ of $f(w)$, calculated with bandwidth H. Finally, let $\hat{F}(w,v)$ be the empirical cumulative distribution estimate.

Consider a nonnegative weighting function $a(w,v)$ and define $A_i \equiv a(W_i,V_i)$. The form of test statistics we consider is

$$\tilde{I} \equiv \frac{1}{N}\sum_{i=1}^N \{\hat{m}_h(W_i,V_i) - \hat{M}_H(W_i)\}^2 A_i, \tag{2.5}$$

that is, we compare $\hat{M}_H(w)$ to $\hat{m}_h(w,v)$ by their squared difference weighted by $a(w,v)$. Notice that $\tilde{I}$ is an estimate of $\mathrm{E}[(m(W,V) - M(W))^2 a(W,V)]$.

The introduction of the weighting function $a(w,v)$ allows us to focus goodness-of-fit testing on particular ranges of the predictor variables. By choosing an appropriate $a(w,v)$, specification tests can be tailored to the empirical question of interest. For example, in our application to option prices in Section 6, we may be interested in differences only for low values of W, corresponding to out-of-the-money put options.

Specifically, we will consider functions $a(w,v)$ which are bounded with compact support $S \subset R^{p+q}$ strictly contained in the support of the density $f(w,v)$; in typical examples, a will be the indicator function of a compact set S. As a result, we can only detect deviations between $m(w,v)$ and $M(w)$ that arise on S.[6] Another approach to this general technical problem is to use a clever but specific data-dependent choice of weighting function proposed by Fan and Li (1996). That is, their approach is equivalent to setting $A_i \equiv \hat{f}_h(W_i,V_i) \times \hat{f}_H^2(W_i)$, which obviously weights heavily towards the empirical

[6] Our results extend to sequences $a_N(w,v)$, each with compact support, converging to a weighting function $a(w,v)$ with unbounded support.

mode(s) of the predictor variables. Kernel regression estimators are ratios of local averages which in general make for a relatively complicated technical analysis, whereas specific density-based weighting, such as that just described, can effectively cancel out the denominator and greatly simplify the analysis at the cost of constraining the empirical researcher to a specific weighting of the squared deviations. We choose instead to deal with an arbitrary weighting function $a(w,v)$ but then must restrict attention to compact support weighting (or to sequences with compact support converging to a weighting function with unbounded support) in order to avoid estimating conditional expectations in areas of low density of the conditioning variables, where the estimates are very imprecise. We discuss this further in Section 3.4.

We will show that the properties of our test statistic can be derived from the properties of the squared error goodness-of-fit functional

$$\Gamma(f_1, f_2, F_3) \equiv \iint \left\{ \frac{\int y f_1(y,w,v)\,\mathrm{d}y}{f_1(w,v)} - \frac{\int y f_2(y,w)\,\mathrm{d}y}{f_2(w)} \right\}^2 a(w,v)\,\mathrm{d}F_3(w,v), \tag{2.6}$$

where f_1 and f_2 range over convex sets of densities on R^{1+p+q} and R^{1+q}, respectively, which will be defined precisely in the statement of Lemma 1 and F_3 is a (nondegenerate but otherwise unrestricted) cumulative distribution function on R^{p+q}.

Under the null hypothesis H_0, we have $\Gamma(f,f,F_3)=0$ for all F_3, and under the alternative H_1, we have $\Gamma(f,f,F_3)>0$ for some F_3.[7] Further, our test statistic is

$$\tilde{\Gamma} = \Gamma(\hat{f}_h, \hat{f}_H, \hat{F}).$$

Since $\hat{f}_h$, $\hat{f}_H$ and $\hat{F}$ are all close to their population counterparts, a test of H_0 using $\tilde{\Gamma}$ can be expected to be consistent (against alternatives where $m(w,v)\neq M(w)$ on S), as we shall show.[8]

Since the restricted regression is nested in the general regression as

$$M(w) = \mathrm{E}[m(W,V)|W=w]$$

[7] The notation $\Gamma(f,f,F_3)$ means that we are evaluating the functional $\Gamma(.,.,.)$ at $f_1(y,w,v)=f(y,w,v)$ and $f_2(y,w)=f(y,w)$, the true densities, and the cumulative density function $F_3(w,v)$.

[8] It is clear that our testing approach is a nonparametric analogue to traditional χ^2 and F-tests of coefficient restrictions in linear regression analysis, and the same intuitive interpretation carries over from the parametric context. Specifically, an F-test is performed by using an estimate of $\Gamma(f,f,F)$ divided by an estimate of residual variance, scaled for degrees of freedom.

by the law of iterated expectations, we have an alternative formulation of the goodness-of-fit functional. Specifically

$$\mathrm{E}[((Y - M(W))^2 - (Y - m(W,V))^2)|W = w, V = v]$$
$$= (m(w,v) - M(w))^2$$

and therefore

$$\Gamma(f, f, F) = \mathrm{E}[((Y - M(W))^2 - (Y - m(W,V))^2)a(W,V)] \tag{2.7}$$

so that $\Gamma(f, f, F) = 0$ is associated with no improvement in residual variance (or least squares goodness-of-fit) from including V in the regression analysis of Y on W.

3. The distribution of the test statistic

3.1. Assumptions and the main result

Our assumptions are as follows.

Assumption 1. The data $\{(Y_i, W_i, V_i);\ i = 1, \ldots, N\}$ are i.i.d. with cumulative distribution function $F(y,w,v)$ and density $f(y,w,v)$.

Assumption 2. We have that:

1. The density $f(y,w,v)$ is $r+1$ times continuously differentiable, $r \geqslant 2$. f and its derivatives are bounded and square-integrable. Let $\mathscr{D}$ denote the space of densities with these properties.
2. $f(w,v)$ is bounded away from zero on the compact support S of a. Hence $\inf_S f(w,v) \equiv b > 0$.[9] If we let PS denote the projection of the support S for the w dimension, i.e., $PS \equiv \{w\colon \exists v \,|\, (w,v) \in S\}$, it follows that $f(w)$ is bounded away from zero on PS.
3. $\mathrm{E}[(Y - m(W,V))^4] < \infty$
   and
   $$\sigma^2(w,v) \equiv \mathrm{E}[(Y - m(W,V))^2|W = w, V = v] \tag{3.1}$$
   is square-integrable on S. The restricted conditional variance
   $$\sigma^2(w) \equiv \mathrm{E}[(Y - M(W))^2|W = w] \tag{3.2}$$
   is square-integrable on PS.

[9] In the case of squared-density weighting, this is not required because the denominators in the kernel regression are effectively eliminated. All conditions involving S are dropped.

Assumption 3. For kernel estimation:

1. The kernel K is a bounded function on R, symmetric about 0, with $\int |K(z)|\,dz < \infty$, $\int K(z)\,dz = 1$, $\int z^j K(z)\,dz = 0$ for $1 \leqslant j < r$. Further,
$$r > 3(p+q)/4. \tag{3.3}$$
2. As $N \to \infty$, the unrestricted bandwidth sequence $h = O(N^{-1/\delta})$ is such that
$$2(p+q) < \delta < 2r + (p+q)/2, \tag{3.4}$$
while the restricted bandwidth $H = O(N^{-1/\Delta})$ satisfies
$$p < \Delta \leqslant 2r + p \tag{3.5}$$
as well as
$$\delta p/(p+q) \leqslant \Delta < \delta. \tag{3.6}$$

Note from (3.3) that there is no need to use a high-order kernel ($r > 2$) unless the dimensionality of the unrestricted model, $p+q$, is greater than or equal to 3. Under the assumptions made on the bandwidth sequence, we have in particular that $Nh^{(p+q)/2+2r} \to 0$, $Nh^{p+q} \to \infty$, $NH^p \to \infty$, $NH^{p+2r} \to R$ for some $0 \leqslant R < \infty$, $H/h \to 0$ and $h^{(p+q)}/H^p \to 0$. So asymptotically we have that $h^p \gg H^p \gg h^{p+q}$.

Our main result is that the test statistic is asymptotically normally distributed with an asymptotic bias. For stating the result and giving the derivation, we define the further notation

$$\alpha(y,w,v) \equiv [y - m(w,v)]/f(w,v), \quad \beta(y,w) \equiv [y - M(w)]/f(w), \tag{3.7}$$

$$\gamma_{12} \equiv C_{12} \iint \left[\int \alpha(y,w,v)^2 f(y,w,v)\,dy\right] f(w,v)a(w,v)\,dw\,dv$$
$$= C_{12} \iint \sigma^2(w,v)a(w,v)\,dw\,dv, \tag{3.8}$$

$$\gamma_{22} \equiv -2C_{22} \iint \left[\int \alpha(y,w,v)\beta(y,w)f(y,w,v)\,dy\right] f(w,v)a(w,v)\,dw\,dv$$
$$= -2C_{22} \iint \sigma^2(w,v)\{f(w,v)/f(w)\}a(w,v)\,dw\,dv, \tag{3.9}$$

$$\gamma_{32} \equiv C_{32} \iint \left[\int \beta(y,w)^2 f(y,w)\,dy\right] f(w)a(w)\,dw$$
$$= C_{32} \int \sigma^2(w)a(w)\,dw, \tag{3.10}$$

$$\sigma_{11}^2 \equiv 2C_{11} \iint \left[\int \alpha(y,w,v)^2 f(y,w,v)\,\mathrm{d}y \right]^2 f(w,v)^2 a(w,v)^2\,\mathrm{d}w\,\mathrm{d}v$$
$$= 2C_{11} \iint \sigma^4(w,v) a(w,v)^2\,\mathrm{d}w\,\mathrm{d}v, \tag{3.11}$$

where $a(w) \equiv \mathrm{E}[a(W,V)|W=w] = \int a(w,v) f(w,v)\,\mathrm{d}v / f(w)$. In each of these equations, the equality results from repeated application of the law of iterated expectations. The C_{ij}'s are constants determined by the kernel function as

$$C_{12} \equiv \iint K(w,v)^2\,\mathrm{d}w\,\mathrm{d}v, \quad C_{22} \equiv K(0), \quad C_{32} \equiv \int K(w)^2\,\mathrm{d}w, \tag{3.12}$$

$$C_{11} \equiv \iint \left[\iint K(w,v) K(w+\tilde{w}, v+\tilde{v})\,\mathrm{d}w\,\mathrm{d}v \right]^2 \mathrm{d}\tilde{w}\,\mathrm{d}\tilde{v}. \tag{3.13}$$

Note that under Assumptions 2–3, γ_{j2}, $j=1,2,3$ and σ_{11}^2 are finite, and so is $\Gamma(f,f,F)$.

Our result is now stated as follows.

Theorem 1. Under Assumptions 1–3, *we have that under* H_0

$$\sigma_{11}^{-1}[Nh^{(p+q)/2}\tilde{\Gamma} - h^{-(p+q)/2}\gamma_{12} - h^{(q-p)/2}\gamma_{22} - h^{(p+q)/2}H^{-p}\gamma_{32}]$$
$$\to \mathrm{N}(0,1). \tag{3.14}$$

To implement the test, we require estimates $\hat{\sigma}_{11}^2$ of σ_{11}^2 and $\hat{\gamma}_{j2}$ of γ_{j2}, $j=1,2,3$ which we give in Section 3.2. We then compare

$$\hat{\tau} \equiv \hat{\sigma}_{11}^{-1}(Nh^{(p+q)/2} \cdot \tilde{\Gamma} - h^{-(p+q)/2}\hat{\gamma}_{12} - h^{(q-p)/2}\hat{\gamma}_{22} - h^{(p+q)/2}H^{-p}\hat{\gamma}_{32}) \tag{3.15}$$

to the critical value z_α from the $\mathrm{N}(0,1)$ distribution, i.e., $z_{0.05} = 1.64$ and $z_{0.10} = 1.28$ since the test is one-sided, and reject H_0 when $\hat{\tau} > z_\alpha$.

Our statistic and the hypotheses also make sense if $p=0$, i.e., there is no w. In that case $\hat{f}_H$ is not defined but we simply set $M(w) \equiv \mathrm{E}[Y]$ (the unconditional expectation), $\hat{M}_H(w)$ to be the sample unconditional mean, and our results continue to hold with $\gamma_{22} = \gamma_{32} = 0$ and the elimination of any conditions involving the restricted bandwidth sequence H. Of course, $\hat{\gamma}_{22}$ and $\hat{\gamma}_{32}$ are also 0.

3.2. Estimation of critical values

The quantities γ_{j2} and σ_{11}^2 depend on (3.1) and (3.2), the conditional variance of y given w and v in the compact support S of a and for that, we can

use any nonparametric estimator, for instance

$$\hat{\sigma}_h^2(w,v) = \frac{\sum_{i=1}^N K_h(w - W_i, v - V_i) Y_i^2}{\sum_{i=1}^N K_h(w - W_i, v - V_i)} - \hat{m}_h(w,v)^2 \tag{3.16}$$

for the unrestricted regression, and

$$\hat{\sigma}_H^2(w) = \frac{\sum_{i=1}^N K_H(w - W_i) Y_i^2}{\sum_{i=1}^N K_H(w - W_i)} - \hat{M}_H(w)^2 \tag{3.17}$$

for the restricted regression. With this estimator, we can define estimates of σ_{11}^2 and γ_{j2}, $j = 1,2,3$ as

$$\hat{\sigma}_{11}^2 = \frac{2C_{11}}{N} \sum_{i=1}^N \frac{\hat{\sigma}_h^4(W_i, V_i) A_i^2}{\hat{f}_h(W_i, V_i)}, \qquad \hat{\gamma}_{12} = \frac{C_{12}}{N} \sum_{i=1}^N \frac{\hat{\sigma}_h^2(W_i, V_i) A_i}{\hat{f}_h(W_i, V_i)},$$

$$\hat{\gamma}_{22} = -\frac{2C_{22}}{N} \sum_{i=1}^N \frac{\hat{\sigma}_h^2(W_i, V_i) A_i}{\hat{f}_H(W_i)}, \qquad \hat{\gamma}_{32} = \frac{C_{32}}{N} \sum_{i=1}^N \frac{\hat{\sigma}_H^2(W_i) \tilde{A}_i}{\hat{f}_H(W_i)},$$

where, in $\hat{\gamma}_{32}$,

$$\tilde{A}_i \equiv \frac{\sum_{j=1}^N K_H(W_i - W_j) A_j}{\sum_{j=1}^N K_H(W_i - W_j)}$$

estimates $a(W_i)$.[10]

We will show that $\hat{\sigma}_{11}^2$ and the respective $\hat{\gamma}_{j2}$, $j = 1,2,3$, can be substituted for σ_{11}^2 and γ_{j2} in Theorem 1 with no effect on the asymptotic distribution. Finally, the constants C_{ij} are determined by the kernel chosen as in (3.12) and (3.13) and are easily computed. For example, of the Gaussian product kernel of order $r = 2$ (density of $N(0,1)$), we have that

$$C_{12} = 1/(2\sqrt{\pi})^{p+q}, \quad C_{22} = 1/(\sqrt{2\pi})^p,$$

$$C_{32} = 1/(2\sqrt{\pi})^p, \qquad C_{11} = 1/(2\sqrt{2\pi})^{p+q}. \tag{3.18}$$

This complete the description of the test statistic $\hat{\tau}$ in (3.15).

We give a proof of Theorem 1 in the appendix in which Section A.2 gives some intuition for the result. In the next two sections, we study the consistency and asymptotic power of the test, and then discuss possible variations on the choice of the bandwidth parameters as well as other variations in the design of the test statistic.

[10] Here and in what follows, if we were to replace our weighting function a by squared-density weighting, the same expressions hold with A_i and $\tilde{A}_i$ defined according to the squared-density estimates evaluated at the data point (W_i, V_i).

3.3. Consistency and local power properties

In this section, we start by studying the consistency of the test, i.e., its ability to reject a false null hypothesis with probability 1 as $N \to \infty$. We then examine its power, i.e., the probability of rejecting a false hypothesis, against sequences of alternatives that get closer to the null as $N \to \infty$. This is given more precisely as follows.

Proposition 1. Under Assumptions 1–3, *the test based on the statistic* (3.15) *is consistent for F such that* $\Gamma(f, f, F) > 0$. *Note that this is equivalent to* $(m(w,v) - M(w))a(w,v) \neq 0$ *in a region of positive density mass.*

We now examine the power of our test against the sequence of local alternatives defined by a density $f^{[N]}(y,w,v)$ such that, if we use superscripts $[N]$ to indicate dependence on $f^{[N]}$:

$$\mathrm{H}_{1N}\colon\ \sup\{|m^{[N]}(w,v) - M^{[N]}(w) - \varepsilon_N \Delta(w,v)|\colon (w,v) \in S\} = o(\varepsilon_N) \tag{3.19}$$

where

$$||f^{[N]} - f||_\infty = o(N^{-1}h^{-(p+q)/2})$$

and the function $\Delta(w,v)$ satisfies

$$\delta_2 \equiv \iint \Delta^2(w,v) f(w,v) a(w,v)\,\mathrm{d}w\,\mathrm{d}v < \infty \tag{3.20}$$

and

$$\int \Delta(w,v) f(w,v)\,\mathrm{d}v = 0. \tag{3.21}$$

The following proposition shows that our test can distinguish alternatives H_{1N} that get closer to H_0 at rate $N^{-1/2}h^{-(p+q)/4}$ while maintaining a constant power level

Proposition 2. Under Assumptions 1–3, *suppose that the local alternative* (3.19) *converges to the null in the sense that*

$$\varepsilon_N = N^{-1/2}h^{-(p+q)/4}. \tag{3.22}$$

Then, asymptotically, the power of the test is given by

$$\Pr(\hat{\tau} \geqslant z_\alpha | \mathrm{H}_{1N}) \to 1 - \Phi(z_\alpha - \delta_2/\sigma_{11}), \tag{3.23}$$

where $\Phi(\cdot)$ *designates the* $\mathrm{N}(0,1)$ *distribution function.*

3.4. Important variations and relation to other approaches

3.4.1. Bandwidth conditions

It is important to note that we obtain our result under the bandwidth choices given by Assumption 3.2, notably (3.4). Other limiting conditions on the bandwidths will result in different terms for bias in the procedure. For instance, if we consider conditions applicable to pointwise optimal estimation, that is if

$$\delta = (p+q) + 2r \tag{3.24}$$

instead of (3.4), then $h = O(N^{-1/\delta})$ would minimize the mean integrated square error (MISE) of the regression estimate $\hat{m}(w,v)$

$$h = \arg\min \int_w \int_v \mathrm{E}[(\hat{m}_h(w,v) - m(w,v))^2]\,\mathrm{d}w\,\mathrm{d}v.$$

The rates exhibited by bandwidths chosen by cross-validation would satisfy (3.24).[11] In that case, the asymptotic distribution of $\tilde{\Gamma}$ would be driven by an additional term. That in turn would lead to an additional component of variance whose estimation would require the difficult estimation of rth order derivatives of the regression and density in order to construct a test statistic. This is the situation considered for the testing of a parametric regression by Härdle and Mammen (1993).

3.4.2. Bias correction

Our results produce a testing procedure based on a direct comparison of the nonparametric regression estimators $\hat{m}_h(w,v)$ and $\hat{M}_h(w)$. Our motivation was practical—this produces to our minds an easily interpretable statistic, and the assumed weighting procedure allows weighting to focus on particular data ranges of interest. It is possible to select the weighting function to put zero weight on certain regions of the support of the density: for example, in the case of option pricing (see our application below), it is sensible to eliminate options that are far out of the money since these are very thinly traded and their prices are less reliable. The technical cost of this orientation is in the bias features of our procedure, which lead to the bias correction terms γ_{j2}, $j = 1,2,3$ in (3.14). However, the second and third terms of the bias in our test statistic can be removed by appealing to the clever centering device of Härdle and Mammen (1993) in our setting. We would first compute the restricted regression, $\hat{M}_H(w)$, then compute the kernel regression of $\hat{M}_H(w)$

[11] Our results apply to the case of a nonstochastic bandwidth sequence h; however, we conjecture that the test is valid for data-driven bandwidths $\hat{h}$, as long as plim $\hat{h}/h = 1$ but do not formally address the issue here.

on (w,v), say $\hat{E}m_{h,H}(w,v)$, and base testing on their difference. Note that

$$\mathrm{E}[\hat{m}_h(w,v)\,|\,X_1,\ldots,X_n] = \sum_{i=1}^{N} K_h(w-W_i, v-V_i)m(W_i,V_i)/\hat{f}_h(w,v).$$

Under H_0, $m(W_i,V_i) = M(W_i)$ and we are thus led to replace $\hat{M}_H(w)$ in our statistic by

$$\hat{E}m_{h,H}(w,v) \equiv \sum_{i=1}^{N} K_h(w-W_i, v-V_i)\hat{M}_H(W_i)/\hat{f}_h(w,v)$$

to form $\tilde{\Gamma} \equiv N^{-1}\sum_{i=1}^{N}\{\hat{m}_h(W_i,V_i) - \hat{E}m_{h,H}(W_i,V_i)\}^2 A_i$.

This procedure differs from ours in that the restricted regression is replaced by a kernel-smoothed version of it. The last two of the three bias terms in Theorem 1 will become asymptotically negligible. In practical terms, this would reduce the need to do bias correction down from three to one term and could allow the setting of larger bandwidths in practice.[12] We establish the following under our assumptions.

Corollary 1. Under Assumptions 1–3, *we have that under* H_0

$$\sigma_{11}^{-1}[Nh^{(p+q)/2}\tilde{\Gamma} - h^{-(p+q)/2}\gamma_{12}] \to \mathrm{N}(0,1). \tag{3.25}$$

3.4.3. Leave-out estimators and squared-density weighting

The second variation arises from noting that the leading bias term in our statistic is due to the simultaneous use of ith observation to compute the regression functions *and* to evaluate them, e.g. the term $\hat{m}_h(W_i,V_i)$ uses (Y_i,W_i,V_i) in its construction. This leading bias term can be avoided by using the well-known "leave-out i" regression estimators. In particular, Fan and Li (1996) use that device (which is not essential to their approach) as well as a specific weighting function $a(W,V)$ (which plays a critical role) in their analysis. In our framework, their statistic would be based on estimating $\Gamma(f,f,F)$ by

$$\mathrm{E}[(Y-M(W))\mathrm{E}[(Y-M(W))\,|\,W,V]a(W,V)] \tag{3.26}$$

but with the very specific choice that $a(W,V) = f^2(W)f(W,V)$. Indeed, note that in that case (3.26) becomes

$$\mathrm{E}[(Y-M(W))f(W)\mathrm{E}[(Y-M(W))f(W)|W,V]f(W,V)]$$

[12] As in Härdle and Mammen (1993)'s simpler case, we can also extend (3.25) for a bandwidth h such that $Nh^{(p+q)/2+2r}$ does not tend to zero, but at the price of additional bias terms that would have to be estimated, and in practice can be difficult to estimate because they involve derivatives of the regression function.

to be estimated quite naturally by

$$\frac{1}{N^2}\sum_{i=1}^{N}(Y_i-\hat{M}_H(W_i))\hat{f}_H(W_i)$$

$$\left\{\sum_{j=1}^{N}K_h(W_i-W_j,V_i-V_j)(Y_j-\hat{M}_H(W_j))\hat{f}_H(W_j)\right\},$$

which no longer contains a denominator. In other words, the specific choice of weighting function has the effect of transforming a rational function into a polynomial. Fan and Li (1996) further simplify the asymptotic distribution calculations by replacing $\hat{f}_H(W_i)$ by the leave-out estimator $\hat{f}_{H,(-i)}(W_i)$, and $\hat{M}_H(W_i)$ by $\hat{M}_{H,(-i)}(W_i)$ where $(-i)$ indicates that all observations from 1 to N except the ith have been used to form the estimator. In practice, this variation is more complicated to apply, since the leave-out estimator requires that a separate regression computation needs to be done for each observation, or summand in the test statistic. On the other hand, it results in a simpler asymptotic distribution—not requiring the functional expansion we must use here—but only for the specific choice of weighting function noted above. As noted earlier, Lavergne and Vuong (2000) give a test statistic that uses squared-density weighting but avoids using "leave-out" estimators, and has significantly better bias properties than the Fan and Li (1996) proposal.

3.4.4. Relation to the bootstrap

One can develop suitable versions of the bootstrap or other resampling methods to avoid the bias estimation problem, while retaining the computational simplicity of our basic test statistic. Härdle and Mammen (1993) show the validity of a version of the bootstrap (which they refer to as the "wild bootstrap") to obtain critical values for their statistic. In fact, it is easy to see from Bickel et al. (1997) that, in both our situation and theirs, the following bootstrap will produce correct critical values: we resample m out of the N observations, (W_i^*, V_i^*, Y_i^*), $1 \leqslant i \leqslant m$, compute by Monte-Carlo the distribution of $mh^{*(p+q)/2}\tilde{\Gamma}_m^*$ where

$$\tilde{\Gamma}_m^* \equiv \frac{1}{m}\sum_{i=1}^{m}\{\hat{m}_{h^*}^*(W_i^*,V_i^*)-\hat{M}_{H^*}^*(W_i^*)\}^2 A_i^*$$

and the bandwidths h^* and H^* are selected as in Assumption 2 but for a sample size of size m, with $m \to \infty$ such that $m/N \to 0$.[13] The advantage of

[13] Data-dependent rules for selecting m are discussed in Sakov (1998). Another possibility is discussed in Bickel et al. (1998).

the bootstrap, other than providing a potentially more accurate critical value for the test statistic, is to obviate the need for estimating the bias terms γ_{j2} and the standard deviation σ_{11} as in Section 3.2.

4. Some useful corollaries

Theorem 1 covers the distribution of the test statistic when the null hypothesis involves omitting the variables V from the regression of Y on (W, V), where the observed data are a random sample. While this case will suffice for many empirical issues, our testing procedure is potentially applicable to a much wider range of situations. We now discuss several corollaries that generalize the basic result above.

4.1. Index models

Our test above applies directly as a test of dimension reduction—it checks whether a smaller number of variables suffices for the regression. Our first set of corollaries indicates how our test is applicable to other methods of dimension reduction. In particular, one might wish to create a smaller set of variables by combining predictor variables in a certain, interpretable way, and then test that the regression depends only on the combination. If the method of combining variables is known exactly (e.g., take their sum), then the above results apply immediately. However, if the method of combining variables involves parameters that are estimated, then we must check how the earlier results would change. We argue heuristically that they apply with only minor additional smoothness and bandwidth assumptions.

A principal example of this kind of structure arises when a weighted average of the predictors is used in the regression, where the weights must be estimated. Here $W = X'\theta$, and we rewrite the predictor vector X as its (invertible) transformation $(W, V) = (X'\theta, V)$. A single index model of the regression is then $m(x) = M^*(x'\theta)$. If the unknown function M^* were assumed to be invertible, this is the standard form of a generalized linear model. We note in passing that we could summarize the impacts of a subset of variables via an index, leaving some others to have unrestricted effects—a partial index model would take $W = (X'\theta, W_{-1})$, where again the predictor X is an invertible transformation of $(X'\theta, W_{-1}, V)$. In these examples, if θ is known, then our previous results apply for testing the null hypothesis that an index model is valid. When θ is not known, it must be estimated, and an estimator $\hat{\theta}$ can often be found that is $\sqrt{N}$ consistent for its limit. For index models with continuous regressors, such estimators are given by average derivative estimators, among many others. For generalized linear models, maximum rank

correlation or penalized maximum likelihood methods give $\sqrt{N}$ consistent estimators of θ.[14]

We consider a more general framework, whereby the vector W is allowed to depend generally on a finite vector θ of parameters as $W \equiv w(X,\theta)$ and the restricted (null) regression model is

$$\mathrm{H}_0: \ \Pr[m(W,V) = M(w(X,\theta))] = 1, \tag{4.1}$$

where again, M is unknown but w is known up to the parameter vector θ. That is, there exists a differentiable and invertible map $w: \ X \mapsto (w(X,\theta), v(X,\theta))$, for each θ, where w takes values in R^p, v in R^q, $q > 0$, which satisfies the following.

Assumption 4. The map w and its Jacobian $J(x,\theta)$ are continuous as functions of x and θ. Further $J \neq 0$ for all $x \in S$, $\theta \in \Theta$.

Of interest is the application of our test statistic where an estimate $\hat{\theta}$ of θ is used, or where $\hat{W} \equiv w(X,\hat{\theta})$ is used in place of the true W in the test statistic. Our discussion above pointed out how the relevant variation for our test statistic is determined by the dimensionality of the alternative hypothesis. Consequently, it is natural to conjecture that the use of a $\sqrt{N}$ consistent estimator $\hat{\theta}$ will not change the limiting distribution at all, so that we can ignore the fact that $\hat{\theta}$ is estimated for the purposes of specification testing.

Assumption 5. The estimate $\hat{\theta}$ is $\sqrt{N}$-consistent, that is for all θ in a compact parameter space Θ, $\hat{\theta} - \theta = O_{\mathrm{p}_\theta}(N^{-1/2})$.

We then obtain the following.

Corollary 2. Under Assumptions 4, 5 *and* $\delta > 13(p+q)$, *the conclusion of Theorem* 1 *can be applied to*

$$\tilde{\Gamma} \equiv \frac{1}{N}\sum_{i=1}^{N}\{\hat{m}_h(w(X_i,\hat{\theta}), v(X_i,\hat{\theta})) - \hat{M}_H(w(X_i,\hat{\theta}))\}^2 A_i. \tag{4.2}$$

4.2. Parametric, semiparametric and other models

In the above section, we have proposed a variation to the basic testing result, namely permitting the use of estimated parameters in the restricted set of regressors. Much of the previous work has focused on testing a specific parametric model against flexible nonparametric alternatives. Our results are directly relevant to this setting, by noting an obvious but quite important

[14] Stoker (1992) discusses these and other methods.

feature of our test. The rate of convergence and asymptotic variance of our test statistic depends only on the dimensionality of the alternative hypothesis, and there is no reason why we cannot restrict attention to null hypotheses that are parametric models. This adds the test statistic to the toolbox of diagnostic methods for parametric modeling. While failure to reject is rather weak evidence for a parametric hypothesis, the test can detect significant departures in unexpected directions. More specifically, consider the case of a parametric model as null hypothesis, with

$$\mathrm{H}_0: \ \Pr[m(W,V) = M(W;\theta)] = 1, \tag{4.3}$$

where the function $M(\cdot;\theta)$ is known, but the parameter vector θ is unknown. An estimator $\hat{\theta}$ of θ satisfying Assumption 5 can be obtained under smoothness assumptions from nonlinear least squares estimation of (4.3) or like methods. Assume the following.

Assumption 6. $M(\cdot;\theta)$ is differentiable in θ, with derivative uniformly bounded for $w \in S$, and θ in a neighborhood of the true parameter value in Θ.

With regard to single index models, we could, for instance, test the null hypothesis that the regression is a linear model in the predictor variables, or that $\mathrm{E}[Y|W=w, V=v] = w'\theta$. This is almost a specialization of the results of the previous section if we consider the case $p=0$ for which $M_\theta(w) \equiv 0$ and then note that replacing 0 by $M(\cdot;\hat{\theta})$ has a lower order effect.

We shall establish as a special case a result analogous to that obtained by Härdle and Mammen (1993):

Corollary 3. Under the additional Assumptions 5-6, *Theorem* 1 *can be applied, with* γ_{22} *and* γ_{32} *replaced by* 0, *to*

$$\tilde{\Gamma} = \frac{1}{N}\sum_{i=1}^{N} \{\hat{m}_h(W_i,V_i) - M(W_i,\hat{\theta})\}^2 A_i,$$

under the null (4.3).

Note that the validity of Theorem 1 implies that the variance of $\hat{\theta}$ does not affect the limiting distribution of $\tilde{\Gamma}$. The logic applies when the restricted model involves lower dimensional estimated functions as well as estimated parameters. For instance, consider the semiparametric model

$$\mathrm{H}_0: \ \Pr[m(W,V) = M(W) + M(V;\theta)] = 1, \tag{4.4}$$

where the function $M(\cdot;\theta)$ is known, but $M(\cdot)$ and the finite-dimensional parameter vector θ are unknown (see Robinson, 1988 for an estimation strategy

for this model with $M(v;\theta) \equiv v'\theta$). Specifically, the model can be written as

$$\mathrm{E}[Y - M(V;\theta) \mid W = w, V = v] = \mathrm{E}[Y - M(V;\theta) \mid W = w].$$

Thus the appropriate test statistic given an estimate $\hat{\theta}$ satisfying Assumption 5 is

$$\tilde{I}^* \equiv \frac{1}{N}\sum_{i=1}^{N} \{\hat{m}_h^*(W_i, V_i, \hat{\theta}) - \hat{M}_H^*(W_i)\}^2 A_i, \tag{4.5}$$

where $\hat{m}_h^*, \hat{M}_H^*$ are $\hat{m}_h, \hat{M}_H$ applied to the observations $(Y_i - M(V_i;\hat{\theta}), W_i, V_i)$. As in Corollary 2, we obtain the following.

Corollary 4. Under the additional Assumptions 5-6*, Theorem* 1 *can be applied without modification to* $\tilde{I}^*$ *under the null* (4.4).

Again, the estimation of $\hat{\theta}$ should not affect the limiting distribution of $\tilde{I}$, while, under our bandwidth choices, the nonparametric estimation of $M(\cdot)$ gives rise to the bias adjustment terms γ_{22} and γ_{32}. As before, when the restricted model depends on estimated functions that converge at rates faster than the general model, the distribution of the test statistic should be determined solely by the general model given by the alternative hypothesis.

4.3. Extensions to more general data types

4.3.1. Limited dependent variables

We have made reference to the joint density $f(y,w,v)$ to facilitate the functional expansion in a natural way. However, there is no explicit use of the continuity of the dependent variable Y in the derivations. In particular, the joint density $f(y,w,v)$ can be replaced everywhere by $f(w,v)\,\mathrm{d}F(y|w,v)$ without changing any of the derivations. This is more than a superficial change, as it allows the application of our test statistic to any situation involving limited dependent variables. For instance, Y may be a discrete response, with the regression a model of the probability that Y takes on a given value. Alternatively, Y could be a more complicated censored or truncated version of a continuous (latent) variable.

4.3.2. Dependent data

We have regarded the observed data above as a random sample, which is appropriate for analysis of survey data or other kinds of data based on unrelated observation units. However, for many settings, the ordering or other kind of connections between observations must be taken into account. Examples include the analysis of macroeconomic or financial time series data.

For testing for regression structure in this context, what complications would dependent data raise for our results?

The moment calculations we have used in our derivation continue to hold for (Y_i, W_i, V_i) stationary ergodic and in a mixing context remainders are still of smaller order. The result follows from our observation that the distribution of the test statistic is characterized by the behavior of the general regression model. In particular, the covariations induced by the dependence in the data are of higher order, and therefore have no impact, as in the standard result on the distribution of the kernel regression estimator itself (see Robinson, 1989). Specifically, Theorem 1 holds provided that the amount of serial dependence in the data decays sufficiently fast.

Assumption 7. With $Z_i \equiv (Y_i, W_i, V_i)$:

1. The data $\{Z_i; i = 1, \ldots, N\}$ are strictly stationary and β-mixing with $\beta_N = O(N^{-\kappa}), \kappa > 19/2$.
2. The joint density $f_{1,j}(\cdot,\cdot)$ of (Z_1, Z_{1+j}) exists for all j and is continuous on $(R \times S)^2$.

We then obtain the following.

Proposition 3. Theorem 1, *Propositions* 1 *and* 2, *and Corollaries* 2–4 *are unmodified if Assumption* 1 *is replaced by Assumption* 7.

5. Finite sample properties: A Monte Carlo study

To give a brief description of the finite sample performance of the test statistic, we present simulation results for one and two-dimensional testing situations. We begin with a one-dimensional study of functional form, where the true model is $\mathrm{E}[Y|W = w] = w\theta$ with $\theta = 1$, for w distributed as $\mathrm{N}(0, 1)$. In particular, we constructed samples with

$$Y = W\theta + \sigma(W)\varepsilon, \tag{5.1}$$

where ε is distributed as $\mathrm{N}(0, 1)$ and $\sigma^2(w) = 0.5625 \exp(-w^2)$. With reference to Corollary 3, we have that $p = 1$, $q = 0$. For the general regression we use a univariate normal kernel function, and compute the bandwidth as $h_0 N^{-1/\delta}$ with $\delta = 4.25$, and h_0 is set to 0.50, 0.60 and 0.75, and a is the indicator function of the interval $S = \{w \in R / -1.5 \leqslant w \leqslant 1.5\}$. The restricted model is estimated by ordinary least squares (OLS). We simulated 1000 samples for each case. Table 1 reports the observed rejection rates for 5% and 10% critical values ($z_{0.05} = 1.64$ and $z_{0.10} = 1.28$), and the standard deviation of the standardized test statistic $\hat{\tau}$ (which is 1 asymptotically).

Table 1
Monte Carlo results: one dimension

		$h_0 = 0.50$	$h_0 = 0.60$	$h_0 = 0.75$
$N = 500$	5%	5.7	5.6	6.0
	10%	8.9	8.4	9.4
	Stan. Dev. ($\hat{\tau}$)	0.94	0.94	0.93
$N = 1000$	5%	5.6	5.3	6.0
	10%	8.9	8.6	9.6
	Stan. Dev. ($\hat{\tau}$)	0.98	0.95	0.94
$N = 5000$	5%	5.5	5.7	6.6
	10%	9.0	8.9	10.5
	Stan. Dev. ($\hat{\tau}$)	0.98	0.97	0.96
$N = 10,000$	5%	5.2	5.7	7.0
	10%	8.6	9.2	11.0
	Stan. Dev. ($\hat{\tau}$)	0.96	0.97	0.99
$N = 15,000$	5%	5.0	5.6	6.7
	10%	8.7	9.3	11.0
	Stan. Dev. ($\hat{\tau}$)	0.96	0.98	0.98

Table 2
Monte Carlo results: two dimensions

		NPGEN-NPREST	NPGEN-PARAM
$N = 500$	5%	7.4	5.2
	10%	13.6	8.6
	S.D. ($\hat{\tau}$)	0.89	0.93
$N = 1000$	5%	7.2	5.1
	10%	12.1	9.1
	S.D. ($\hat{\tau}$)	0.90	0.96
$N = 5000$	5%	6.2	4.7
	10%	11.8	8.6
	S.D. ($\hat{\tau}$)	0.95	0.97

Table 1 shows a reasonable correspondence between the finite sample performance of the test statistic and the asymptotic results for the one-dimensional design. There is some tendency of the test statistic to over-reject, and that tendency arises with larger bandwidth values. But in any case, the results are close to the expected values.

To study the performance of the test statistic in two dimensions, we generate samples using the same model (5.1) as above, and test for the presence of an additional regressor V, which is distributed as $N(0,1)$, independently of W and ε. We study the performance of the test statistic in two settings: first a comparison of nonparametric estimates of the general nonparametric regression $E[Y|W=w, V=v]$ and the restricted nonparametric regression $E[Y|W=w]$ ("NPGEN-NPREST" in Table 2) and then a comparison of nonparametric estimates of $E[Y|W=w, V=v]$ to the OLS fitted values of

regressing parametrically Y on W ("NPGEN-PARAM" in Table 2). With reference to Theorem 1, we have $p=1$ and $q=1$. We use standard normal (one- and two-dimensional) kernel functions, and set bandwidths as $h=h_0N^{-1/\delta}$, $\mathrm{H}=\mathrm{H}_0N^{-1/\Delta}$, where $\delta=4.75, \Delta=4.25, h_0=0.75, \mathrm{H}_0=0.79$. The weighting function a is the indicator function of the disk $S=\{(w,v)\in R^2/\sqrt{w^2+v^2}\leqslant 1.5\}$ in R^2. We simulated 500 samples for each case. Table 2 reports the observed rejection rates for 5% and 10% critical values, and the standard deviation of the standardized test statistic $\hat{\tau}$.

In Table 2 we see that the tendency of the test statistic to over-reject is more pronounced. When comparing general and restricted kernel estimates, the observed standard deviation of the test statistic is less than one, which is associated with the tendency to over-reject, but the same spread compression is not very evident in the test statistics comparing the general nonparametric regression to OLS estimates of the parametric regression. In any case, we conclude that there is a general correspondence between the finite sample performance of the test statistic and the theoretical results, with some tendency toward the test statistic over-rejecting, or producing confidence regions that are too small.

6. An empirical application: modeling the deviations from the Black–Scholes formula

To illustrate our testing procedure, we present an empirical analysis of options prices. In particular, we study the goodness-of-fit of the classical Black and Scholes (1973) option pricing formula and its extensions. While we reject the standard parametric version of this formula against a general nonparametric alternative, we fail to reject a version of the formula that employs a semiparametric specification of the implied volatility surface. Our data sample consists of $N=2910$ observations on daily S&P 500 index options traded at the Chicago Board Options Exchange during the year 1993. The S&P 500 index option market is extremely liquid (as one of the most active among option markets in the United States), the options are European, and within this market we have chosen the most actively traded options with maturities ranging from 3 to 9 months. For further information on the basic data set, see Aït-Sahalia and Lo (1998).

We denote the call option price as C, its strike price as X, its time-to-expiration as T, the S&P 500 futures price for delivery at expiration as F, the risk-free interest rate between the current date and T as r.[15] The Black–

[15] In our sample, the risk-free rate of interest is constant at $r=3.05\%$ (short term interest rates were quite stable during 1993).

Scholes option pricing formula is

$$C = \mathrm{e}^{-rT} F \left[\Phi(d_1) - \frac{X}{F} \Phi(d_2) \right], \tag{6.1}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function, and

$$d_1 = \frac{\ln(F/X) + (s^2/2)T}{s\sqrt{T}}, \quad d_2 = d_1 - s\sqrt{T}$$

for a value of the volatility parameter s constant across different moneyness values (defined as X/F) and time-to-expiration T. The put option price is given by

$$P = C + (X - F)\mathrm{e}^{-rT}.$$

The industry's standard convention is to quote option prices in terms of their implied volatilities, so an option "trades at $s = 16\%$" rather than "for $C = \$3.125$". In other words, for each option price in the database, with characteristics $(X/F, T)$, (6.1) can be inverted to produce the option's implied volatility. This is the unique value of s, as a function of $(X/F, T)$, that would make $C(X/F, T, s)$ on the right-hand side of (6.1) equal to the observed market price of the option. Using (6.1) to compute s just represents an invertible transformation of the price data; a nonlinear change in scale.

Of course, market prices may not satisfy the Black–Scholes formula, in which case the implied volatility s of options with different moneyness X/F and time-to-expiration T will not be identical across different options, but would depend on X/F and T. Moreover, there are a number of possible sources of noise in the market data: option data might not match perfectly with the market price F (S&P 500 futures are traded at the Chicago Mercantile Exchange), both F and C are subject to a bid-ask spread, the interest rate r used to calculate s may not correspond to the market's perception of the cost of riskless borrowing and lending at that point in time, etc. We summarize these potential sources of noise as an additive residual in implied volatilities. Namely, we pose the model [16]

$$s = m(X/F, T) + \varepsilon \quad \text{with } \mathrm{E}[\varepsilon | X/F, T] = 0. \tag{6.2}$$

[16] Measurement errors in the regressors, notably through F, can potentially lead to biased estimation of the model's parameters. Hence, as is the standard practice in finance, our modeling choice is to summarize all measurement errors as an additive residual ε in implied volatilities and assume that all variables are measured without error. In addition, the term ε does create the appearance of arbitrage opportunities in the system of options. These opportunities are however spurious since differences in s from the values $m(X/F, T)$ represent measurement error on the part of the econometrician and cannot be traded upon. Finally, because of the nonlinear nature of the transformation between implied volatilities and prices, an additive measurement-error noise in implied volatilities is no longer an additive noise in prices—except as a first order Taylor approximation.

If the Black–Scholes formula were a correct depiction of how an ideal market operates, then $m(X/F, T)$ would be a constant independent of X/F and T. It is now recognized in the literature that this is not the case, especially since the October 1987 market crash. Regression patterns, known as "volatility smiles", have been identified in the data, whereby it is typical for out-of-the-money puts, i.e., put options with moneyness $X/F < 1$, to trade at higher implied volatilities than at-the-money options and out-of-the-money calls ($X/F > 1$).[17] The pattern is nonlinear: as a function of X/F, the implied volatility s tends to decrease below $X/F = 1$ (at-the-money), and then flattens out above 1. The level of s also generally decreases as a function of time-to-expiration T, although this effect is not as salient in the data.

Our objective is to determine a parsimonious model for $\mathrm{E}[s|X/F, T] = m(X/F, T)$. In particular, we are interested in learning whether the full generality of the two-dimensional nonparametric regression function $m(X/F, T)$ is needed to adequately model implied volatilities. For this, we consider five versions of the model: BS refers to the standard Black–Scholes specification of constant volatility, PARAM refers to a quadratic model in X/F as is common in the modeling of a "volatility smile", NPREST refers to the restricted (one-dimensional) nonparametric specification of volatility as a function of X/F, SEMIPAR refers to the partially linear specification with a nonparametric additive structure for X/F and a quadratic structure for time T, and NPGEN refers to the general unrestricted two-dimensional regression. The five models are given in Table 3.

Estimation for the various specifications is as follows: the parametric models (BS and PARAM) are estimated by OLS regression, the nonparametric models (NPREST and NPGEN) are estimated by setting smoothing parameters as in the Monte Carlo analysis,[18] and SEMIPAR is estimated by using Robinson's (1988) difference method (with the same smoothing parameters as for NPREST). We summarize the results for fitting these models in Table 4. To illustrate the nonparametric component of the SEMIPAR, we also include in Fig. 1 a graph of the estimated function $\hat{g}(X/F)$ (where X/F values are in standardized form). Each dot represents an option, and all the maturities are included.

[17] Various arguments have been proposed to explain this phenomenon: for example, puts are more expensive, on a volatility basis, because of the excess demand for protective puts—an option strategy which would cap the losses of a stock portfolio in the event of a market downturn. See Aït-Sahalia and Lo (1998, 2000) for a discussion.

[18] In particular, normal kernels are used, and smoothing is done after the data is standardized (centered by removing the mean and divided by standard deviation). In this application, $w = X/F$ (so $p = 1$) and $v = t$ (so $q = 1$). For NPGEN, we set $h = h_0 N^{-1/\delta}$ for $h_0 = 0.65$, $\delta = 4.75$, and for NPREST, we set $H = \mathrm{H}_0 N^{-1/\Delta}$ for $\mathrm{H}_0 = 0.50$, $\Delta = 4.25$. The weighting function a is the indicator function of the set $S = [-1.5, 1.5] \times [-0.75, 0.75]$ in R^2.

Table 3
Models to be estimated and tested[a]

BS:	$m(X/F,T)=\theta_0$
PARAM:	$m(X/F,T)=\theta_1+\theta_2 X/F+\theta_3(X/F)^2$
NPREST:	$m(X/F,T)=g_1(X/F)$
SEMIPAR:	$m(X/F,T)=g(X/F)+\theta_4 T+\theta_5 T^2$
NPGEN:	$m(X/F,T)$ unrestricted

[a]The parametric models BS and PARAM are estimated by ordinary least squares. The two nonparametric models NPREST and NPGEN are estimated by one- and two-dimensional kernel regressions, respectively. Finally, the semiparametric model SEMIPAR implies that $s-\mathrm{E}[s|X/F]=\theta_4(T-\mathrm{E}[T|X/F])+\theta_5(T^2-\mathrm{E}[T^2|X/F])+\eta$ with $\mathrm{E}[\eta|X/F,T]=0$, so we estimate first the univariate regressions $m_s\equiv\mathrm{E}[s|X/F]$, $m_T=\mathrm{E}[T|X/F]$ and $m_{T^2}=\mathrm{E}[T^2|X/F]$ by kernel regression, estimate next θ_4 and θ_5 by regressing $s_i-\hat{m}_s(X_i/F_i)$ on $T_i-\hat{m}_T(X_i/F_i)$ and $T_i^2-\hat{m}_{T^2}(X_i/F_i)$ and finally form $\hat{g}=\hat{m}_s-\hat{\theta}_4\hat{m}_T-\hat{\theta}_5\hat{m}_{T^2}$. The fitted values from the semiparametric model are $\hat{s}=\hat{g}+\hat{\theta}_4 T+\hat{\theta}_5 T^2$, the fitted residuals $\hat{\varepsilon}=s-\hat{s}$, and, assuming that the residuals are conditionally homoskedastic, the asymptotic variance covariance matrix of the parameter estimators $(\hat{\theta}_4,\hat{\theta}_5)$ is $N^{-1}(\hat{\varepsilon}'\hat{\varepsilon})[Z'Z]^{-1}$ where Z is the matrix of parametric regressors $(T-\hat{m}_T, T^2-\hat{m}_{T^2})$.

Table 4
Estimated models for the implied volatility smile[a]

BS:	$\hat{m}(X/F,T)=$	9.09		
		(447)		
PARAM:	$\hat{m}(X/F,T)=$	8.94	$-0.89\cdot X/F$	$+0.15\cdot(X/F)^2$
		(508)	(−75)	(12)
SEMIPAR:	$\hat{m}(X/F,T)=$	$\hat{g}(X/F)$	$+0.17\cdot T$	$+0.001\cdot T^2$
			(9)	(0.1)

[a]t-statistics are in parentheses. We treat the full year of data as being drawn from the same distribution, so successive options in the sample are not temporally correlated (otherwise two successive options with the same strike price X would have days-to-maturity of T and $T-1$, respectively). For the purpose of calculating these standard errors, we treat both regressors X/F and T as random variables (a natural alternative would be to treat the variable T as a fixed design).

We then carry out the testing by computing $\hat{\Gamma}$ for each comparison, and then carry out the bias correction and scaling to compute $\hat{\tau}$ for each comparison. The results are given in Table 5. The bottom line is that all specializations of the model are rejected except for SEMIPAR. The low p-values for the restricted versions of the model (aside from SEMIPAR) could reflect the tendency of the test statistic to over-reject as noted in Section 5. The fact that the test statistics for NPREST and PARAM are close suggests that the main improvement in SEMIPAR comes from the inclusion of the time-to-maturity explanatory variable, as opposed to the nonparametric nature of the function $g(X/F)$. In summary, a semiparametric model permitting a flexible "volatility smile" as well as an additive time effect is a statistically adequate depiction of the implied volatility data.

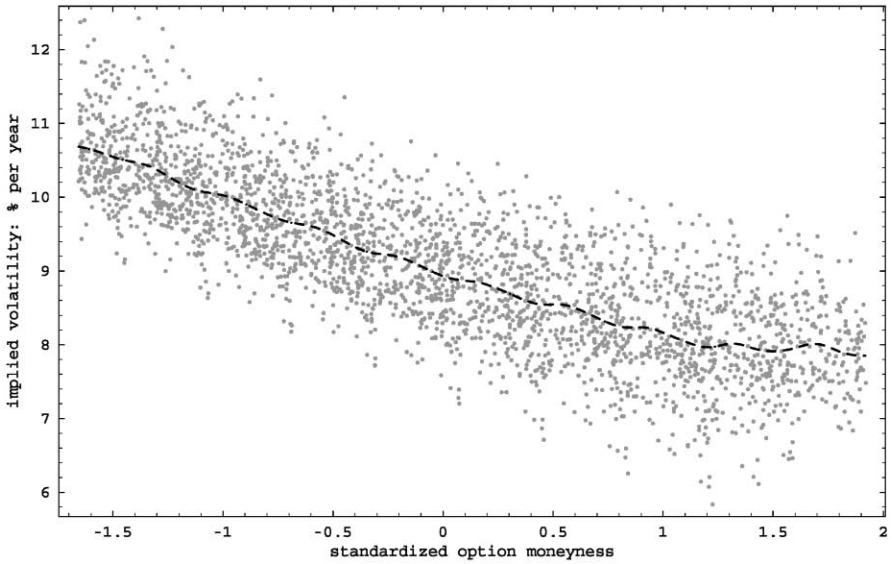

Fig. 1. Implied volatility curve for the semiparametric model: $g(X/F)$.

Table 5
Test statistics for the different implied volatility models[a]

	$\hat{\tau}$	p-value
NPGEN-BS	514.4	0.00
NPGEN-PARAM	7.2	0.00
NPGEN-NPREST	6.6	0.00
NPGEN-SEMIPAR	1.1	0.12

[a]The critical values appropriate for $\hat{\tau}$ are 1.64 (5%) and 1.28 (10%). The test NPGEN-NPREST is a basic application of Theorem 1. The tests NPGEN-BS and NPGEN-PARAM are based on Corollary 3. The test NPGEN-SEMIPAR is based on Corollary 4.

7. Conclusions

In this paper, we have developed a general specification test for parametric, semiparametric and nonparametric regression models against alternative models of higher dimension. The key contribution in our work involves the analysis of the variation of sum-of-squared residuals, in noting how the asymptotic distribution depends primarily on the generality of the alternative models permitted. As such, the test we propose is applicable to virtually any situation where the model under the null hypothesis is of lower dimension than the possible alternatives. However, the heuristics we propose need to be checked more carefully.

Our results are restricted to the use of standard kernel estimators, which include procedures that use standard positive kernels as well as higher order kernels for bias reduction. The asymptotic distribution we have derived does depend on aspects intrinsic to the kernel estimators (for instance, the constants C_{ij}), and there is no obvious reason why a similar asymptotic distribution would be applicable when other nonparametric estimators are used, such as truncated series expansions. In particular, as in Bickel and Rosenblatt (1973), differences may arise between tests based on different nonparametric estimators. The characterization of such differences, as well as questions involving the choice of the best nonparametric techniques for model testing, provide a rich field of issues for future research.

Moreover, the practical properties of our test as well as related tests need to be understood. In particular, given the richness of possible nonparametric alternatives, one might conjecture that such tests will have limited power relative to tests based on (fortunately chosen) parametric alternatives. Alternatively, while we have derived the asymptotic distribution based on the leading second order terms of the asymptotic expansion, the fact that the first order terms are nonzero under any fixed alternative suggests further study of the practical performance of test statistics of this kind.

Acknowledgements

This paper was first circulated in 1994 and was presented at the Econometric Society meetings in January 1995. The authors wish to acknowledge the numerous helpful comments they received from colleagues, seminar participants and especially the Editor and referees. Aït-Sahalia thanks the NSF (grant SBR-9996023) and an Alfred P. Sloan Research Fellowship for support.

Appendix A.

A.1. Steps of the proof

We begin by studying the asymptotic properties of the functional Γ evaluated at $(\hat{f}_h, \hat{f}_H, F)$, using the functional delta method. The only difference between $\Gamma(\hat{f}_h, \hat{f}_H, F)$ and $\tilde{\Gamma}$ of (2.5) is that the latter is an average over the empirical cumulative distribution function $\hat{F}$ instead of F . We will show in Lemma 7 that this difference is inconsequential for the asymptotic distribution of the test statistic, and hence start by studying $\Gamma(\cdot, \cdot, F)$. To bound the remainder term in the functional expansion of $\Gamma(\cdot, \cdot, F)$, we define

the seminorms [19]

$$||g_1|| \equiv \max\left(\sup_{(w,v)\in S} \left| \int y g_1(y,w,v)\,\mathrm{d}y \right|, \sup_{(w,v)\in S} |g_1(w,v)| \right),$$

$$||g_2|| \equiv \max\left(\sup_{w\in PS} \left| \int y g_2(y,w)\,\mathrm{d}y \right|, \sup_{w\in PS} |g_2(w)| \right). \tag{A.1}$$

The main ingredient in the proof of the theorem is the functional expansion of Γ, summarized as follows.

Lemma 1. Let F_3 be a c.d.f. on R^{p+q}. Let Ω_1 (resp. Ω_2) be the set of all bounded functions g_1 (resp. g_2) such that $||g_1|| < b/2$ (resp. $||g_2|| < b/2$) and $\iint \int g_1(y,w,v)\,\mathrm{d}y\,\mathrm{d}w\,\mathrm{d}v = 0$ (resp. $\iint g_2(y,w)\,\mathrm{d}y\,\mathrm{d}w = 0$). Then, under Assumption 2 and H_0, $\Gamma(\cdot,\cdot,F_3)$ has an expansion on $\Omega_1 \times \Omega_2$ about (f,f) given by,

$$\begin{aligned}
&\Gamma(f+g_1, f+g_2, F_3) \\
&\quad = \Gamma(f,f,F_3) + \iint \left[\int (\alpha(y,w,v) g_1(y,w,v) \right. \\
&\quad \left. - \beta(y,w) g_2(y,w))\,\mathrm{d}y \right]^2 a(w,v)\,\mathrm{d}F_3(w,v) \\
&\quad + R(g_1,g_2,F_3)
\end{aligned}$$

where

$$\sup\{|R(g_1,g_2,F_3)|/(||g_1||^3 + ||g_2||^3) : (g_1,g_2) \in \Omega_1 \times \Omega_2\} < \infty.$$

Consequently, to apply the functional expansion to our test statistic, we need to be able to bound the remainder term R, i.e., we need to characterize the nonparametric approximating properties for the kernel estimators $\hat{f}_h$ and $\hat{f}_H$:

Lemma 2. Under Assumptions 1–3, *we have*

$$||\hat{f}_h - f|| = O_{\mathrm{p}}(h^r + N^{-1/2} h^{-(p+q)/2} \ln(N)).$$

$$||\hat{f}_H - f|| = O_{\mathrm{p}}(H^r + N^{-1/2} H^{-p/2} \ln(N)).$$

We will choose the bandwidth sequence h in such a way that

$$Nh^{(p+q)/2} ||\hat{f}_h - f||^3 = O_{\mathrm{p}}(Nh^{(p+q)/2}[h^{3r} + N^{-3/2} h^{-3(p+q)/2} \ln^3(N)]) = o_{\mathrm{p}}(1). \tag{A.2}$$

[19] In the case of squared-density weighting, the relevant norm to use would be the L_p norm with that weighting. The expansion that follows is otherwise unmodified; Lemma 2 is also modified to account for this different norm.

and also $Nh^{(p+q)/2}||\hat{f}_H - f||^3 = o_p(1)$. This is ensured by the bandwidth choices given in Assumptions 3-2, as they clearly imply $\delta < 3r + (p+q)/2$ and $\delta > 2(p+q)$.

By applying Lemma 1 with $g_1(y,w,v) = \hat{f}_h(y,w,v) - f(y,w,v)$ and $g_2(y,w) = \hat{f}_H(y,w) - f(y,w)$, and using Lemma 2, we obtain the following

Lemma 3. Under Assumptions 1–3, *we have for any c.d.f.* F_3,

$$\Gamma(\hat{f}_h, \hat{f}_H, F_3) \tag{A.3}$$

$$= \Gamma(f, f, F_3) + \iint \left[\int \alpha(y,w,v) \hat{f}_h(y,w,v)\, dy - \int \beta(y,w) \hat{f}_H(y,w)\, dy \right]^2 a(w,v)\, dF_3(w,v)$$

$$+ O_p(||\hat{f}_h - f||^3 + ||\hat{f}_H - f||^3). \tag{A.3}$$

We then show that the difference between $\tilde{\Gamma} = \Gamma(\hat{f}_h, \hat{f}_H, \hat{F})$ *and* $\Gamma(\hat{f}_h, \hat{f}_H, F)$ *is negligible. That is, we shall show that*

$$\tilde{\Gamma} = \Gamma(\hat{f}_h, \hat{f}_H, F) + \Delta_N + \Lambda_N, \tag{A.4}$$

where from Lemma 2

$$\Lambda_N = O_p(||\hat{f}_h - f||^3 + ||\hat{f}_H - f||^3) = o_p(N^{-1} h^{-(p+q)/2}) \tag{A.5}$$

and

$$\Delta_N \equiv \iint \left\{ \int \alpha(y,w,v) \hat{f}_h(y,w,v)\, dy - \int \beta(y,w) \hat{f}_H(y,w)\, dy \right\}^2 \times a(w,v) d(\hat{F}(w,v) - F(w,v)), \tag{A.6}$$

which we will bound later in Lemma 7.

The next step is to study

$$\Gamma(\hat{f}_h, \hat{f}_H, F) = I_N + o_p(N^{-1} h^{-(p+q)/2}) \tag{A.7}$$

where

$$I_N \equiv \iint \left\{ \int \alpha(y,w,v) \hat{f}_h(y,w,v)\, dy - \int \beta(y,w) \hat{f}_H(y,w)\, dy \right\}^2 \times a(w,v)\, dF(w,v). \tag{A.8}$$

Define, for $X \equiv (W, V)$,

$$r_N(Y_1, X_1; x) \equiv \int \alpha(y, x) K_h(y - Y_1) K_h(x - X_1)\, \mathrm{d}y - \int \beta(y, w) K_H(y - Y_1) K_H(w - W_1)\, \mathrm{d}y,$$

$$\tilde{r}_N(Y_1, X_1; x) \equiv r_N(Y_1, X_1; x) - \mathrm{E}[r_N(Y_1, X_1; x)] \tag{A.9}$$

and decompose I_N according to

$$\begin{aligned} I_N &= N^{-2} \int \left(\sum_{j=1}^{N} r_N(Y_j, X_j; x) \right)^2 a(x)\, \mathrm{d}F(x) \\ &= N^{-2} \sum_{j,k=1}^{N} \int r_N(Y_j, X_j; x) r_N(Y_k, X_k; x) a(x)\, \mathrm{d}F(x) \\ &= N^{-2} \left\{ 2 \sum_{1 \leqslant j < k \leqslant N} \int \tilde{r}_N(Y_j, X_j; x) \tilde{r}_N(Y_k, X_k; x) a(x)\, \mathrm{d}F(x) \right. \\ &\quad + \sum_{j=1}^{N} \int r_N^2(Y_j, X_j; x) a(x)\, \mathrm{d}F(x) \\ &\quad + 2(N-1) \sum_{j=1}^{N} \int \tilde{r}_N(Y_j, X_j; x) \mathrm{E}[r_N(Y_1, X_1; x)] a(x)\, \mathrm{d}F(x) \\ &\quad \left. + N(N-1) \int \mathrm{E}[r_N(Y_1, X_1; x)]^2 a(x)\, \mathrm{d}F(x) \right\} \\ &\equiv I_{N1} + I_{N2} + I_{N3} + I_{N4}. \end{aligned} \tag{A.10}$$

We shall show under our assumptions that all these terms are asymptotically Gaussian but that I_{N3}, I_{N4} are asymptotically negligible while I_{N2} gives a bias term. The centered term I_{N1} gives the asymptotic variance. For that purpose, we make use of the following.

Lemma 4 (Hall, 1984). *Let* $\{Z_i : i = 1, \ldots, N\}$ *be an i.i.d. sequence. Suppose that the U-statistic* $U_N \equiv \sum_{1 \leqslant i < j \leqslant N} \tilde{P}_N(Z_i, Z_j)$ *with symmetric variable function* $\tilde{P}_N$ *is centered* (*i.e.*, $\mathrm{E}[\tilde{P}_N(Z_1, Z_2)] = 0$) *and degenerate* (*i.e.*, $\mathrm{E}[\tilde{P}_N(Z_1, Z_2) | Z_1 = z_1] = 0$ *almost surely for all* z_1). *Let*

$$\sigma_N^2 \equiv \mathrm{E}[\tilde{P}_N(Z_1, Z_2)^2], \qquad \tilde{\Pi}_N(z_1, z_2) \equiv \mathrm{E}[\tilde{P}_N(Z_i, z_1) \tilde{P}_N(Z_i, z_2)].$$

Then if

$$\lim_{N\to\infty} \frac{\mathrm{E}[\tilde{I\!I}_N(Z_1,Z_2)^2] + N^{-1}\mathrm{E}[\tilde{P}_N(Z_1,Z_2)^4]}{(\mathrm{E}[\tilde{P}_N(Z_1,Z_2)^2])^2} = 0, \tag{A.11}$$

we have that as $N \to \infty$

$$\frac{2^{1/2}}{N\sigma_N} U_N \to \mathrm{N}(0,1).$$

We can then obtain the following.

Lemma 5. Under Assumptions 1–3

$$Nh^{(p+q)/2}[I_N - N^{-1}h^{-(p+q)}\gamma_{12} - N^{-1}h^{-p}\gamma_{22} - N^{-1}H^{-p}\gamma_{32})] \to \mathrm{N}(0,\sigma_{11}^2), \tag{A.12}$$

where σ_{11} *and the* γ_{j2} *are given above.*

By Lemma 2 and (A.2), the result (A.12) is also valid if we replace I_N by $\Gamma(\hat{f}_h, \hat{f}_H, F)$ there. We then show that $\hat{\sigma}_{11}^2$ and the respective $\hat{\gamma}_{j2}$, $j = 1,2,3$, can be substituted for σ_{11}^2 and γ_{j2} in Theorem 1 with no effect on the asymptotic distribution.

Lemma 6. Under Assumptions 1–3, $\hat{\sigma}_{11}^2 - \sigma_{11}^2 = o_{\mathrm{p}}(1)$, $\hat{\gamma}_{12} - \gamma_{12} = o_{\mathrm{p}}(h^{(p+q)/2})$, $\hat{\gamma}_{22} - \gamma_{22} = o_{\mathrm{p}}(h^{(p+q)/2}H^{-p})$ *and* $\hat{\gamma}_{32} - \gamma_{32} = o_{\mathrm{p}}(h^{(p+q)/2}H^{-p})$.

The theorem follows in view of (A.4) and the following lemma.

Lemma 7. Given Assumptions 1–3, *we have that*

$$\begin{aligned}\Delta_N &= O_{\mathrm{p}}(N^{-3}(h^{-3(p+q)} + H^{-3p}) + N^{-1}(H^{2r} + h^{2r})) \\ &= o_{\mathrm{p}}(N^{-2}h^{-(p+q)}).\end{aligned} \tag{A.13}$$

A.2. The asymptotic normality of the test statistic

There is really one key feature that drives the structure of our results. In particular, the limiting distributional structure of the test statistic is determined by the nonparametric estimation of the unrestricted model. For instance, the rate of convergence is determined by the dimensionality and bandwidth for $\hat{m}_h(w,v)$, the general regression. The place where this arises in the proof is in the decomposition (A.10) of I_N, where the presence of both w and v makes for a slower rate of convergence for $\hat{m}_h(w,v)$. In other words, the specification of the unrestricted model is the only factor determining the asymptotic behavior

of the test statistic. This feature is a strength of our approach, because similar distributional features will arise for a wide range of null hypotheses. Indeed, we explored several variations in Section 4.

Another noteworthy feature of the result concerns the normality of the limiting distribution. Typically, second order terms of the von Mises expansion will be distributed as an infinite weighted sum of chi squared variables. This structure is associated with condition (A.11)—under this condition, the degenerate U-statistic U_N has a limiting normal distribution, but otherwise, it would typically have a weighted sum of chi-squares distribution. The normal distribution occurs here because the eigenvalues λ_{jN}, $j=1,2,\ldots,\infty$, of the linear operator Ψ_N on L_2 defined by

$$\psi \in L_2 \mapsto (\Psi_N \psi)(z) \equiv \mathrm{E}[\tilde{P}_N(Z_i, z)\psi(z)]$$

have the asymptotic negligibility property that

$$\lim_{N\to\infty} \frac{(\sum_{j=1}^{\infty} |\lambda_{jN}|^4)^{1/2}}{\sum_{j=1}^{\infty} |\lambda_{jN}|^2} = 0.$$

Rather than attempt to check this condition directly, which would be practically impossible since the eigenvalues are not known explicitly, we relied on the sufficient characterization (A.11).

A.3. Proofs

Proof of Lemma 1. Define

$$\Psi(t) \equiv \iint \left(\frac{\int y f(y,w,v)\,\mathrm{d}y + t\int y g_1(y,w,v)\,\mathrm{d}y}{f(w,v) + t g_1(w,v)} - \frac{\int y f(y,w)\,\mathrm{d}y + t\int y g_2(y,w)\,\mathrm{d}y}{f(w) + t g_2(w)} \right)^2 a(w,v)\,\mathrm{d}F_3(w,v),$$

where (g_1, g_2) are such that $(tg_1, tg_2) \in \Omega_1 \times \Omega_2$ for all $0 \leqslant t \leqslant 1$. From the explicit expression above and the properties of f, g_1 and g_2, it follows that Ψ is three times continuously differentiable in t on $[0,1]$. We start by applying Taylor's formula with Lagrange reminder to Ψ:

$$\Psi(t) = \Psi(0) + t\Psi'(0) + t^2\Psi''(0)/2 + t^3\Psi'''(\vartheta(t))/6, \tag{A.14}$$

where $0 \leqslant \vartheta(t) \leqslant t$. Note that $\Psi(0) = 0$ under H_0. Then define

$$\varphi(t,w,v) \equiv \frac{\int y f(y,w,v)\,\mathrm{d}y + t\int y g_1(y,w,v)\,\mathrm{d}y}{f(w,v) + t g_1(w,v)} - \frac{\int y f(y,w)\,\mathrm{d}y + t\int y g_2(y,w)\,\mathrm{d}y}{f(w) + t g_2(w)}. \tag{A.15}$$

It is immediate to compute that

$$\Psi'(t) = 2 \iint \varphi(t,w,v) \frac{\partial \varphi(t,w,v)}{\partial t} a(w,v)\, \mathrm{d}F_3(w,v)$$

and therefore $\Psi'(0) = 0$ since $\varphi(0,w,v) \equiv 0$ under H_0. Note that since $\Psi'(0) = 0$ the terms linear in (g_1, g_2) vanish from the functional expansion under the null hypothesis, which is the source of the singularity for the test statistic.

Next, we calculate

$$\Psi''(t) = 2 \iint \left\{ \varphi(t,w,v) \frac{\partial^2 \varphi(t,w,v)}{\partial t^2} + \left[\frac{\partial \varphi(t,w,v)}{\partial t} \right]^2 \right\} a(w,v)\, \mathrm{d}F_3(w,v)$$

and therefore $\Psi''(0) = 2 \iint [\partial \varphi / \partial t(0,w,v)]^2 a(w,v)\, \mathrm{d}F_2(w,v)$ under H_0.

To characterize the reminder term, we finally need to compute

$$\Psi'''(t) = 2 \iint \left\{ \varphi(t,w,v) \frac{\partial^3 \varphi(t,w,v)}{\partial t^3} + 3 \frac{\partial \varphi(t,w,v)}{\partial t} \frac{\partial^2 \varphi(t,w,v)}{\partial t^2} \right\} \times a(w,v)\, \mathrm{d}F_3(w,v) \tag{A.16}$$

and bound each term. Note that $\varphi(t,w,v)$ and $\partial \varphi / \partial t$ are linear in (g_1, g_2), while $\partial^2 \varphi / \partial t^2$ is quadratic and $\partial^3 \varphi / \partial t^3$ is cubic. Indeed

$$\frac{\partial \varphi(t,w,v)}{\partial t} = \frac{f(w,v) \int y g_1(y,w,v)\, \mathrm{d}y - g_1(w,v) \int y f(y,w,v)\, \mathrm{d}y}{\{f(w,v) + t g_1(w,v)\}^2} - \frac{f(w) \int y g_2(y,w)\, \mathrm{d}y - g_2(w) \int y f(y,w)\, \mathrm{d}y}{\{f(w) + t g_2(w)\}^2},$$

similarly

$$\frac{\partial^2 \varphi(t,w,v)}{\partial t^2} = -2 \frac{\{f(w,v) \int y g_1(y,w,v)\, \mathrm{d}y - g_1(w,v) \int y f(y,w,v)\, \mathrm{d}y\} g_1(w,v)}{\{f(w,v) + t g_1(w,v)\}^3} + 2 \frac{\{f(w) \int y g_2(y,w)\, \mathrm{d}y - g_2(w) \int y f(y,w)\, \mathrm{d}y\} g_2(w)}{\{f(w) + t g_2(w)\}^3}$$

and

$$\frac{\partial^3 \varphi(t,w,v)}{\partial t^3} = 6 \frac{\{f(w,v) \int y g_1(y,w,v)\, \mathrm{d}y - g_1(w,v) \int y f(y,w,v)\, \mathrm{d}y\} g_1^2(w,v)}{\{f(w,v) + t g_1(w,v)\}^4} - 6 \frac{\{f(w) \int y g_2(y,w)\, \mathrm{d}y - g(w) \int y f(y,w)\, \mathrm{d}y\} g_2^2(w)}{\{f(w) + t g_2(w)\}^4}.$$

It is then easy to show that each term in (A.16) is of order $O(||g_1||^3+||g_2||^3)$, by noting that

$$\frac{1}{|f(w,v)+tg_1(w,v)|} \leqslant \frac{1}{|f(w,v)|-|g_1(w,v)|} \leqslant \frac{2}{b}$$

since $0 \leqslant t \leqslant 1$, $|f(w,v)| \geqslant b$ and $g_1(w,v)| \leqslant b/2$.[20] Therefore

$$|\Psi'''(\vartheta(t))| = O(||g_1||^3 + ||g_2||^3)$$

as long as $||g_1|| \leqslant b/2$ and $||g_2|| \leqslant b/2$.

From (A.14), evaluated at $t=1$, we have obtained that

$$\begin{aligned}\Psi(t) &= \Psi''(0)/2 + \Psi'''(\vartheta(t))/6 \\ &= \iint \left[\frac{\partial\varphi}{\partial t}(0,w,v)\right]^2 a(w,v)\,\mathrm{d}F_3(w,v) + O(||g_1||^3 + ||g_2||^3),\end{aligned}$$

where

$$\begin{aligned}&\frac{\partial\varphi(0,w,v)}{\partial t} \\ &\quad= \left\{f(w,v)\int yg_1(y,w,v)\,\mathrm{d}y - g_1(w,v)\int yf(y,w,v)\,\mathrm{d}y\right\}\Big/ f(w,v)^2 \\ &\qquad- \left\{f(w)\int yg_2(y,w)\,\mathrm{d}y - g_2(w)\int yf(y,w)\,\mathrm{d}y\right\}\Big/ f(w)^2 \\ &\quad= \int \alpha(y,w,v)g_1(y,w,v)\,\mathrm{d}y - \int \beta(y,w)g_2(y,w)\,\mathrm{d}y,\end{aligned}$$

since

$$\int \alpha(y,w,v)f(y,w,v)\,\mathrm{d}y = \beta(y,w)f(y,w)\,\mathrm{d}y = 0$$

and the lemma follows. □

Proof of Lemma 2. The bound on the L_∞ deviations of the kernel density estimator

$$||\hat{f}_h(w,v) - f(w,v)||_\infty \equiv \sup_{(w,v)\in S} |\hat{f}_h(w,v) - f(w,v)|$$

[20] In the case of squared-density weighting, these terms are absent (hence no bounds) because the weighting effectively cancels out the denominators.

is a classical result: see Stone (1983, Lemmas 2 and 8). For the other part of the seminorm $||\cdot||$ defined in (A.1),

$$\sup_{(w,v)\in S}\left|\int y(\hat{f}_h(y,w,v)-f(y,w,v))\,\mathrm{d}y\right|$$

$$\leqslant ||\hat{f}_h(w,v)-f(w,v)||_\infty ||m(w,v)||_\infty + ||\hat{m}_h(w,v)-m(w,v)||_\infty ||\hat{f}_h(w,v)||_\infty$$

and for N large enough $||\hat{f}_h(w,v)||_\infty \leqslant 2||f(w,v)||_\infty$. Then

$$||\hat{m}_h(w,v)-m(w,v)||_\infty = O_{\mathrm{p}}(h^r + N^{-1/2}h^{-(p+q)/2}\ln(N)) \tag{A.17}$$

goes to zero at the same rate as $||\hat{f}_h(w,v)-f(w,v)||_\infty$: see Härdle (1990, Section 4). The bound follows for $||\hat{f}_h(w,v)-f(w,v)||$.

Finally note that under 2 we have $h^p \gg H^p \gg h^{p+q}$ (so both the asymptotic bias and variance of $\hat{f}_H(w)$ are smaller than those of $\hat{f}_h(w,v)$) and thus all the norms involving $\hat{f}_H(w)-f(w)$ are strictly smaller than those involving $\hat{f}_h(w,v)-f(w,v)$. □

Proof of Lemma 3. Apply Lemma 1 with $g_1(y,w,v)=\hat{f}_h(y,w,v)-f(y,w,v)$ and $g_2(y,w)=\hat{f}_H(y,w)-f(y,w)$. This can be done since by Lemma 2,

$$\Pr\left[||\hat{f}_h(w,v)-f(w,v)||_\infty \geqslant \frac{b}{2} \text{ and } ||\hat{f}_H(w)-f(w)||_\infty \geqslant \frac{b}{2}\right] \to 0$$

so that

$$\Pr[(g_1,g_2)\in \Omega_1\times\Omega_2] \to 1. \quad \square$$

Proof of Lemma 5. We begin with some essential bounds and expansions:

$$\mathrm{E}[r_N(Y_1,X_1;x)] = \int\int f(y_1,x_1)\left[\int \alpha(y,x)K_h(y-y_1)K_h(x-x_1)\,\mathrm{d}y - \int \beta(y,w)K_h(y-y_1)K_H(w-w_1)\,\mathrm{d}y\right]\mathrm{d}x_1\,\mathrm{d}y_1.$$

Changing first the order of integration and then the variables from y_1 to $u=(y-y_1)/h$, and from x_1 to $s=(x-x_1)/h$ in the first term and similarly in the second we get, if $x=(w,v)$ and $s=(\tau,\sigma)$,

$$\mathrm{E}[r_N(Y_1,X_1;x)] = \int\left[\int\int f(y-uh,x-hs)\alpha(y,x-hs)K(u)K(s)\,\mathrm{d}u\,\mathrm{d}s\right.$$

$$- \int\int f(y - uH, x - Hs)\beta(y, w - H\tau)K(u)K(\tau)\,\mathrm{d}u\,\mathrm{d}\tau\Big]\,\mathrm{d}y$$

$$= \int f(y,x)(\alpha(y,x) - \beta(y,w))\,\mathrm{d}y + O(h^r) + O(H^r)$$

$$= O(h^r) + O(H^r), \tag{A.18}$$

uniformly in x on S by Assumption 3. On the other hand,

$$\mathrm{E}[|\tilde{r}_N(Y_1, X_1; x)|] = O(1) \tag{A.19}$$

by the same argument. Similarly

$$\int \mathrm{E}[r_N^2(Y_1, X_1; x)]a(x)\,\mathrm{d}F(x)$$

$$= \iint\int \Bigg[\left\{\int \alpha(y,x)K_h(y - y_1)K_h(x - x_1)\,\mathrm{d}y\right\}^2$$

$$-2\left\{\int \alpha(y,x)K_h(y - y_1)K_h(x - x_1)\,\mathrm{d}y\right\}$$

$$\left\{\int \beta(y,w)K_H(y - y_1)K_H(w - w_1)\,\mathrm{d}y\right\}$$

$$+\left\{\int \beta(y,w)K_H(y - y_1)K_H(w - w_1)\,\mathrm{d}y\right\}^2\Bigg]$$

$$f(y_1, x_1)\,\mathrm{d}x_1\,\mathrm{d}y_1 a(x)f(x)\,\mathrm{d}x. \tag{A.20}$$

Changing variables to $u = (y - y_1)/h$, $t = (x - x_1)/h$ in the first term, to $u = (y - y_1)/h$, $u_2 = (y_2 - y_1)/H$, $s = (w - w_1)/H$ and $t = (v - v_1)/h$ in the second and $u = (y - y_1)/H$, $t = (w - w_1)/H$ in the third we obtain

$$N\,\mathrm{E}[I_{N2}] = \int \mathrm{E}[r_N^2(Y_1, X_1; x)]a(x)\,\mathrm{d}F(x)$$

$$= \gamma_{12}h^{-(p+q)}(1 + O(h^{2r})) + \gamma_{22}h^{-p}(1 + O(h^{2r}))$$

$$+ \gamma_{32}H^{-p}(1 + O(H^{2r})), \tag{A.21}$$

where we also note that

$$\mathrm{E}[r_N^2(Y_1, X_1; x)] = O(h^{-(p+q)}). \tag{A.22}$$

Then

$$\operatorname{Var}\left[N^{-2}\sum_{j=1}^{N}\int r_N^2(Y_j,X_j;x)a(x)\,\mathrm{d}F(x)\right]$$

$$= N^{-3}\operatorname{Var}\left[\int r_N^2(Y_1,X_1;x)a(x)\,\mathrm{d}F(x)\right]$$

and

$$\operatorname{Var}\left[\int r_N^2(Y_1,X_1;x)a(x)\,\mathrm{d}F(x)\right] = \mathrm{E}\left[\left(\int r_N^2(Y_1,X_1;x)a(x)\,\mathrm{d}F(x)\right)^2\right]$$

$$-\left(\mathrm{E}\left[\int r_N^2(Y_1,X_1;x)a(x)\,\mathrm{d}F(x)\right]\right)^2$$

$$= O(h^{-2(p+q)}),$$

(where the first term dominates). By Chebyshev's Inequality it follows that

$$Nh^{(p+q)/2}\{I_{N2} - \gamma_{12}N^{-1}h^{-(p+q)} - \gamma_{22}N^{-1}h^{-p} - \gamma_{32}N^{-1}H^{-p}\} = o_{\mathrm{p}}(1). \tag{A.23}$$

which characterizes the asymptotic bias term.

Next, we express I_{N1} in U-statistic form where $Z_i \equiv (Y_i, W_i)$ as

$$I_{N1} \equiv \sum_{1\leqslant j<k\leqslant N}\tilde{P}_{N1}(Z_j,Z_k)$$

with

$$\tilde{P}_{N1}(Z_j,Z_k) \equiv \frac{2}{N^2}\int \tilde{r}_N(Y_j,X_j;x)\tilde{r}_N(Y_k,X_k;x)a(x)\,\mathrm{d}F(x), \tag{A.24}$$

which by construction verifies the centering and degeneracy conditions. We compute

$$\mathrm{E}[\tilde{P}_{N1}(Z_1,Z_2)^2] = \left(\frac{1}{N^2h^{2(p+q)}}\right)^2 h^{3(p+q)}\sigma_{11}^2,$$

$$\mathrm{E}[\tilde{I\!I}_{N1}(Z_1,Z_2)^2] \equiv \mathrm{E}_{Z_1,Z_2}\{\mathrm{E}_{Z_i}[\tilde{P}_{N1}(Z_i,Z_1)\tilde{P}_{N1}(Z_i,Z_2)]^2\}$$

$$= O\left(\left(\frac{1}{N^2h^{2(p+q)}}\right)^4 h^{7(p+q)}\right),$$

and

$$\mathrm{E}[\tilde{P}_{N1}(Z_1,Z_2)^4] = O\left(\left(\frac{1}{N^2h^{2(p+q)}}\right)^4 h^{5(p+q)}\right)$$

so that condition (A.11) follows from $h^{(p+q)} + 1/(Nh^{(p+q)}) \to 0$. Therefore, from Lemma 4, we have

$$Nh^{(p+q)/2} I_{N1} \to \mathrm{N}(0, \sigma_{11}^2). \tag{A.25}$$

From (A.19) we have

$$\mathrm{E}\left[\left(\int \tilde{r}_N(Y_1, X_1; x)\mathrm{E}[r_N(Y_2, X_2; x)]a(x)\,\mathrm{d}F(x)\right)^2\right] = O(h^{2r})$$

and it follows that

$$\mathrm{E}[(Nh^{(p+q)/2} I_{N3})^2] = O(Nh^{(p+q)}h^{2r}) = o(1)$$

and then again from Chebyshev's Inequality we have that

$$Nh^{(p+q)/2} I_{N3} = o_{\mathrm{p}}(1). \tag{A.26}$$

Finally from (A.18) we conclude for the deterministic term I_{N4} that

$$Nh^{(p+q)/2} I_{N4} = Nh^{(p+q)/2} O(h^{2r} + H^{2r}) = o(1). \tag{A.27}$$

The lemma follows for gathering the four terms I_{Ni}, $i = 1, \ldots, 4$. □

Proof of Lemma 6. Write for $(w, v) \in S$

$$R(w, v) \equiv \frac{\hat{\sigma}_h^4(w, v)}{\hat{f}_h(w, v)} - \frac{\sigma^4(w, v)}{f(w, v)}.$$

By definition, for $(w, v) \in S$,

$$|R(w, v)| \leqslant 2\frac{\max\{\hat{\sigma}_h^2(w, v), \sigma^2(w, v)\}}{f(w, v)}|\hat{\sigma}_h^2(w, v) - \sigma^2(w, v)|$$

$$+ \frac{\hat{\sigma}_h^4(w, v)}{\hat{f}_h(w, v) f(w, v)}|\hat{f}_h(w, v) - f(w, v)|. \tag{A.28}$$

It is easy to verify that

$$\frac{2C_{11}}{N}\sum_{i=1}^{N}\frac{\sigma^4(W_i, V_i)A_i}{f(W_i, V_i)} = \sigma_{11}^2 + O_{\mathrm{p}}(N^{-1/2}).$$

Therefore to establish $\hat{\sigma}_{11}^2 = \sigma_{11}^2 + o_{\mathrm{p}}(1)$ it certainly suffices to show that

$$\sup_{(w,v)\in S}|\hat{\sigma}_h^2(w, v) - \sigma^2(w, v)| = O_{\mathrm{p}}(h^r + N^{-1/2}h^{-(p+q)/2}\ln(N)). \tag{A.29}$$

This follows from noting that (3.16) is nothing other than

$$\hat{\sigma}_h^2(w, v) = \hat{E}_h[Y^2 | W = w, V = v] - (\hat{E}_h[Y | W = w, V = v])^2,$$

where $\hat{E}_h$ denotes the Nadaraya–Watson regression estimator and then applying the bound (A.17) to the conditional regression of Y^2 instead of that of Y.

Finally, we have that

$$\left|\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{\sigma}_h^4(W_i,V_i)A_i}{\hat{f}_h(W_i,V_i)}-\frac{\sigma^4(W_i,V_i)A_i}{f(W_i,V_i)}\right)\right| \leqslant \sup_{(w,v)\in S}|R(w,v)a(w,v)|$$

and the argument is complete. The same argument will also yield

$$\hat{\gamma}_{12}=\gamma_{12}+O_{\mathrm{p}}(h^{-(p+1)/2})$$

by Assumption 3 and the other terms are dealt with in the same way. □

Proof of Lemma 7. We have that

$$\begin{aligned}\Delta_N &= N^{-2}\sum_{j,k=1,\ldots,n}\int r_N(Y_j,X_j;x)r_N(Y_k,X_k;x)a(x)\mathrm{d}(\hat{F}-F)(x)\\ &= N^{-3}\sum_{j,k,l=1,\ldots,n}\left\{ r_N(Y_j,X_j;X_l)r_N(Y_k,X_k;X_l)a(X_l)\right.\\ &\quad\left.-\int r_N(Y_j,X_j;x)r_N(Y_k,X_k;x)a(x)\,\mathrm{d}F(x)\right\}.\end{aligned}$$

The proof that Δ_N is small is direct, and by brute force. For this we write

$$\Delta_N=\Delta_{N1}+\Delta_{N2}+\Delta_{N3}+\Delta_{N4},$$

where

$$\begin{aligned}\Delta_{N1} &= N^{-3}\sum_{l\neq j,k}\{r_N(Y_j,X_j;X_j)r_N(Y_k,X_k,X_l)a(X_l)\\ &\quad-\mathrm{E}[r_N(Y_j,X_j;X_l)r_N(Y_k,X_k;X_l)a(X_l)\,|\,X_j,X_k)]\}\end{aligned}$$

are the centered terms with $l\neq j,k$, where we used the identity

$$\begin{aligned}&\mathrm{E}[r_N(Y_j,X_j;X_l)r_N(Y_k,X_k;X_l)a(X_l)\,|\,X_j,X_k]\\ &\quad=\int r_N(Y_j,Y_j;x)r_N(Y_k,X_k;x)a(x)\,\mathrm{d}F(x)\end{aligned}$$

if $l\notin\{j,k\}$;

$$\Delta_{N2}=2N^{-3}\sum_{j\neq k}r_N(Y_j,X_j;X_j)r_N(Y_k,X_k;X_j)a(X_j)$$

are the terms with $l=j\neq k$, and $k\neq j$;

$$\Delta_{N3}=N^{-3}\sum_{j}r_N^2(Y_j,X_j;X_j)a(X_j)$$

are the terms with $l=j=k$ and

$$\Delta_{N4}=-N^{-3}\sum_{j,k}\int r_N(Y_j,X_j;x)r_N(Y_k,X_k;x)a(x)\,\mathrm{d}F(x)$$

which are the centering terms for Δ_{N2} and Δ_{N3}. Dispensing with the simpler ones first, we have

$$\Delta_{N4} = -N^{-1}\Gamma(\hat{f}_h, \hat{f}_H, F) = o_p(N^{-1}h^{-(p+q)/2}).$$

by Lemma 5.

Further

$$\begin{aligned}\mathrm{E}[\Delta_{N3}] &= N^{-2}\mathrm{E}[r_N^2(Y_1, X_1; X_1)a(X_1)] \\ &= o(N^{-1}h^{-(p+q)/2}),\end{aligned}$$

since $\mathrm{E}[r_N^2(Y_1, X_1; X_1)a(X_1)] = O(h^{-(p+q)})$ by (A.22).

The other terms are more difficult, and require explicit consideration of their components. Taking up the main term, let $z = (y, x)$, and write

$$\Delta_{N1} = \Delta_{N11} + \Delta_{N12},$$

$$\Delta_{N11} = N^{-3}\sum_{l\neq j,k} W^*_{jkl},$$

$$\Delta_{N12} = N^{-3}\sum_{l\neq j,k} \mathrm{E}[W_{jkl} \mid X_l],$$

where

$$W_{jkl} = r_N(Z_j; X_l)r_N(Z_k; X_l)a(X_l) - \mathrm{E}[r_N(Z_j; X_l)r_N(Z_k; X_l)a(X_l) \mid X_j, X_k],$$

$$\begin{aligned}W^*_{jkl} &= W_{jkl} - \mathrm{E}[W_{jkl} \mid X_l] \\ &= r_N(Z_j; X_l)r_N(Z_k; X_l)a(X_l) - \mathrm{E}[r_N(Z_j; X_l)r_N(Z_k; X_l)a(X_l)|x_l] \\ &\quad - \mathrm{E}[r_N(Z_j; X_l)r_N(Z_k; X_l)a(X_l) \mid X_j, X_k] \\ &\quad + \mathrm{E}[r_N(Z_j; X_l)r_N(Z_k; X_l)a(X_l)].\end{aligned}$$

Note that, if $l \neq j, k$,

$$\mathrm{E}[W_{jkl}] = 0.$$

Now, by (A.22)

$$\begin{aligned}\mathrm{E}[\Delta_{N12}] &= O(N^{-2}\mathrm{E}[r_N^2(Z_1; X_2)]) \\ &= O(N^{-2}h^{-(p+q)}) \\ &= o(N^{-1}h^{-(p+q)/2})\end{aligned}$$

and

$$\mathrm{Var}[\Delta_{N12}] = O(N^{-1})$$

since $\mathrm{E}[\mathrm{E}[W_{123}|X_3]^2] = O(1)$.

For the first term, we have

$$\mathrm{E}[\Delta_{N11}] = 0,$$
$$\mathrm{E}[\Delta_{N11}^2] = N^{-6} \sum_{l \notin \{j,k,j',k'\} \text{ and } \{j,k\} \frown \{j',k'\} \neq \emptyset} \mathrm{E}[W_{jkl}^* W_{j'k'l}^*].$$

Since $\mathrm{E}[W_{jkl}^* \,|\, X_l] = 0$ and $\mathrm{E}[W_{jkl}^* \,|\, X_j, X_k] = 0$ if $l \neq j,k$. Now, by arguing as for (A.20), we have $O(N^3)$ terms of the form

$$\mathrm{E}[W_{jkl}^{*2}] \leqslant \mathrm{E}[\mathrm{E}[r_N(Z_1; X_2 \,|\, X_2]^2]$$
$$= O(h^{-2(p+q)}), \quad j \neq k,$$

$O(N^2)$ terms of the form

$$\mathrm{E}[W_{jjl}^{*2}] \leqslant \mathrm{E}[r_N(Z_1, X_2; X_2)^4]$$
$$= O(h^{-3(p+q)}).$$

and $O(N^4)$ terms of the form

$$\mathrm{E}[W_{jkl}^* W_{jk'l}^*] = O(h^{-(p+q)}).$$

In any case, combining all of these gives

$$\mathrm{E}[\Delta_{N11}^2] = o(N^{-2}h^{-(p+q)})$$

so that we are finished with Δ_{N1}.

The remaining term Δ_{N2} is analyzed in exactly the same fashion, namely by squaring (obtaining the factor N^{-6}) and keeping track of expectations of terms of the form

$$r_N(Y_j, X_j; X_j) r_N(Y_k, X_k; X_j) a(X_j) r_N(Y_{j'}, X_{j'}; X_{j'}) r_N(Y_{k'}, X_{k'}; X_{j'}) a(X_{j'}),$$

where $j \neq k$ and $j' \neq k'$. There are $O(N^4)$ terms with $j \neq k \neq j' \neq k'$ each of $O(1)$, giving an overall contribution of $O(N^{-2}) \to 0$. There are $O(N^2)$ terms with $j = j'$ and $k = k'$ each of $O(h^{-2(p+q)})$, giving an overall contribution of $O(N^{-4}h^{-2(p+q)}) \to 0$. There are $O(N^3)$ terms with $j = j'$ and $k \neq k'$ each of $O(h^{-(p+q)})$, giving an overall contribution of $O(N^{-3}h^{-(p+q)}) \to 0$. In any case, $\mathrm{E}[\Delta_{N2}^2] = o(N^{-2}h^{-(p+q)})$, and the lemma is proved. □

Proof of Proposition 1. The analysis is similar to that of Lemmas 1 and 7, keeping now the additional terms that were not present under the null. That is, as before

$$\Psi(t) = \Psi(0) + t\Psi'(0) + t^2\Psi''(0)/2 + t^3\Psi'''(\vartheta(t))/6 \tag{A.30}$$

with $\Psi(0) = \Gamma(f, f, F_3)$, $\varphi(0, w, v) = m(w, v) - M(w)$,

$$\Psi'(0) = 2 \iint \varphi(0, w, v) \frac{\partial \varphi(0, w, v)}{\partial t} a(w, v)\, \mathrm{d}F_3(w, v)$$

and

$$\Psi''(0) = 2 \iint \left\{ \varphi(0, w, v) \frac{\partial^2 \varphi(0, w, v)}{\partial t^2} + \left[\frac{\partial \varphi(0, w, v)}{\partial t} \right]^2 \right\} a(w, v)\, \mathrm{d}F_3(w, v),$$

where

$$\frac{\partial \varphi(0, w, v)}{\partial t} = \int \alpha(y, w, v) g_1(y, w, v)\, \mathrm{d}y - \int \beta(y, w) g_2(y, w)\, \mathrm{d}y \tag{A.31}$$

and

$$\frac{\partial^2 \varphi(0, w, v)}{\partial t^2} = -2 \frac{g_1(w, v)}{f(w, v)} \int \alpha(y, w, v) g_1(y, w, v)\, \mathrm{d}y + 2 \frac{g_2(w)}{f(w)} \int \beta(y, w) g_2(y, w)\, \mathrm{d}y.$$

To characterize the reminder term, we have that

$$\Psi'''(t) = 2 \iint \left\{ \varphi(t, w, v) \frac{\partial^3 \varphi(t, w, v)}{\partial t^3} + 3 \frac{\partial \varphi(t, w, v)}{\partial t} \frac{\partial^2 \varphi(t, w, v)}{\partial t^2} \right\} a(w, v)\, \mathrm{d}F_3(w, v) \tag{A.32}$$

and bound each term as in Lemma 1 to obtain

$$|\Psi'''(\vartheta(t))| = O(||g_1||^3 + ||g_2||^3).$$

It follows that

$$\tilde{\Gamma} = \Gamma(f, f, F) + \tilde{\Pi}$$

where $\tilde{\Pi}$ is such that $\hat{\sigma}_{11}^{-1}(Nh^{(p+q)/2} \cdot \tilde{\Pi} - h^{-(p+q)/2}\hat{\gamma}_{12} - h^{(q-p)/2}\hat{\gamma}_{22} - h^{(p+q)/2} H^{-p}\hat{\gamma}_{32}) = O_{\mathrm{p}}(1)$. Then from the definition of $\hat{\tau}$ in (3.15), it follows that

$$\hat{\tau} = \hat{\sigma}_{11}^{-1}(Nh^{(p+q)/2} \cdot \Gamma(f, f, F)) + O_{\mathrm{p}}(1)$$

and therefore $\hat{\tau} \xrightarrow{\mathrm{p}} \infty$ if $\Gamma(f, f, F) > 0$, i.e., if the null hypothesis is false. The test is therefore consistent. □

Proof of Proposition 2. Our assumptions are such that the preceding arguments remain valid even when Z_{iN}, $1 \leqslant i \leqslant N$, are a double array. Expansion (A.3) and subsequent lemmas, with the additional terms arising under the alternative, continue to hold. Thus, we have that under H_{1N},

$$\hat{\tau} - \hat{\sigma}_{11}^{-1} Nh^{(p+q)/2} \Gamma(f, f, \hat{F}) \to \mathrm{N}(0, 1).$$

Moreover, $\hat{\sigma}_{11} \xrightarrow{\mathrm{P}^{[N]}} \sigma_{11}$ and

$$\begin{aligned}\Gamma(f^{[N]}, f^{[N]}, \hat{F}) &= \frac{1}{N}\sum_{i=1}^{N}\{m^{[N]}(X_i) - M^{[N]}(W_i)\}^2 a(X_i) \\ &= \mathrm{E}[(m^{[N]}(X_1) - M^{[N]}(W_1))^2]a(X_i) + O_{\mathrm{p}}(N^{-1/2}) \\ &= \delta_2 N^{-1}h^{-(p+q)/2} + o_{\mathrm{p}}(N^{-1}h^{-(p+q)/2}). \end{aligned} \tag{A.33}$$

The proposition follows. □

Proof of Corollary 1. Define

$$Em_h(w,v) \equiv \sum_{i=1}^{N} K_h(w - W_i, v - V_i)M(W_i)/\hat{f}_h(w,v)$$

so that

$$\begin{aligned}\tilde{\Gamma} &= N^{-1}\sum_{i=1}^{N}\{\hat{m}_h(W_i,V_i) - Em_{h,H}(W_i,V_i)\}^2 A_i \\ &\quad + 2N^{-1}\sum_{i=1}^{N}\{\hat{m}_h(W_i,V_i) - Em_{h,H}(W_i,V_i)\} \\ &\quad \{\hat{E}m_{h,H}(W_i,V_i) - \hat{E}m_{h,H}(W_i,V_i)\}A_i \\ &\quad + N^{-1}\sum_{i=1}^{N}\{\hat{E}m_{h,H}(W_i,V_i) - Em_{h,H}(W_i,V_i)\}^2 A_i \\ &\equiv J_{N1} + J_{N2} + J_{N3}.\end{aligned}$$

First, by considering $\tilde{Y}_i \equiv Y_i - M(W_i)$ and observations $(W_i, V_i, \tilde{Y}_i)$ we have by Proposition 1 in Härdle and Mammen (1993) that

$$Nh^{(p+q)/2}J_{N1} = O_{\mathrm{p}}(1) \tag{A.34}$$

and has the requisite distribution. Next, note that J_{N3} can further be written as

$$\begin{aligned}J_{N3} &= N^{-1}\sum_{i=1}^{N}\left\{\sum_{j=1}^{N}K_h(W_i - W_j, V_i - V_j)(\hat{M}_H(W_j) - M(W_j))\right/ \\ &\quad \left.\hat{f}_h(W_j,V_j)\right\}^2 A_i \\ &\leqslant \sup_{(w,v)\in S}(\hat{M}_H(w) - M(w))^2 a(w,v) = O_{\mathrm{p}}(H^{2r} + N^{-1}H^{-(p+q)}\ln^2(N))\end{aligned}$$

from Lemma 2, and hence $Nh^{(p+q)/2}J_{N3} = o_p(1)$ under Assumption 2. Finally by Schwartz' Inequality we have that $J_{N2} \leqslant J_{N1}^{1/2} J_{N3}^{1/2}$ and therefore $Nh^{(p+q)/2} J_{N2} = o_p(1)$. The result follows. □

Proof of Corollary 2. Let

$$\tilde{\Gamma}(\theta) \equiv \frac{1}{N}\sum_{i=1}^{N} \{\hat{m}_h(w(x_i,\theta), v(x_i,\theta)) - \hat{M}_H(w(x_1,\theta))\}^2 A_i. \tag{A.35}$$

Then, Theorem 1 applies to $\tilde{\Gamma}(\theta_0)$ where θ_0 is the true value of θ. Consider the statistic we intend to use, $\tilde{\Gamma}(\hat{\theta})$, and apply Taylor's formula with Lagrange remainder:

$$\begin{aligned}\tilde{\Gamma}(\hat{\theta}) &= \tilde{\Gamma}(\theta_0) + \tilde{\Gamma}'(\theta_0)(\hat{\theta}-\theta_0) + \tfrac{1}{2}\tilde{\Gamma}''(\theta_0)(\hat{\theta}-\theta_0, \hat{\theta}-\theta_0) \\ &\quad + \tfrac{1}{6}\tilde{\Gamma}'''(\tilde{\theta})(\hat{\theta}-\theta_0, \hat{\theta}-\theta_0, \hat{\theta}-\theta_0) \\ &\equiv \tilde{\Gamma}(\theta_0) + \Delta_{N1} + \Delta_{N2} + \Delta_{N3},\end{aligned} \tag{A.36}$$

where $\tilde{\Gamma}'$, etc., are differentials with respect to θ, and $\tilde{\theta}$ is between θ_0 and $\hat{\theta}$. Now, without loss of generality, $w(x,\theta_0) = w$, $v(x,\theta_0) = v$ and we can write $\tilde{\Gamma}'(\theta_0) = \vartheta(\hat{f}_h, \hat{f}_H, \hat{f}_h', \hat{f}_H')$ where

$$\begin{aligned}\vartheta(f_1, f_2, f_1', f_2') = 2\int \left\{ \frac{\int y f_1(y,x)\,\mathrm{d}y}{f_1(x)} - \frac{\int y f_2(y,w)\,\mathrm{d}y}{f_2(w)} \right\} \\ \{\dot{M}(w,x,f_1,f_1') - \dot{m}(w,x,f_2,f_2')\} a(x)\,\mathrm{d}\hat{F}(x),\end{aligned}$$

f_j' are the gradients with respect to x and w, respectively, of f_j and

$$\begin{aligned}\dot{m}(w,x,f_2,f_2') \equiv \int y f_2'(y,x)\,\mathrm{d}y \dot{x}(\theta_0) \Big/ \int f_2(y,x)\,\mathrm{d}y \\ - \left(\int y f_2(y,x)\,\mathrm{d}y \int f_2'(y,x)\,\mathrm{d}y \right) \dot{x}(\theta_0) \Big/ \\ \left(\int f_2(x,y)\,\mathrm{d}y \right)^2.\end{aligned}$$

$\dot{M}$ is defined similarly and $\dot{x}(\theta_0)$ is the derivative of W at θ_0, with vector and matrix multiplication properly defined. Now, under (4.1), $\vartheta(f,f,f_1',f_2') = 0$ for all f_1', f_2'. Thus

$$\vartheta(\hat{f}_h, \hat{f}_H, \hat{f}_h', \hat{f}_H') = O_p(N^{-1}h^{-3(p+q)}) \tag{A.37}$$

by analogy with Lemma 5 since

$$(\hat{f}_h - f)^2 = O_p(N^{-1}h^{-(p+q)})$$

but $(\hat{f}_h' - f')^2 = O_p(N^{-1}h^{-3(p+q)})$, and similarly for $\hat{f}_H$ and $\hat{f}_H'$.

Combining (A.37) with $\hat{\theta}-\theta_0 = O_p(N^{-1/2})$ we conclude that Δ_{N1} is $O_p(N^{-3/2} h^{-3(p+q)})$ so that $Nh^{(p+q)/2}\Delta_{N1} = o_p(1)$. Similarly $\tilde{\Gamma}''(\theta_0) = O_p(1)$, thus

$$Nh^{(p+q)/2}\Delta_{N2} = O_p(h^{(p+q)/2}) = o_p(1),$$

and

$$\sup\{|\tilde{\Gamma}'''(\tilde{\theta})| : |\tilde{\theta} - \theta_0| \leqslant \varepsilon\} = O_p(h^{-7(p+q)}).$$

Thus the conclusion of Theorem 1 will continue to hold if $\delta > 13(p+q)$, which insures that

$$Nh^{(p+q)/2}\Delta_{N3} = o_p(1). \quad \square$$

Proof of Corollary 3. Note that

$$\tilde{\Gamma} = \Gamma(\hat{f}_h, f(\cdot;\hat{\theta}), \hat{F}),$$

where $f(\cdot;\theta)$ is such that $Y|X=x$ has the parametric model distribution $F(\cdot|x;\theta)$. The argument of Theorem 1 with $\hat{f}_H$ replaced by $f(\cdot;\theta_0)$ where θ_0 is the true value of θ, yields that $\Gamma(\hat{f}_h, f(\cdot;\theta_0), \hat{F})$ obeys the conclusion of Theorem 1 (and subsequent propositions) with $\gamma_{12} = \gamma_{32} = 0$ (and no restrictions on Δ which does not appear since $p=0$). But, by Lemma 6,

$$\begin{aligned}\Gamma(\hat{f}_h, f(\cdot;\hat{\theta}), \hat{F}) = \Gamma(\hat{f}_h, f(\cdot;\theta_0), \hat{F}) + \iint \beta(y,w)(f(y,w;\hat{\theta}) \\ - f(y,w;\theta_0))^2 \,\mathrm{d}y\, a(w)\,\mathrm{d}\hat{F}(w) \\ + O_p(||f(\cdot;\hat{\theta}) - f(\cdot;\theta_0)||^3). \qquad \text{(A.38)}\end{aligned}$$

Under our assumption it is easy to see that the first term is $O_p(N^{-1})$ and the second $O_p(N^{-3/2})$. The corollary follows. $\square$

Proof of Corollary 4. Note that

$$\tilde{\Gamma}^* = \Gamma(\hat{f}_h(\cdot,\cdot;\hat{\theta}), \hat{f}_H(\cdot,\cdot;\hat{\theta}), \hat{F}),$$

where $\hat{f}_h(\cdot,\cdot;\theta)$ is the smoothed density of $(Y_i(\theta), X_i)$ where $Y_i(\theta) \equiv Y_i - m(V_i;\theta)$ and similarly for $\hat{f}_H(\cdot,\cdot;\theta)$. Thus,

$$\begin{aligned}\tilde{\Gamma}^* = \frac{1}{N}\sum_{i=1}^{N}\{\hat{m}_h^*(W_i,V_i;\theta_0) - \hat{M}_H^*(W_i;\theta_0)\}^2 A_i \\ - \frac{2}{N}\sum_{i=1}^{N}(\hat{m}_h^*(W_i,V_i;\theta_0) - \hat{M}_H^*(W_i;\theta_0))(\hat{C}_N(V_i) - \hat{D}_N(V_i))A_i \\ + \frac{1}{N}\sum_{i=1}^{N}\{\hat{C}_N(V_i) - \hat{D}_N(V_i)\}^2 A_i, \qquad \text{(A.39)}\end{aligned}$$

where

$$\hat{C}_N(V_i) = \sum_{j=1}^{N} (m(V_i;\hat{\theta}) - m(V_i;\theta_0))K_h(X_i - X_j) \Big/ \sum_{j=1}^{N} K_h(X_i - X_j),$$

$$\hat{D}_N(V_i) = \sum_{j=1}^{N} (m(V_i;\hat{\theta}) - m(V_i;\theta_0))K_H(W_i - W_j) \Big/ \sum_{j=1}^{N} K_H(W_i - W_j).$$

With the parameter vector one-dimensional for simplicity of notation, we have that

$$\hat{C}_N(V_i) - \hat{D}_N(V_i) = (\hat{\theta} - \theta_0)\left(\frac{\partial m}{\partial \theta}(V_i;\theta_0) - \mathrm{E}\left[\frac{\partial m_\theta}{\partial \theta}(V_i;\theta_0) \mid W_i\right]\right) (1 + o_{\mathrm{p}}(1)).$$

Thus, the third term in (A.39) is $O_{\mathrm{p}}(N^{-1})$. The second is

$$O_{\mathrm{p}}\left(2\left\{\frac{1}{N}\sum_{i=1}^{N}\{\hat{m}_h^*(W_i,V_i;\theta_0) - \hat{M}_H^*(W_i;\theta_0)\}^2 A_i\right\}^{1/2} \times \left\{\frac{1}{N}\sum_{i=1}^{N}\{\hat{C}_N(V_i) - \hat{D}_N(V_i)\}^2 A_i\right\}^{1/2}\right)$$

$$= O_{\mathrm{p}}(N^{-1/2}h^{-(p+q)/4}N^{-1/2}) = o_{\mathrm{p}}(N^{-1}h^{-(p+q)/2}),$$

since the first obeys Theorem 1 under (4.4). □

Proof of Proposition 3. The conclusion of Lemma 4 remains valid under the additional Assumption 7, in light of Theorem 1 in Khashimov (1992), which provides a generalization to absolutely regular sequences of the central limit theorem for degenerate U-statistics with sample-dependent functions. The bounds for the density errors in Lemma 2 such as $||\hat{f}_h - f||_\infty = O(h^r + N^{-1/2}h^{-(p+q)/2}\log N)$ and similarly for the regression function are given in the mixing context by Gyorfi et al. (1989, Section III. 3). The covariations induced by the dependence in the data are of higher order, and therefore have no impact. The additional terms arising in the asymptotic distribution of I_{N1} because of the time-series nature of the data, are of the form $\mathrm{E}[\tilde{P}_{N1}(z_i,z_j)\tilde{P}_{N1}(z_k,z_l)]$ (defined in (A.24)) and can be shown to be of order $O(h^{4p})$—while the terms already present in the i.i.d. context are dominant (from (A.25)). Our other supporting lemmas are unaffected and the result therefore follows. □

References

Aït-Sahalia, Y., 1996. Nonparametric pricing of interest rate derivative securities. Econometrica 64, 527–560.

Aït-Sahalia, Y., 1998. The delta method for nonparametric kernel functionals. Working paper, Graduate School of Business, University of Chicago.

Aït-Sahalia, Y., Lo, A.W., 1998. Nonparametric estimation of state price densities implicit in financial asset prices. Journal of Finance 53, 499–547.

Aït-Sahalia, Y., Lo, A.W., 2000. Nonparametric risk management and implied risk aversion. Journal of Econometrics 94, 9–51.

Bickel, P.J., Rosenblatt, M., 1973. On some global measures of the deviations of density function estimates. Annals of Statistics 1, 1071–1096.

Bickel, P.J., Götze, F., van Zwet, W.R., 1997. Resampling fewer than n observations: gains, losses and remedies for losses. Statistica Sinica 1, 1–31.

Bickel, P.J., Ritov, Y., Stoker, T.M., 1998. Testing and the method of sieves, Working paper, Department of Statistics, University of California at Berkeley.

Bierens, H.J., 1990. A consistent conditional moment test of functional form. Econometrica 58, 1443–1458.

Bierens, H.J., Ploberger, W., 1997. Asymptotic theory of integrated conditional moment tests. Econometrica 65, 1129–1151.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of Political Economy 81, 637–654.

Blundell, R., Duncan, A., 1998. Kernel regression in empirical microeconomics. Journal of Human Resources 33, 62–87.

Chen, X., Fan, Y., 1999. Consistent hypothesis testing in semiparametric and nonparametric models of economic time series. Journal of Econometrics 91, 373–401.

Christoffersen, P., Hahn, J., 1998. Nonparametric testing of ARCH for option pricing. Working paper, University of Pennsylvania.

Doksum, K., Samarov, A., 1995. Global functionals and a measure of the explanatory power of covariates in nonparametric regression. Annals of Statistics 23, 1443–1473.

Ellison, G., Fisher-Ellison, S., 2000. A simple framework for nonparametric specification testing. Journal of Econometrics 96, 1–23.

Eubank, R., Spiegelman, C., 1990. Testing the goodness-of-fit of linear models via nonparametric regression techniques. Journal of the American Statistical Association 85, 387–392.

Fan, J., 1994. Testing the goodness-of-fit of a parametric density function by kernel method. Econometric Theory 10, 316–356.

Fan, J., Li, Q., 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. Econometrica 64, 865–890.

Fernandes, M., 1999. Nonparametric tests for Markovian dynamics. Working paper, European University Institute in Florence.

Fernholz, L., 1983. Von Mises Calculus for Statistical Functionals. Lecture Notes in Statistics, Vol. 19. Springer, New York.

Filippova, A.A., 1962. Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. Theoretical Probability and Applications 7, 24–57.

Fraga, M., 1999. Parametric and semiparametric estimation of sample selection models: an application to the female labor force. Working paper, Université Libre de Bruxelles.

Gozalo, P.L., 1993. A consistent model specification test for nonparametric estimation of regression functions models. Econometric Theory 9, 451–477.

Gyorfi, L., Härdle, W., Sarda, P., Vieu, P., 1989. Nonparametric curve estimation from time series. Lecture Notes in Statistics, Vol. 60. Springer, New York.

Hall, P., 1984. Central limit theorem for integrated squared error of multivariate nonparametric density estimators. Journal of Multivariate Analysis 14, 1–16.
Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge, UK.
Härdle, W., Mammen, E., 1993. Comparing nonparametric vs. parametric regression fits. Annals of Statistics 21, 1926–1947.
Heckman, J., Ichimura, I., Smith, J., Todd, P., 1998. Characterization of selection bias using experimental data. Econometrica 66, 1017–1098.
Hidalgo, J., 1992. A general non-parametric misspecification test. Working paper, Department of Economics, London School of Economics.
Hong, Y., White, H., 1995. Consistent specification testing via nonparametric series regression. Econometrica 63, 1133–1160.
Horowitz, J.L., Härdle, W., 1994. Testing a parametric model against a semiparametric alternative. Econometric Theory 10, 821–848.
de Jong, R.M., Bierens, H.J., 1994. On the limit behavior of a chi-squared type test if the number of conditional moments tested approaches infinity. Econometric Theory 10, 70–90.
Khashimov, Sh.A., 1992. Limiting behavior of generalized U-statistics of weakly dependent stationary processes. Theory of Probability and Its Applications 37, 148–150.
Lavergne, P., Vuong, Q.H., 1996. Nonparametric selection of regressors: the nonnested case. Econometrica 64, 207–219.
Lavergne, P., Vuong, Q.H., 2000. Nonparametric significance testing. Econometric Theory 16, 576–601.
Lee, B.J., 1988. Nonparametric tests using a kernel estimation method. Ph.D. Dissertation. Department of Economics, University of Wisconsin at Madison.
Lewbel, A., 1991. Applied consistent tests of nonparametric regression and density restrictions. Working paper, Department of Economics, Brandeis University.
Li, Q., 1999. Consistent model specification tests for time series econometric models. Journal of Econometrics 92, 101–147.
von Mises, R., 1947. On the asymptotic distribution of differentiable statistical functions. Annals of Mathematical Statistics 18, 309–348.
Reeds, J.A., 1976. On the definition of von Mises functionals. Ph.D. Dissertation. Harvard University, Cambridge.
Robinson, P.M., 1988. Root n semiparametric regression. Econometrica 56, 931–954.
Robinson, P.M., 1989. Hypothesis testing in semiparametric and nonparametric models for econometric time series. Review of Economic Studies 56, 511–534.
Rodriguez, D., Stoker, T.M., 1992. A regression test of semiparametric index model specification. Working paper, Sloan School of Management, MIT.
Sakov, A., 1998. Using the m out of n bootstrap in hypothesis testing. Ph.D. Dissertation, Department of Statistics, University of California at Berkeley.
Staniswalis, J.G., Severini, T.A., 1991. Diagnostics for assessing regression models. Journal of the American Statistical Association 86, 684–691.
Stoker, T.M., 1992. Lectures on Semiparametric Econometrics. CORE Foundation, Louvain-la-Neuve.
Stone, C.J., 1983. Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. Recent Advances in Statistics, Vol. 393–406. Academic Press, New York.
Vieu, P., 1994. Choice of regressors in nonparametric estimation. Computational Statistics and Data Analysis 17, 575–594.
Whang, Y.-J., Andrews, D.W.K., 1991. Tests of specification for parametric and semiparametric models. Journal of Econometrics 57, 277–318.
White, H., Hong, Y., 1993. M-testing using finite and infinite dimensional parameter estimators. Working paper, University of California at San Diego.

Wooldridge, J., 1992. A test for functional form against nonparametric alternatives. Econometric Theory 8, 452–475.
Yatchew, A.J., 1992. Nonparametric regression tests based on an infinite dimensional least squares procedure. Econometric Theory 8, 435–451.
Zhang, P., 1991. Variable selection in nonparametric regression with continuous covariates. Annals of Statistics 19, 1869–1882.
Zheng, J.X., 1996. A consistent test of functional form via nonparametric estimation function. Journal of Econometrics 75, 263–290.