# GOODNESS-of-FIT TESTS for REGRESSION USING KERNEL METHODS


*** Yacine Aït-Sahalia**
**** Peter J. Bickel**
***** Thomas M. Stoker**



**SWP#3970** **Revised June 1997**



* University of Chicago
** University of California at Berkeley
*** Massachusetts Institute of Technology



November 1994

# Goodness-of-Fit Tests for Regression Using Kernel Methods

Yacine Aït-Sahalia
University of Chicago

Peter J. Bickel
University of California at Berkeley

Thomas M. Stoker
Massachusetts Institute of Technology

November 1994. Revised June 1997.

**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.

## 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 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

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

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 is based on goodness-of-fit, via residual sums of squares under the null and alternative regression hypotheses.

We derive the expansion of the functional for the sum-of-squared departures between the restricted regression and the unrestricted regression, along the method of von Mises (1947).[2] We carry out the second order expansion, because the distribution of our test statistic has a singularity when the null hypothesis is true.[3] In particular, the first order (influence) terms vanish under the null hypothesis, 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 results are related to a fairly recent but growing literature in econometrics on hypothesis testing with nonparametric methods. This literature has focused on testing a specific parametric model against flexible alternatives, with variations in the type of approach and particular specifications of null and alternative hypotheses. 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 non-nested nonparametric alternative, and Lavergne and Vuong (1994) propose a residual based test for specification of regressors under the similar guidelines; see also Doksum and Samarov (1993). Yatchew (1992) and Whang and Andrews (1991) propose methods based on "sample splitting," or using different parts of the original data sample for estimation and testing. Also related is work based on the cross validation function used for choosing smoothing parameters in non-

[2] See also Filipova (1962), Reeds (1976), Fernholz (1983), and Ait-Sahalia (1995) among others.

[3] 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 (1993), Aït-Sahalia (1996b) and Bierens and Ploberger (1994), among others. We are not aware however of applications of the Von Mises expansion to the testing problem, that we carry out.

parametric regression: see Zhang (1991) and Vieu (1993). 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. Recent analysis of second order terms of tests of parametric models against general non-parametric alternatives are carried out by Bierens and Ploberger(1994) and de Jong and Bierens (1994) for general orthogonality restrictions, by Zheng (1996) for tests of parametric models and Ellison and Fisher-Ellison (1992) for tests of linear models. Analysis of 'density-weighted' test statistics using kernel estimators are given in Staniswalis and Severini (1991), Hidalgo (1992), White and Hong (1996), and Fan and Li (1996). 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, Ichimura, Smith and Todd (1994) 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.

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. As such, our results are generally applicable, 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., as in the existing literature, although some regularization parameters need to be specified.

## 2 Basic Framework

Suppose that we are studying a response $y$ as a function of a vector of predictor variables $z$, where the data sample consists of $N$ independent and identically distributed observations $(y_i, z_i), i = 1, ..., N$.

Our base case concerns testing whether some of the predictor variables can be omitted from the regression of $y$ on $z$. In particular, suppose $z = (w, v)$, where $w$ is a vector of dimension $p$ and $v$ is a vector of dimension $q$. The joint density (resp. cumulative distribution function) of $(y, w, v)$ is denoted $f$ (resp. $F$). Below we need to make reference to several marginal densities from $f(y, w, v)$ and estimates of $f$ which we denote via the list of arguments — for example $f(w, v) \equiv \int_y f(y, w, v)dy$, where $\int_y$ denotes integration with respect to $y$, etc. While this notation is compact, we feel that it is sufficiently unambiguous.

Our interest is in comparing the regression of $y$ on $(w, v)$ to the regression of $y$ on $w$. The regression of $y$ on $(w, v)$ is defined as

$$m(w, v) \equiv E(y|w, v) = \frac{\int_y y f(y, w, v)dy}{f(w, v)} \tag{2.1}$$

and the regression of $y$ on $w$ as

$$M(w) \equiv E(y|w) = \frac{\int_y y f(y, w)dy}{f(w)}.$$

We are interested in whether $v$ can be omitted from the regression $m(w,v)$, namely the null hypothesis is

$$H_0 : \Pr\left[m\left(w,v\right) = M\left(w\right)\right] = 1.$$

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

$$H_1 : \Pr\left[m\left(w,v\right) = M\left(w\right)\right] < 1.$$

Our testing approach is to assess the significance of squared differences in nonparametric kernel estimates of the functions $m$ and $M$; or in particular, to measure the mean squared difference $E\left\{\left[m\left(w,v\right) - M\left(w\right)\right]^2\right\}$. We first introduce the kernel estimators, and then the test statistic of interest.

Based on a kernel function $\mathcal{K}$ and bandwidth $h$, the standard Nadaraya-Watson kernel regression estimator of $m\left(w,v\right)$ is

$$\hat{m}\left(w,v\right) \equiv \frac{\sum_{i=1}^{N} \mathcal{K}\left(\frac{w-w_i}{h}, \frac{v-v_i}{h}\right) y_i}{\sum_{i=1}^{N} \mathcal{K}\left(\frac{w-w_i}{h}, \frac{v-v_i}{h}\right)} \tag{2.2}$$

while

$$\hat{M}\left(w\right) \equiv \frac{\sum_{i=1}^{N} \mathcal{K}\left(\frac{w-w_i}{H}\right) y_i}{\sum_{i=1}^{N} \mathcal{K}\left(\frac{w-w_i}{H}\right)} \tag{2.3}$$

is the standard estimator of $M\left(w\right)$ with bandwidth $H$. [4]

It is convenient to represent $\hat{m}$ and $\hat{M}$ in terms of two density estimates,

$$\hat{f}(y,w,v) = \frac{1}{N}\sum_{i=1}^{N} h^{-(p+q+1)}\mathcal{K}\left(\frac{y-y_i}{h}, \frac{w-w_i}{h}, \frac{v-v_i}{h}\right)$$

$$\hat{\mathrm{f}}(y,w) = \frac{1}{N}\sum_{i=1}^{N} H^{-(p+1)}\mathcal{K}\left(\frac{y-y_i}{H}, \frac{w-w_i}{H}\right)$$

Similarly, we define the estimates $\hat{f}(y,w)$ of $f(w,v)$, calculated with bandwidth $h$, and $\hat{\mathrm{f}}(w)$ of $f(w)$, calculated with bandwidth $H$. For simplicity, we take each $\mathcal{K}$ to be a product kernel with the same component kernel $\mathcal{K}$ which is of order $r$.

Note that if we indicate the dependence on $f$ in (2.1) by $m(w,v) \equiv m(w,v,f)$, and similarly for $M$ then $\hat{m}(w,v) = m(w,v,\hat{f})$, $\hat{M}(w) = M(w,\hat{\mathrm{f}})$.

For technical reasons, we choose to focus on comparing $\hat{M}\left(w\right)$ to $\hat{m}\left(w,v\right)$ in areas where there is sufficient density. Rather than assuming that the support of the density $f$ is compact, and $f$ is bounded away from zero on its support, we only compare $\hat{M}$ to $\hat{m}$ on a compact set where the density is known to be bounded away from zero. The conclusions we derive are undoubtedly valid under weaker hypotheses. Let $a\left(w,v\right)$ be a bounded weighting function with compact support $S \subset R^{p+q}$; for example, the indicator function of a compact set. Define $a_i \equiv a\left(w_i, v_i\right)$. Our test statistic is

$$\tilde{\Gamma} \equiv \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{m}\left(w_i,v_i\right) - \hat{M}\left(w_i\right)\right\}^2 a_i \tag{2.4}$$

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

that is, we compare $\hat{M}(w)$ to $\hat{m}(w,v)$ everywhere the comparison of the conditional expectations is practically meaningful.[5] In fact, a slight modification of (2.4) where $\hat{m}(w,v)$ is replaced by $\hat{m}(w,v) - \int \hat{m}(w,v)\hat{f}(w,v)dwdv$ and $\hat{M}(w)$ is replaced by $\hat{M}(w) - \int \hat{M}(w)\hat{\mathrm{f}}(w)dw$ is more appropriate but this makes no difference in practice.

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

$$\Gamma(F_{11}, F_{12}, F_2) \equiv \int_v \int_w \left\{ \frac{\int_y y f_{11}(y,w,v)dy}{f_{11}(w,v)} - \frac{\int_y y f_{12}(y,w)dy}{f_{12}(w)} \right\}^2 a(w,v)\, dF_2(w,v) \tag{2.5}$$

defined formally on $\mathcal{F}_\infty \times \mathcal{F}_\infty \times \mathcal{F}_\in$. Here $F_2 \in \mathcal{F}_\in$, the set of all probabilities on $R^{p+1}$ and $F_{11}, F_{12} \in \mathcal{F}_\infty$, a suitable convex set of functions of bounded variation on $R^{p+q+1}$ which are absolutely continuous with (Radon Nikodym) derivatives. If $f_{1j}$, $j = 1, 2$ are the derivatives of $F_{1j}$ then the quantities appearing in the definition (2.5), $m(w, v, f_{11})$, $M(w, f_{12})$ are assumed bounded as $F_{1j}$ range over $\mathcal{F}_\infty$. Under the null hypothesis $H_0$, we have $\Gamma[F, \mathrm{F}, F_2] = 0$ for all $F_2$, and under the alternative $H_1$, we have $\Gamma[F, \mathrm{F}, F_2] > 0$ for some $F_2$.

Since the restricted regression is nested in the general regression as

$$M(w) = E_F[m(w,v)\,|w]$$

we have several alternative formulations of the goodness-of-fit functional. First,

$$\Gamma(F, F, F) = E_F\{m(w,v)[y - M(w)]\, a(w,v)\}$$

so that $\Gamma(F, \mathrm{F}, F) = 0$ coincides with the orthogonality restriction that $m(w,v)$ is uncorrelated with the restricted residual $y - M(w)$. By construction, this is equivalent to the fact that $y - M(w)$ is uncorrelated with any function of $w$ and $v$. Second, we have that

$$\Gamma(F, F, F) = E_F\left\{\left([y - M(w)]^2 - [y - m(w,v)]^2\right) a(w,v)\right\}$$

so that $\Gamma(F, \mathrm{F}, F) = 0$ is associated with no improvement in residual variance (least squares goodness-of-fit) from including $v$ in the regression analysis of $y$.

Further, $\tilde{\Gamma} = \Gamma(\hat{F}, \hat{\mathrm{F}}, \tilde{F})$ where $\tilde{F}$ is the empirical c.d.f. of the data. Since $\hat{F}, \hat{\mathrm{F}}$ and $\tilde{F}$ are all consistent a test of $H_0$ using $\tilde{\Gamma}$ can be expected to be consistent.[6]

# 3 The Distribution of the Test Statistic

## 3.1 Assumptions and the Main Result

Our assumptions are:

**Assumption 1** *The data $\{(y_i, w_i, v_i); i = 1, \ldots, N\}$ are i.i.d. with distribution $F$.*

[5] As mentioned in the Introduction, 'density weighted' sums-of-squares tests have been analyzed for certain problems. Here, a 'density weighted' version of (2.4) would be constructed by replacing $a_i$ by $\hat{f}(w_i, v_i)\, a_i$.

[6] It is clear that our testing approach is a nonparametric analogue to traditional $\chi^2$ 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 $\Gamma[F, \mathrm{F}, F]$ divided by an estimate of residual variance, scaled for degrees of freedom.

**Assumption 2** *We have that*

1. *The density* $f(y,w,v)$ *is* $r+1$ *times continuously differentiable,* $r \geq 2$. $f$ *and its derivatives are bounded and in* $L_2\left(R^{1+p+q}\right)$. *Let* $\mathcal{D}$ *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* $\inf_S f(w,v) \equiv b > 0$.

3. $$E(y - m(w,v))^4 < \infty$$

   *and*

   $$\sigma^2(w,v) \equiv E\left[(y - m(w,v))^2 | w, v\right] \tag{3.1}$$

   *satisfies* $\sigma^4(w,v) \in L_2\left(R^{p+q}\right)$. *The restricted conditional variance*

   $$\sigma^2(w) \equiv E\left[(y - M(w))^2 | w\right] \tag{3.2}$$

   *satisfies* $\sigma^2(w) \in L_2(R^p)$.

**Assumption 3** *For kernel estimation*

1. *The kernel* $\mathcal{K}$ *is a bounded function on* $R$, *symmetric about* 0, *with* $\int |\mathcal{K}(z)|dz < \infty$, $\int \mathcal{K}(z)dz = 1$, $\int z^j \mathcal{K}(z)dz = 0$ $1 \leq j < r$. *Further,*

   $$r > 3(p+q)/4. \tag{3.3}$$

2. *As* $N \to \infty$, *the unrestricted bandwidth sequence* $h = O\left(N^{-1/\delta}\right)$ *is such that*

   $$2(p+q) < \delta < 2r + (p+q)/2 \tag{3.4}$$

   *while the restricted bandwidth* $H = O\left(N^{-1/\Delta}\right)$ *satisfies*

   $$p < \Delta \leq 2r + p \tag{3.5}$$

   *as well as*

   $$\delta p/(p+q) \leq \Delta < \delta. \tag{3.6}$$

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 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 $0 \leq R < \infty$, $H/h \to 0$ 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, define the further notation

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

$$\begin{aligned} \gamma_{12} &\equiv C_{12} \int_v \int_w \left[\int_y \alpha(y,w,v)^2 f(y,w,v)\,dy\right] f(w,v)\,a(w,v)\,dwdv \\ &= C_{12} \int_v \int_w \sigma^2(w,v)\,a(w,v)\,dwdv \end{aligned} \tag{3.8}$$

$$\begin{aligned}\gamma_{22} &\equiv -2C_{22}\int_v\int_w\left[\int_y \alpha(y,w,v)\,A(y,w)\,f(y,w,v)\,dy\right]f(w,v)\,a(w,v)\,dwdv \\ &= -2C_{22}\int_v\int_w \sigma^2(w,v)\,\{f(w,v)/f(w)\}\,a(w,v)\,dwdv\end{aligned} \tag{3.9}$$

$$\gamma_{32} \equiv C_{32}\int_v\int_w\left[\int_y A(y,w)^2 f(y,w)\,dy\right]f(w)\,a(w)\,dw = C_{32}\int_w \sigma^2(w)\,a(w)\,dw \tag{3.10}$$

$$\begin{aligned}\sigma_{11}^2 &\equiv 2C_{11}\int_v\int_w\left[\int_y \alpha(y,w,v)^2 f(y,w,v)\,dy\right]^2 f(w,v)^2\,a(w,v)^2\,dwdv \\ &= 2C_{11}\int_v\int_w \sigma^4(w,v)\,a(w,v)^2\,dwdv\end{aligned} \tag{3.11}$$

where $a(w) \equiv E[a(w,v)\,|w] = \int_v a(w,v)\,f(w,v)\,dv/f(w)$ and the $C_{ij}$'s are constants determined by the kernel as

$$C_{12} \equiv \int_v\int_w \mathcal{K}(w,v)^2\,dwdv,\ C_{22} \equiv K_w(0),\ C_{32} \equiv \int_w K_w(w)^2\,dw, \tag{3.12}$$

$$C_{11} \equiv \int_{\tilde{v}}\int_{\tilde{w}}\left[\int_v\int_w \mathcal{K}(w,v)\,\mathcal{K}(w+\tilde{w},v+\tilde{v})\,du_w du_v\right]^2 d\tilde{w}d\tilde{v}. \tag{3.13}$$

Note that under Assumption 2-3, $\gamma_{j2}$, $j = 1,2,3$ and $\sigma_{11}^2$are finite, as well as $\Gamma(F,F,F)$.

Our result is now stated as:

**Theorem 1** *Under Assumptions 1, 2 and 3, we have that under $H_0$*

$$\sigma_{11}^{-1}\left[Nh^{(p+q)/2}\tilde{\Gamma} - h^{-(p+q)/2}\gamma_{12} - h^{(q-p)/2}\gamma_{22} - h^{(p+q)/2}H^{-p}\gamma_{32}\right] \longrightarrow \mathcal{N}(0,1)$$

To implement the test, we require estimates $\hat{\sigma}_{11}^2$ of $\sigma_{11}^2$ and $\hat{\gamma}_{j2}$ of $\gamma_{j2}$, $j = 1,2,3$. We then compare

$$\hat{\tau} = \hat{\sigma}_{11}^{-1}\left(Nh^{(p+q)/2}\cdot\tilde{\Gamma} - h^{-(p+q)/2}\hat{\gamma}_{12} - h^{(q-p)/2}\hat{\gamma}_{22} - h^{(p+q)/2}H^{-p}\hat{\gamma}_{32}\right) \tag{3.14}$$

to the critical value $z_\alpha$, the quantile at level $\alpha$ of the $\mathcal{N}(0,1)$ distribution.

Our statistic and the hypotheses also make sense if $p = 0$, i.e., there is no $w$. In that case $\hat{f}$ is not defined but we simply take $M(w) \equiv \hat{M}(w) \equiv 0$ 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 $\gamma_{j2}$ and $\sigma_{11}^2$ depend on (3.1), 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

$$\hat{\sigma}^2(w,v) = \frac{\sum_{i=1}^N \mathcal{K}\left(\frac{w-w_i}{h},\frac{v-v_i}{h}\right)y_i^2}{\sum_{i=1}^N \mathcal{K}\left(\frac{w-w_i}{h},\frac{v-v_i}{h}\right)} - \hat{m}(w,v)^2$$

for the unrestricted regression, and

$$\hat{\sigma}^2(w) = \frac{\sum_{i=1}^N \mathcal{K}\left(\frac{w-w_i}{H}\right)y_i^2}{\sum_{i=1}^N \mathcal{K}\left(\frac{w-w_i}{H}\right)} - \hat{M}(w)^2$$

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

$$\hat{\sigma}_{11}^2 = \frac{2C_{11}}{N}\sum_{i=1}^N \frac{\hat{\sigma}^4(w_i,v_i)a_i^2}{\hat{f}(w_i,v_i)},\quad \hat{\gamma}_{12} = \frac{C_{12}}{N}\sum_{i=1}^N \frac{\hat{\sigma}^2(w_i,v_i)a_i}{\hat{f}(w_i,v_i)},$$

$$\hat{\gamma}_{22} = -\frac{2C_{22}}{N}\sum_{i=1}^{N}\frac{\hat{\sigma}(w_i,v_i)^2 a_i}{\hat{f}(w_i)}, \quad \hat{\gamma}_{32} = \frac{C_{32}}{N}\sum_{i=1}^{N}\frac{\hat{\sigma}(w_i)^2 a_i}{\hat{f}(w_i)}.$$

The following lemma shows that $\hat{\sigma}_{11}^2$ and the respective $\hat{\gamma}_{j2}$, $j = 1, 2, 3$, can be substituted for $\sigma_{11}^2$ and $\gamma_{j2}$ in Theorem 1 with no effect on the asymptotic distribution:

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

Finally, the constants $C_{ij}$ are determined by the kernel chosen as in ( 3.12) and (3.13) and are easily computed. For example, for the Gaussian product kernel of order $r = 2$ (density of $\mathcal{N}(0,1)$), we have that

$$C_{12} = 1/(2\sqrt{\pi})^{p+q}, \; C_{22} = 1/(\sqrt{2\pi})^{p}, \; C_{32} = 1/(2\sqrt{\pi})^{p}, \; C_{11} = 1/(2\sqrt{2\pi})^{p+q}. \quad (3.15)$$

This complete the description of the test statistic $\hat{\tau}$ in ( 3.14).

We now give a proof of Theorem 1. Sections 3.4 and 3.5 give some intuition for the result, while section 3.6 studies the consistency and asymptotic power of the test.

### 3.3 Proof of the Theorem

We begin by studying the asymptotic properties of the functional $\Gamma$ evaluated at $\left(\hat{F}, \hat{\mathrm{F}}, F\right)$, using the functional delta method. The only difference between $\Gamma\left(\hat{F}, \hat{\mathrm{F}}, F\right)$ and $\tilde{\Gamma}$ of (2.4) is that the latter is an average over the empirical c.d.f. $\tilde{F}$ instead of $F$ . We then show in Lemma 7 that this difference is inconsequential for the asymptotic distribution of the test statistic. To bound the remainder term in the functional expansion of $\Gamma$, define the norms

$$\begin{aligned}
\|yg(y,w,v)\|_2^2 &\equiv \sup_{w,v} a(w,v)\left(\int yg(y,w,v)dwdv\right)^2 \\
\|yg(y,w)\|_2 &\equiv \left[\int_w \left[\int_y yg(y,w)\,dy\right]^2 dw\right]^{1/2} \\
\|yg(y,w)\|_2 &\equiv \left[\int_v \int_w g^2(w,v)\,dwdv\right]^{1/2} \\
\|g(w,v)\|_2 &\equiv \left[\int_w g^2(w)\,dw\right]^{1/2} \\
\|g\|_\infty &\equiv \sup\left(g(w,v)\,|(w,v) \in S\right)
\end{aligned}$$

and

$$\|g\| \equiv \max\left(\|g\|_\infty, \|yg(y,w,v)\|_2, \|yg(y,w)\|_2, \|g(w,v)\|_2, \|g(w)\|_2\right).$$

The main ingredient in the proof of the theorem is the functional expansion of $\Gamma(\cdot,\cdot,\cdot)$, summarized as:

**Lemma 2** *Let* $\mathcal{F}_\infty = \{\mathcal{F} + \mathcal{G} : \mathcal{G} \in \mathcal{V}\}$ *where* $\mathcal{V}$ *is the set of all absolutely continuous functions of bounded variation such that if* $G = \int g d\lambda$ *where* $\lambda$ *is Lebesgue measure then* $\|g\|_\infty < b/2$. *Then, under Assumption 2 and* $H_0$, $\Gamma(\cdot,\cdot,F_2)$ *has an expansion on* $\mathcal{F}_\infty \times \mathcal{F}_\infty$ *about* $(F,F)$ *given by,*

$$\begin{aligned}
\Gamma(F+G_1, F+G_2, F_2) = \Gamma(F,F,F_2) \\
+ \int_v \int_w \Big[\int_y (\alpha(y,w,v)g_1(y,w,v) - \\
A(y,w)g_2(y,w))dy\Big]^2 dF_2(w,v) + R(g_1,g_2,F_2)
\end{aligned}$$

*where*

$$\sup\{|R(G_1,G_2,F_2)|\,/(\|g_1\|^3 + \|g_2\|^3) : G_1, G_2 \in \mathcal{V}_\infty\} = O_p(1).$$

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

**Lemma 3** *Under Assumptions 1, 2 and 3, we have*

$$\left\|\hat{f}-f\right\| = O_p\left(h^r + N^{-1/2}h^{-(p+q)/2}\ln(N)\right).$$

*The same result can be applied to $\hat{\mathrm{f}}$ with $H$ replacing $h$ and $p/2$ replacing $(p+q)/2$.*

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

$$Nh^{(p+q)/2}\left\|\hat{f}-f\right\|^3 = O_p\left(Nh^{\frac{p+q}{2}}\left[h^{3r}+N^{-\frac{3}{2}}h^{-3(p+q)/2}\ln^3(N)\right]\right) = o_p(1). \tag{3.16}$$

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

From Lemma 2 and 3 we obtain:

**Lemma 4** *Under Assumptions 1, 2 and 3, we have for any df $G$,*

$$\begin{aligned}\Gamma\left(\hat{F},\hat{F},G\right) = \Gamma(F,F,G) + \textstyle\int_v\int_w\left[\int_y \alpha(y,w,v)\hat{f}(y,w,v)\,dy - \int_y A(y,w)\hat{f}(y,w)\,dy\right]^2 \\ .\, a(w,v)\,d\tilde{F}(w,v) \;+\; O_p\left(\left\|\hat{f}-f\right\|^3 + \|\hat{\mathrm{f}}-f\|^3\right)\end{aligned} \tag{3.17}$$

Apply Lemmas 3 and 4 to $\Gamma(\hat{F},\hat{\mathrm{F}},F)$ and $\Gamma(\hat{F},\hat{\mathrm{F}},\tilde{F}) = \tilde{\Gamma}$ to obtain

$$\tilde{\Gamma} = \Gamma(\hat{F},\hat{\mathrm{F}},F) + \Delta_N + o_p(N^{-1}h^{-(p+q)/2}) \tag{3.18}$$

where

$$\begin{aligned}\Delta_N &= \textstyle\int_v\int_w\left\{\int_y \alpha(y,w,v)\hat{f}(y,w,v)dy\right. \\ &- \left.\textstyle\int_y A(y,w)\hat{\mathrm{f}}(y,w)dy\right\}^2 a(w,v) \\ &\quad d(\tilde{F}(w,v)-F(w,v))\end{aligned}$$

and

$$\Gamma(\hat{F},\hat{\mathrm{F}},F) = I_N + o_p(N^{-1}h^{-(p+q)/2}) \tag{3.19}$$

where

$$\begin{aligned}I_N &\equiv \textstyle\int_v\int_w\left\{\int_y \alpha(y,w,v)\hat{f}(y,w,v)dy\right. \\ &- \left.\textstyle\int_y A(y,w)\hat{\mathrm{f}}(y,w)dy\right\}^2 a(w,v)f(w,v)dwdv.\end{aligned}$$

Define, for $(w,v)\equiv x$

$$\begin{aligned}a_N(y_1,x_1,x_2) &\equiv a(x_2)\textstyle\int_t\{\alpha(t,x_2)K_h^{(1)}(t-y_1)K_h^{(p+q)}(x_2-x_1) \\ &- A(t,w_2)K_H^{(1)}(t-y_1)K_H^{(p)}(w_2-w_1)\}dt, \\ a_N^*(y_1,x_1,x) &\equiv a_N(y_1,x_1,x) - Ea_N(y_1,x_1,x)\end{aligned}$$

and write

$$\begin{aligned} I_N &= N^{-2}\sum_{j,k}\int a_N(y_j,x_j;x)a_N(y_k,x_k;x)dF(x) \\ &= N^{-2}\{\sum_{j\neq k}\int a_N^*(y_j,x_j;x)a_N^*(y_k,x_k,x)dF(n) \\ &+ \sum_j\int [a_N]^2(y_j,x_j;x)dF(x) \\ &+ 2(N-1)\sum_j\int a_N(y_j,x_j;x)Ea_N^*(y_1,x_1,x)dF(x) \\ &+ N(N-1)\int E^2a_N(y_1,x_1,x)dF(x)\} \\ &\equiv I_{N1}+I_{N2}+I_{N3}+I_{N4}. \end{aligned} \tag{3.20}$$

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. Summarizing, we have:

**Lemma 5** *Under Assumptions 1, 2 and 3*

$$Nh^{(p+q)/2}[I_N - N^{-1}h^{-(p+q)}\gamma_{12} - N^{-1}H^{-p}(\gamma_{22}+\gamma_{32})] \to N(0,\sigma_{11}^2) \tag{3.21}$$

*where* $\sigma_{11}$ *and the* $\gamma_{ij}$ *are given above.*

The proof of this lemma will utilize:

**Lemma 6** *(Hall (1984)) Let* $\{z_i|i=1,\ldots,N\}$ *be an i.i.d. sequence. Suppose that the U-statistic* $U_N \equiv \sum_{1\le i<j\le N}\tilde{P}_N(z_i,z_j)$ *with symmetric variable function* $\tilde{P}_N$ *is centered (i.e.,* $E\left[\tilde{P}_N(z_1,z_2)\right]=0$*) and degenerate (i.e.,* $E\left[\tilde{P}_N(z_1,z_2)|z_1\right]=0$ *almost surely for all* $z_1$*). Let*

$$\sigma_N^2 \equiv E\left[\tilde{P}_N(z_1,z_2)^2\right],\ \tilde{\Pi}_N(z_1,z_2)\equiv E_{z_i}\left[\tilde{P}_N(z_i,z_1)\tilde{P}_N(z_i,z_2)\right].$$

*Then if*

$$\lim_{N\to\infty}\frac{E[\tilde{\Pi}_N(z_1,z_2)^2]+N^{-1}E[\tilde{P}_N(z_1,z_2)^4]}{(E[\tilde{P}_N(z_1,z_2)^2])^2}=0 \tag{3.22}$$

*we have that as* $N\to\infty$

$$\frac{2^{1/2}}{N\sigma_N}U_N \to \mathcal{N}(\prime,\infty).$$

By Lemma 3 and (3.16), the result (3.21) is also valid if we replace $I_N$ by $\Gamma(\hat{F},\hat{\mathrm{F}},F)$ there. The theorem follows in view of (3.18) and:

**Lemma 7** *Given Assumptions 1, 2 and 3, we have that*

$$\begin{aligned} \Delta_N &= O_p(N^{-3}(h^{-3(p+q)}+H^{-3p})+N^{-1}(H^{2r}+h^{2r})) \\ &= o_p(N^{-2}h^{-(p+q)}). \end{aligned} \tag{3.23}$$

### 3.4 Intuition for Theorem and Bandwidth Choice

There is really one key feature that drives the structure of the results above, as well as the results of the next section. 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 $\hat{m}(w,v)$, the general regression. The place this arises in the proof is in the decomposition (3.20) of $I_N$. There $I_{N1}$ represents the variation of the unrestricted regression (both $w$ and $v$ present), and $I_{N3}$ represents the variation for the restricted regression (only $w$). Since the presence of both $w$ and $v$ makes for a slower rate of convergence for $\hat{m}(w,v)$, the rate of convergence of $I_{N1}$ is slower (see (A.14)) than that of $I_{N3}$ (see (A.12)). Therefore the asymptotic distribution of $I_N$ is driven by (A.14) and (A.12) does not contribute any term. Likewise, $I_{N2}$ (see (A.15)) has no impact. 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. We explore several variations in the next section.

It is worth noting that we obtain this 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; if

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

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

$$h = \arg\min \int_w \int_v E\left[\left(\hat{m}(w,v) - m(w,v)\right)^2\right] dw dv$$

– for instance, rates exhibited by bandwidths chosen by cross-validation.[7] In that case, the terms $I_{N1} - (1/Nh^{p+q})\,\gamma_{12} = O_p\left(N^{-1}h^{-(p+q)/2}\right)$ and $I_{N4} = O_p\left(N^{-1/2}h^r\right)$ would be of the same order, and $\tilde{\Gamma}$ or equivalently $I_N$ would be driven by an additional term which would lead to an additional component of variance whose estimation would require estimation of $r$-th order derivatives of the regression and density to construct a test statistic.

It would also appear – see our computations in the appendix – that bias estimation might be avoided by replacing $\hat{m}(w_i, v_i)$ and $\hat{M}(w_i)$ in $\tilde{\Gamma}$ by, $m_{(-i)}(w_i, v_i)$, $\hat{M}_{(-i)}(w_1)$ where $(-i)$ indicates that the function is computed using all observations *other than* $(y_i, x_i)$. Of course, the statistic is more complicated since a separate computation needs to be done for each summand. We conjecture that suitable versions of the bootstrap or other resampling methods will enable us to avoid the bias estimation problem, while retaining the computational simplicity of the test statistic $\tilde{\Gamma}$, i.e., without requiring that a bias-correction term be included in the definition of the test statistic. The main message from (3.24) is that we should undersmooth the unrestricted model $(h)$ compared to the indications we get from cross-validation.

### 3.5 Intuition for the Asymptotic Normality of the Test Statistic

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 (3.22) — 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 $\lambda_{jN}$, $j = 1, 2, ..., \infty$, of the linear operator $\Psi_N$ on $L_2$ defined by

$$\psi \in L_2 \longmapsto (\Psi_N \psi)(z) \equiv E\left[\tilde{P}_N(z_i, z)\,\psi(z)\right]$$

[7]Our results apply to the case of a non-stochastic 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.

have the asymptotic negligibility property that

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

Rather than attempt to check this condition directly (the eigenvalues are not known explicitly), we relied on the sufficient characterization (3.22).

### 3.6 Consistency and Local Power Properties

In this section, we first note that the test we proposed is consistent, i.e., it tends 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 $N \to \infty$.

**Proposition 1** *Under Assumptions 1, 2 and 3, the test based on the statistic (3.14) is consistent for $F$ such that $\Gamma(F, F, F) > 0$. Note that this is equivalent to $(m(w,v) - M(w))a(w,v) \neq 0$ (a.e.).*

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 subscripts $N$ to indicate dependence on $f_N$:

$$H_{1N} : \sup\{|m_N(w,v) - M_N(w) - \epsilon_N \Delta(w,v)| : (w,v) \in S\} = o(\epsilon_N), \tag{3.25}$$

$$\|f_N - f\| = o(N^{-1}h^{-(p+q)/2}) \tag{3.26}$$

where $\|\cdot\|$ is defined in section 3.1. Further, suppose

$$\delta_2 \equiv \int\int \Delta^2(w,v) f(w,v) a(w,v) dw dv < \infty$$

and

$$\int \Delta(w,v) f(w,v) dv = 0 \tag{3.27}$$

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

**Proposition 2** *Under Assumptions 1, 2 and 3, suppose that the local alternative (3.25) converges to the null in the sense that*

$$\epsilon_N = 1/\left(N^{1/2}h^{(p+q)/4}\right).$$

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

$$\Pr(\hat{\tau} \geq z_\alpha | H_{1N}) \to 1 - \Phi(z_\alpha - \delta_2/\sigma_{11})$$

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

## 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 is 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 indicate 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 rewrite the predictor vector $x$ as its (invertible) transformation $(w, v) = \left(x'\theta, v\right)$. A single index model of the regression is then $m(x) = M^*\left(x'\theta\right)$. 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 $w = \left(x'\theta, w_{-1}\right)$, where again the predictor $x$ is an invertible transformation of $\left(x'\theta, w_{-1}, v\right)$. In these examples, if $\theta$ is known, then our previous results apply for testing the null hypothesis that an index model is valid. When $\theta$ 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 the $\theta$.[8]

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

$$H_0 : \Pr\left[m(w, v) = M(w(x, \theta))\right] = 1 \tag{4.1}$$

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

**Assumption 4** *The map $W$ and its Jacobian $J(x, \theta)$ are continuous as functions of $x$ and $\theta$. 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 $\theta$ is used, or that $\hat{w} \equiv w\left(x, \hat{\theta}\right)$ is used to 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. Consider,

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

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

We give in the Appendix heuristics for:

**Corollary 1** *Under Assumptions 4 and 5 and $\delta > 6(p+q)$ the conclusion of Theorem 1 can be applied to*

$$\tilde{\Gamma} \equiv \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{m}(w(x_i,\hat{\theta}),v(x_i,\hat{\theta})) - \hat{M}(w(x_i,\hat{\theta}))\right\}^2 a_i. \tag{4.2}$$

## 4.2 Parametric and Semiparametric Models

In the above section, we have proposed an interesting variation to the basic testing result, namely permitting the use of estimated parameters in the restricted set of regressors. Much 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 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

$$H_0 : \Pr\left[m\left(w,v\right) = M_\theta\left(w\right)\right] = 1 \tag{4.3}$$

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

**Assumption 6** *$M_\theta\left(w\right)$ is differentiable in $\theta$, with derivative uniformly bounded for $w \in S$, and $\theta$ in a neighborhood of the true parameter value in $\Theta$.*

With regard to single index models, we could, for instance, test the null hypothesis that the egression is a linear model in the predictor variables, or that $E[y|w,v] = w'\theta$. 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 $M_{\hat{\theta}}(w)$ has a lower order effect.

We shall establish a result already appearing in Härdle and Mammen (1993):

**Corollary 2** *Under the additional Assumptions 5-6, Theorem 1 can be applied, with $\gamma_{22}$ and $\gamma_{32}$ replaced by 0, to*

$$\tilde{\Gamma} = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left\{\hat{m}\left(w_i,v_i\right) - M_{\hat{\theta}}\left(w_i\right)\right\}^2 a_i$$

*under the null (4.3).*

Notice that the validity of Theorem 1 implies that the variance of $\hat{\theta}$ does not affect the limiting distribution of $\tilde{\Gamma}$. Further, the logic applies when the restricted model involves lower dimensional estimated functions as well as estimated parameters. For example, our null model could be semiparametric

$$H_0 : \Pr\left[E\left[y|w,v\right] = M\left(w\right) + M_\theta\left(v\right)\right] = 1 \tag{4.4}$$

where the function $M_\theta(\cdot)$ is known, but $M(\cdot)$ and the finite-dimensional parameter vector $\theta$ are unknown (see Robinson (1988) for an estimation strategy for this model with $M_\theta(v) \equiv v'\theta$). Specifically, the model can be written as $E(y - M_\theta(v) \mid w, v) = E(y - M_\theta(v) \mid w)$. Thus the appropriate test statistic given an estimate $\hat{\theta}$ satisfying Assumption 5 is

$$\tilde{\Gamma}^* \equiv \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{m}^*(w_i, v_i, \hat{\theta}) - \hat{M}^*(w_i, \hat{\theta})\right\}^2$$

where $\hat{m}^*, \hat{M}^*$ are $\hat{m}, \hat{M}$ applied to the observations $(y_i - M_{\hat{\theta}}(v_i), w_i, v_i)$. As in Corollary 1, we give in the Appendix heuristics for:

**Corollary 3** *Under the additional Assumptions 5-6, Theorem 1 can be applied without modification to* $\tilde{\Gamma}^*$ *under the null (4.4).*

Again, the estimation of $\hat{\theta}$ should not affect the limiting distribution of $\tilde{\Gamma}$, while, under our bandwidth choices, the nonparametric estimation of $M(\cdot)$ gives rise to the bias adjustment terms $\gamma_{22}$ and $\gamma_{32}$. 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 $f(w, v)\, dF(y|w, v)$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?

It is heuristically clear that the moment calculations we have used in our derivation continue to hold for $(w_i, v_i, y_i)$ stationary ergodic and at least formally that in a suitable mixing context remainders should still be of smaller order. Results such as $\|\hat{f} - f\|_\infty = O(h^r + N^{-1/2}h^{-(p+q)/2}\log N)$, (Gyorfi et al. (1989)), and for $U$ statistics the appropriate extension of Lemma 6 (Khashimov (1992)), are also available. The technical issues are resolvable under suitable mixing conditions (which are of course not verifiable!). That is we conjecture:

| | | $h_0$=0.45 | $h_0$=0.50 | $h_0$=0.55 |
|---|---|---|---|---|
| $N$=500 | 5% | 5.5 | 7.0 | 7.8 |
| | 10% | 8.4 | 10.6 | 8.4 |
| | Stan. Dev. ($\hat{\tau}$) | 0.98 | 1.01 | 1.02 |
| $N$=1000 | 5% | 5.7 | 5.7 | 6.1 |
| | 10% | 8.9 | 8.5 | 9.6 |
| | Stan. Dev. ($\hat{\tau}$) | 0.98 | 0.99 | 0.97 |
| $N$=5000 | 5% | 6.3 | 6.6 | 6.7 |
| | 10% | 10.0 | 10.3 | 10.7 |
| | Stan. Dev. ($\hat{\tau}$) | 1.00 | 0.98 | 1.00 |
| $N$=10,000 | 5% | 4.8 | 6.3 | 6.4 |
| | 10% | 6.6 | 9.8 | 9.9 |
| | Stan. Dev. ($\hat{\tau}$) | 0.99 | 0.98 | 0.98 |
| $N$=15,000 | 5% | 6.0 | 5.5 | 5.9 |
| | 10% | 9.7 | 10.7 | 9.5 |
| | Stan. Dev. ($\hat{\tau}$) | 0.98 | 1.01 | 1.01 |

Table 1: Monte Carlo Results: One Dimension

**Proposition 3** *If Assumption 1 is replaced by suitable smoothness and mixing conditions, Theorem 1, Propositions 1 and 2 and Corollaries 1, 2 and 3 continue to hold.*

We accordingly apply the test statistic to an example in which at best we can think of $y_i$ given $(x_i)$ as consisting of independent components, and we have a semi-parametric structure as in (4.4).

# 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-dimensional and two-dimensional testing situations. We begin with a one-dimensional study of functional form, where the true model is $E[y|w] = w\theta$ with $\theta = 1$, for $w$ distributed as $\mathcal{N}(0,1)$. In particular, we constructed samples with

$$y = w\theta + \sigma(w)\varepsilon \tag{5.1}$$

where $\varepsilon$ is distributed as $\mathcal{N}(0,1)$ and $\sigma^2(w) = .5625\exp(-w^2)$. With reference to Corollary 2, we have that $p = 1$, $q = 0$. We 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 .45, .50 and .55, and $a$ is the indicator function of the interval $S = \{w \in R/ -2 \le w \le 2\}$. To estimate the conditional variance, we computed the bandwidth as $.25N^{-1/\delta}$. 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 ($z_{.05} = 1.64$ and $z_{.10} = 1.28$), and the standard deviation of the standardized test statistic $\hat{\tau}$ (which is 1 asymptotically).

Table 1 shows a fairly close correspondence between the finite sample performance of the test statistic and the asymptotic results for the one-dimensional design. There is some

| | | NPGEN - NPREST | NPGEN - PARAM |
|---|---|---|---|
| $n$=500 | 5% | 8.6 | 8.3 |
| | 10% | 15.4 | 13.3 |
| | Stan. Dev. ($\hat{\tau}$) | 0.93 | 0.97 |
| $n$=1000 | 5% | 7.9 | 6.9 |
| | 10% | 13.7 | 11.2 |
| | Stan. Dev. ($\hat{\tau}$) | 0.94 | 0.99 |
| $n$=5000 | 5% | 7.4 | 6.3 |
| | 10% | 12.9 | 11.0 |
| | Stan. Dev. ($\hat{\tau}$) | 0.95 | 0.98 |

Table 2: Monte Carlo Results: Two Dimensions

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 sample using the same model (5.1) as above, and test for the presence of an additional regressor $v$, which is distributed as $\mathcal{N}(0,1)$, independently of $w$ and $\varepsilon$. We study the performance of the test statistic in two settings: first a comparison of nonparametric estimates of the general nonparametric regression $E(y|w,v)$ and the restricted nonparametric regression $E(y|w)$ ("NPGEN - NPREST" in Table 2) and then a comparison of nonparametric estimates of $E(y|w,v)$ 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 $h=h_0N^{-1/\delta}$, $H=H_0N^{-1/\Delta}$ , where $\delta=4.75$, $\Delta=4.25$, $h_0=0.65$, $H_0=0.50$. The weighting function $a$ is the indicator function of the disk $S=\left\{(w,v)\in R^2/\ \sqrt{w^2+v^2}\leq 2\right\}$ in $R^2$. To estimate conditional variances, we use the bandwidth $.25N^{-1/\Delta}$ for the one-dimensional regression and $.55N^{-1/\delta}$ for the two dimensional regression. 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 that 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-Scholes (1973) option pricing formula. 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 $N = 9,005$ observations on daily S&P 500 index options obtained from the Chicago Board Options Exchange for the year 1993. The S&P 500 index option market is extremely liquid (as one of the most active among options markets in the United States), the options are European, and we have chosen options with open interest above the annual mean open interest level (i.e. we focus on actively traded options) with maturities ranging from 0 to 12 months. For further information on the basic data set, see Aït-Sahalia and Lo (1997).

We denote the call option price as $C$, its strike price as $X$, its time-to-expiration as $T$, the S&P500 futures price for delivery at expiration as $F$, the risk-free interest rate between the current date and $T$ as $r$.[9] The Black-Scholes option pricing formula is

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

where $\Phi(\cdot)$ is the normal c.d.f., 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

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

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), the fact that both $F$ and $C$ are subject to a bid-ask spread, and the fact that settlement prices are computed as representative of the last trading prices of the day. We summarize these potential sources of noise as an additive residual in implied volatilities: namely we pose the model

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

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

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

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 out-of-the-money calls ($X/F > 1$).[10] The pattern is strongly nonlinear since as a function of $X/F$, the implied volatility $s$ tends to decrease between $X/F = 0.85$ and $X/F = 1$ (at-the-money), and then flattens out between 1 and 1.1. 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 $E[s|X/F,T] = m(X/F,T)$. 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

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

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.

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,[11] and SEMIPAR is estimated by using Robinson's (1988) difference method (with the same smoothing parameters as for NPREST).[12]

[10] 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 (1997) for a discussion.

[11] 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 $h_0 = .65$, $\delta = 4.75$, and for NPREST, we set $H = H_0 N^{-1/\Delta}$ for $H_0 = .50$, $\Delta = 4.25$, with $N = 9,005$ in both cases, giving $h = .14$ and $H = .081$. The weighting function $a$ is the indicator function of the set $S = [-1.6, 1.6] \times [-1.0, 0.6]$ in $R^2$. Finally, the bandwidth for estimating conditional variance terms are given above with $h_0 = .55$ and $H_0 = .25$.

[12] We estimate the univariate regressions $m_s \equiv E[s|X/F]$, $m_T = E[T|X/F]$ and $m_{T^2} = E[T^2 \mid X/F]$ by kernel regression, estimate $\theta_4$ and $\theta_5$ 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 then form $\hat{g} = \hat{m}_s - \hat{\theta}_4 \hat{m}_T - \hat{\theta}_5 \hat{m}_{T^2}$.

| | $\hat{\tau}$ | p-value |
|---|---|---|
| NPGEN-BS | 2511.6 | 0.00 |
| NPGEN-PARAM | 146.2 | 0.00 |
| NPGEN-NPREST | 139.6 | 0.00 |
| NPGEN-SEMIPAR | .60 | 0.27 |

Table 3: Testing Results: Implied Volatility

We can summarize the parametric results for fitting these models as:

$$\text{BS:} \quad \hat{m}(X/F,T) = \underset{(464)}{9.69}$$

$$\text{PARAM:} \quad \hat{m}(X/F,T) = \underset{(483)}{9.15} \quad \underset{(-111)}{-1.54 \cdot X/F} \quad \underset{(40.6)}{+.54 \cdot (X/F)^2}$$

$$\text{SEMIPAR:} \quad \hat{m}(X/F,T) = \hat{g}(X/F) \quad \underset{(-20.3)}{-.37 \cdot T} \quad \underset{(23.4)}{+.202 \cdot T^2}$$

where t-statistics are in parentheses. The $R^2$ value for SEMIPAR is .98, indicating a reasonable goodness-of-fit to the implied volatility data.

We 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 3. The critical values appropriate for $\hat{\tau}$ are 1.64 (5%) and 1.28 (10%). The overall bottom line is that all specializations of the model are rejected except for SEMIPAR. The extremely 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.

In any case, we conclude that a model permitting a flexible "volatility smile" as well as an additive time effect is a statistically adequate depiction of the volatility data. To illustrate the estimates of the model SEMIPAR, we also include a graph of the estimated function $\hat{g}(X/F)$ in Figure 1 (where $X/F$ values are in standardized form). The downward slope on the left is evident, and there is a substantially less pronounced upturn on the right than would occur with the quadratic model.

# 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 the 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 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 non-zero under any fixed alternative suggest further study of the practical performance of test statistics of this kind.

# A Appendix

**Proof of Lemma 1:** Write

$$(\hat{\sigma}^4(w,v)/\hat{f}(w,v))a(w,v) = \frac{\sigma^4(w,v)}{E\hat{f}(w,v)} + R(w,v)$$

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

$$\begin{array}{rcl} |R(w,v)| & \leq & 2\max\frac{\{\hat{\sigma}^2(w,v),\sigma^2(w,v)\}}{f(w,v)}|\hat{\sigma}^2(w,v)-\sigma^2(w,v)| \\ & + & \frac{\sigma^4(w,v)}{f(w,v)}|\hat{f}(w,v)-f(w,v)| \end{array} \tag{A.1}$$

It is easy to verify that

$$\frac{2C_{11}}{N}\sum_{i=1}^{N}\frac{\sigma^4(w_i,v_i)a_i}{f(w_i,v_i)} = \sigma_{11}^2 + O_p(N^{-1/2}). \tag{A.2}$$

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

$$\sup_S |\hat{\sigma}^2(w,v) - \sigma^2(w,v)| = O_p(a_N) \tag{A.3}$$

where

$$a_N = N^{-1/2}\log N h^{-(p+q)/2} + h^r$$

and also use Lemma 3. In fact, the same type of argument will also yield

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

by Assumption 3. The same type of calculation yields the other assertions of Lemma 1.

To prove (A.3) write

$$\begin{array}{l} \hat{\sigma}^2(w,v) = \sum_{i=1}^N\{\mathcal{K}\left(\frac{w-w_i}{h},\frac{v-v_i}{h}\right)(y_i-m(w,v))^2 \\ \quad /\sum_{i=1}^N\mathcal{K}\left(\frac{w-w_i}{h},\frac{v-v_i}{h}\right)\} + (\hat{m}(w,v)-m(w,v))^2. \end{array} \tag{A.4}$$

Then, (A.3) follows from

$$\sup_S \frac{\left|\frac{1}{N}\sum_{i=1}^N h^{-(p+q)}\mathcal{K}\left(\frac{w-w_i}{h},\frac{v-v_i}{h}\right)(y_i-m(w_i,v_i))^2 - \sigma^2\hat{f}(w,v)\right|}{\hat{f}(w,v)} = O_p(a_N) \tag{A.5}$$

and

$$\sup_S \frac{\left|\frac{1}{N}\sum_{i=1}^N(m(w,v)-m(w_i,v_i))h^{-(p+q)}\mathcal{K}\left(\frac{w-w_i}{h},\frac{v-v_i}{h}\right)\right|}{\hat{f}(w,v)} = O_p(a_N) \tag{A.6}$$

$$\sup_S \frac{\left|\frac{1}{N}\sum_{i=1}^N(y_i-m(w_i,v_i))h^{-(p+q)}\mathcal{K}\left(\frac{w-w_i}{h},\frac{v-v_i}{h}\right)\right|}{\hat{f}(w,v)} = O_p(a_N). \tag{A.7}$$

By Lemma 3 again we can replace $\hat{f}$ by $f$ in (A.5)–(A.7). Consider the expression equivalent to (A.5 ),

$$\sup_S\left|\frac{1}{N}\sum_{i=1}^N\frac{h^{-(p+q)}\mathcal{K}\left(\frac{w-w_i}{h},\frac{v-v_i}{h}\right)}{f(w,v)}(y_i-m(w_i,v_i))^2-\sigma^2(w_i,v_i))\right| \tag{A.8}$$

$$+\sup_S\left|\frac{\hat{f}(w,v)}{f(w,v)}-1\right|\sigma^2(w,v). \tag{A.9}$$

The second term is $O_p(a_N)$ by Lemma 3. The first is $O_p(N^{-1/2}\sqrt{\log N})$ by Ossiander's bracketing theorem (Ossiander (1987)) and our moment conditions (A.2)–(A.3). The other terms are dealt with in the same way but somewhat more easily.

**Proof of Lemma 2:** For $t \in [0,1]$ let

$$\begin{aligned} m(w,v,t) &\equiv m(w,v,f+tg_1) \\ M(w,t) &\equiv M(w,f+tg_2) \end{aligned}$$

where $g_1, g_2$ the derivatives of $G_1, G_2$ are such that $tG_1, tG_2 \in \mathcal{V}$ for all $|t|$ sufficiently small and strictly positive. Let,

$$\Phi(t) \equiv \int (M(w,t) - m(w,v,t))^2 dF_2(w,v).$$

Then $\Phi$ is infinitely differentiable and $\Phi(0) = \Gamma(F,F,F_2)$. Expand to 3 terms using Taylor's formula to get

$$\Phi(t) = \Phi(0) + t\Phi'(0) + \frac{t^2\Phi''(0)}{2} + \frac{t^3\Phi'''(\theta)}{6} \tag{A.10}$$

where $0 \le \theta(t) \le |t|$. Evidently

$$\Phi'(0) = 2\int (M(w) - m(w,v)) \left(\frac{\partial M(w,0)}{\partial t} - \frac{\partial m}{\partial t}(w,v,0)\right) dF_2(w,v)$$

$$\begin{aligned} \Phi''(0) &= 2\int \Big\{ \left(\tfrac{\partial M}{\partial t}(w,0) - \tfrac{\partial m}{\partial t}(w,v,0)\right)^2 \\ &+ (M(w) - m(w,v))\tfrac{\partial^2}{\partial t^2}(M(w,0) - m(w,v,0))\Big\} \\ & dF_2(w,v) \\ \Phi'''(t) &= 2\Big(\int \Big\{3\left(\tfrac{\partial}{\partial t}(M(w,t) - m(w,v,t)\right)\tfrac{\partial^2}{\partial t^2}(M(w,t) - m(w,v,t)) \\ &+ (M(w,t) - m(w,v,t))\tfrac{\partial^3}{\partial t^3}(M(w,t) - m(w,v,t))\Big\} dF_2(w,v) \end{aligned}$$

If $H_0$ is true $\Phi(0) = \Phi'(0) = 0$ and

$$\begin{aligned} \Phi''(0) &= 2\int \left(\tfrac{\partial M}{\partial t}(w,0) - \tfrac{\partial m}{\partial t}(w,v,0)\right)^2 dF_2(w,v) \\ &= 2\int_{w,v} [\int(\alpha(y,w,v)g_1(y,w,v) \\ &- A(y,w)g_2(y,w))dy]^2 dF_2(w,v) \end{aligned} \tag{A.11}$$

If $tG_j \in \mathcal{V}$, $j = 1,2$ it is easy to check that the integrand in $\Phi'''$ is uniformly bounded by a universal constant times $\|g_1\|^3 + \|g_2\|^3$. For instance,

$$\begin{aligned} \left|\tfrac{\partial M}{\partial t}(w,t)\right| &\le |g_2|(w)\left(\int(y(f(y,w) + tg_2(y,w))dy\right) \\ & /(f - |tg_2|)^2(w)) + \int yg_2(y,w)dy/(f - |tg_2|(w)) \\ &\le \tfrac{4}{b^2}(\|g_2\|_\infty(\|yg_2\|_2 + \|aM\|_\infty) + 2\|yg_2\|_2) \end{aligned}$$

The lemma follows.

**Proof of Lemma 3:** The bounds on the $L_2$ and $L_\infty$ deviations of the kernel density estimator, $\left\|\hat{f}(y,w,v) - f(y,w,v)\right\|_2$ and $\left\|\hat{f}(y,w,v) - f(y,w,v)\right\|_\infty$, are classical results: see e.g., Nadaraya (1983, Chapter 2) and Stone (1983, Lemma 2 and Lemma 8). The same bounds apply to $\left\|y\left(\hat{f}(y,w,v)-f(y,w,v)\right)\right\|_2$. Under 2 we have $h^p \gg H^p \gg h^{p+q}$ (so both the asymptotic bias and variance of $\hat{f}(w)$ are smaller than those of $\hat{f}(w,v)$) and thus all the norms involving $\hat{f}(w)-f(w)$ are strictly smaller than those involving $\hat{f}(w,v)-f(w,v)$. The bounds follow for $\left\|\hat{f}-f\right\|$.

**Proof of Lemma 4**: Apply Lemma 2 with $G=\hat{F}-F$, or $\hat{\mathrm{F}}-F$. This can be done since by Lemma 2,

$$P\left[\|\hat{g}\|_\infty \geq \frac{b}{2}\right] \to 0$$

so that

$$P[\hat{G} \in \mathcal{V}] \to \infty.$$

Lemma 4 follows.

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

$$\begin{aligned} Ea_N(y_1,x_1,x) &= \textstyle\int f(y_1,x_1)\{\int_y \alpha(y,x_1)K_h^{(1)}(y-y_1)dyK_h^{(p+q)}(x-x_1) \\ &- \textstyle\int_y A(y,w_1)K_h^{(p)}(y-y_1)dyK_H^{(p)}(w-w_1)\}dy_1dx. \end{aligned}$$

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

$$\begin{aligned} Ea_N(y_1,x_1,x) &= a(x)[\textstyle\int_{y_1}\int_{u,s} f(y_1,x-hs)\alpha(y_1+hu,x-hs)K_1^{(1)}(u)K_1^{(p+q)}(s)dudsdy_1 \\ &- a(x)[\textstyle\int_{y_1}\int_{u,\tau} f(y_1,x-Hs)A(y_1+Hu,w-H\tau)K_1^{(1)}(u)K_1^{(p)}(\tau)dud\tau]dy_1 \\ &= a(x)\textstyle\int f(y_1,x)(\alpha(y_1,x)-A(y_1,w))dy_1+O(h^r)+O(H^r) \end{aligned} \tag{A.12}$$

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

$$E|a_N(y_1,x_1,x)|=O(1) \tag{A.13}$$

by the same argument. Similarly

$$\begin{aligned} \textstyle\int Ea_N^2(y_1,x_1,x)dF(x) &= \textstyle\int[(\int \alpha(y,x)K_h^{(1)}(y-y_1)K_h^{(p+q)}(x-x_1)dy)^2 f(x)f(y_1,x_1)dx \\ &- 2\textstyle\int(\int \alpha(y,x)K_h^{(1)}(y-y_1)dy) \\ &\times (\textstyle\int A(y_2,w)K_H^{(1)}(y_2-y_1)dy_2)K_h^{(p+q)}(x-x_1)K_H^{(p)}(w-w_1)f(x)dx \\ &+ \textstyle\int(\int A(y,w)K_H^{(1)}(y-y_1)dy)^2[K_H^{(p)}]^2(w-w_1)f(w)dw] \\ &\times f(y_1,x_1)dx_1dy_1. \end{aligned} \tag{A.14}$$

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

$$\begin{aligned} \textstyle\int Ea_N^2(y_1,x_1;x)dF(x) &= \gamma_{12}h^{-(p+q)}(1+O(h^{2r})) \\ &+ \gamma_{22}h^{-p}(1+O(h^{2r}))+\gamma_{32}H^{-p}(1+O(H^{2r})). \end{aligned} \tag{A.15}$$

Next, compute

$$E(\int a_N(y_1,x_1;x)E(a_N(y_2,x_2;x))dF(x))^2 \leq E(\int |a_N(y_1,x_1;x)|dF(x))^2 O(H^{2r}).$$

From (A.15) we conclude

$$E(\int a_N(y_1,x_1;x)Ea_N(y_2,x_2,x)dF(x))^2 = O(H^{2r}). \tag{A.16}$$

Finally a similar computation yields

$$E(\int a_N^*(y_1,x_1;x)a_N^*(y_2,x_2;x)dF(x))^2 = \sigma_{11}^2(1+o(1))h^{-(p+q)} \tag{A.17}$$

$$E(\int a_N^*(y_1,x_1;x)a_N^*(y_2,x_2;x)dF(x))^2 = O(h^{-2(p+q)}) \tag{A.18}$$

and

$$\begin{aligned} &E\{E(\int a_N^*(y_1,x_1,x)a_N^*(y_2,x_2,x)dF(x)\int a_N^*(y_1,x_1,x)a_N^*(y_3,x_3;x)dF(x) \mid z_1)\}^2 \\ &\quad = E(\int a_N^*(y_1,x_1,x)Ea_N^*(y_2,x_2,x)dF(x))^4 \\ &\quad = O(H^{4r}). \end{aligned} \tag{A.19}$$

From (A.12) we conclude that

$$I_{N4} = O(H^{2r}) = o(N^{-1}h^{-(p+q)/2}). \tag{A.20}$$

From (A.16)

$$\begin{aligned} E(I_{N3}^2) &\le 4E(\int a_N(y_2,x_2;x)Ea_N^*(y_1,x_1,x)dF(x))^2 \\ &= O(H^{2r}/N) = o(N^{-2}h^{-(p+q)}). \end{aligned} \tag{A.21}$$

From (A.15) and

$$\begin{aligned} N^{-1}E(\int a_N^2(y_1,x_1;x)dF(x))^2 &\le N^{-1}\int Ea_N^4(y_1,x_1,x)dF(x) \\ &= O(N^{-1}h^{-2(p+q)}) = o(N^{-1}h^{-(p+q)}) \end{aligned}$$

by arguing as for (A.15), we obtain

$$I_{N2} = N^{-1}\{\gamma_{12}h^{-(p+q)} + \gamma_{22}h^{-p} + \gamma_{32}H^{-p}\}(1+o(1)). \tag{A.22}$$

Finally (A.17)-(A.19) enable us to apply Lemma 6 and Lemma 5 follows.

**Proof of Lemma 7:** Write

$$\begin{aligned} \Delta_N &= N^{-3}\sum\{a_N(y_j,x_j;x_j)a_N(y_k,x_k,x_l) \\ &- E(a_N(y_j,x_j;x_l)a_N(y_k,x_k;x_l) \mid z_j,z_k) : l \notin \{j,k\} \\ &+ 2N^{-3}\textstyle\sum_{j\neq k} a_N(y_j,x_j;x_j)a_N(y_k,x_k;x_j) \\ &+ N^{-3}\textstyle\sum_j a_N^2(y_j,x_j;x_j) \\ &- N^{-3}\textstyle\sum_{j,k}\int a_N(y_j,x_j;x)a_N(y_k,x_k;x)dF(x) \end{aligned} \tag{A.23}$$

where we use the identity

$$E\{a_N(y_j,x_j;x_l)a_N(y_k,x_k;x_l) \mid j,k\} = \int a_N(y_j,x_j;x)a_N(y_k,x_k;x)dF(x)$$

if $l \notin \{j,k\}$. Call these terms $\Delta_{Nj}$, $1 \le j \le 4$. Evidently,

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

Further

$$\begin{aligned} E\Delta_{N3} &= N^{-2}Ea_N^2(y_1, x_1; x_1) \\ &= o(N^{-1}h^{-(p+q)/2}). \end{aligned} \tag{A.24}$$

Note that $N\Delta_{N2}$ is a $U$ statistic. From (A.12),

$$\begin{aligned} E\Delta_{N2} &= 2N^{-1}(1+o(1))O(h^{-(p+q)+r}) \\ &= o(N^{-1}h^{-(p+q)/2}). \end{aligned} \tag{A.25}$$

Then, by Hoeffding's inequality (Serfling (1980) p. 153, Lemma A)

$$\begin{aligned} \mathrm{Var}(\Delta_{N2}) &\leq N^3 Ea_N^2(y_1, x_1; x_1)a_N^2(y_2, x_2; x_1) \\ &= O(N^{-3}h^{-3(p+q)}) \\ &= o(N^{-2}h^{-(p+q)}) \end{aligned} \tag{A.26}$$

by arguing as for (A.15). Finally write

$$\begin{aligned} \Delta_{N1} &= \Delta_{N11} + \Delta_{N12} \\ \Delta_{N11} &= N^{-3}\sum\{W_{jkl}^* : l \notin \{j,k\}\} \\ \Delta_{N12} &= N^{-3}\sum\{E(W_{jkl} \mid z_l) : l \notin \{j,k\}\} \end{aligned}$$

where

$$\begin{aligned} W_{jkl} &= a_N(z_j; x_l)a_N(z_k; x_l) - E(a_N(z_j; x_l)a_N(z_k; x_l) \mid z_j, z_k) \\ W_{jkl}^* &= W_{jkl} - E(W_{jkl} \mid z_l) \\ &= a_N(z_j; x_l)a_N(z_k; x_l) - E(a_N(z_j; x_l)a_N(z_k; x_l) \mid x_l) \\ &- E(a_N(z_j; x_l)a_N(z_k; x_l) \mid z_j, z_k) \\ &+ E(a_N(z_j; x_l)a_N(z_k, x_l)). \end{aligned} \tag{A.27}$$

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

$$EW_{jkl} = 0. \tag{A.28}$$

Now, by (A.15)

$$\begin{aligned} E\Delta_{N12} &= O(N^{-2}Ea_N^2(z_1, x_2)) \\ &= O(N^{-2}h^{-(p+q)}) \\ &= o(N^{-1}h^{-(p+q)/2}) \end{aligned} \tag{A.29}$$

and

$$\begin{aligned} Var(\Delta_{N12}) &\leq N^{-1}E\{E^2(W_{123} \mid z_3) + N^{-1}E^2(W_{113} \mid z_3)\} \\ &= O(N^{-1}) \end{aligned}$$

by (A.13). Finally,

$$\begin{aligned} E\Delta_{N11} &= 0 \\ E\Delta_{N11}^2 &= N^{-6}\sum\{E(W_{jkl}^*)^2 : l \notin \{j,k\}\} \end{aligned}$$

since $EW_{jkl}^* W_{j'k'l'}^* = 0$ unless $\{j,k,l\} = \{j',k',l'\}$. Now, by arguing as for (A.15),

$$\begin{aligned} E(W_{jkl}^*) &\leq E(E^2(a_N(z_1; x_2 \mid x_2)) \\ &= O(h^{-2(p+q)}),\ j \neq k \\ E(W_{jjl}^*) &\leq Ea_N^4(z_1, x_2; x_2) \\ &= O(h^{-3(p+q)}). \end{aligned}$$

In any case

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

and the lemma is proved.

**Proof of Proposition 1:** Our analysis in Lemmas 2 and 7 shows that

$$\tilde{\Gamma} = \Gamma(F, F, F) + o_p(1).$$

Then, $\tilde{\tau} \xrightarrow{p} \infty$ if $\Gamma(F, F, F) > 0$ .

**Proof of Proposition 2:** Our assumptions are such that the preceding arguments remain valid even when the $z_{iN}$, $1 \le i \le N$, are a double array. The expansion (3.17) and subsequent lemmas continue to hold. Thus, we have that under $H_{1N}$,

$$(\hat{\tau} - \hat{\sigma}_{11}^{-1} N h^{(p+q)/2} \Gamma(F, F, \tilde{F})) \rightarrow \mathcal{N}(0, 1) .$$

Moreover, $\hat{\sigma}_{11} \xrightarrow{P_N} \sigma_{11}$ and

$$\begin{aligned} \Gamma(F_N, F_N, \tilde{F}) &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \{m_N(x_i) - M_N(w_i)\}^2 a^2(x_i) \\ &= E(m_N(x_1) - M_N(w_1))^2 a^2(x_1) + O_p(N^{-1/2}) \\ &= \delta_2 N^{-1} h^{-(p+q)/2} + o_p(N^{-1} h^{-(p+q)/2}). \end{aligned} \tag{A.30}$$

The proposition follows.

**Heuristics for Corollary 1:** Let

$$\hat{\Gamma}(\theta) \equiv \frac{1}{N} \sum_{i=1}^{N} \left\{ \hat{m}\left(w(x_i, \theta), v(x_i, \theta)\right) - \hat{M}(w(x_1, \theta)) \right\}^2 a_i \tag{A.31}$$

Then, Theorem 1 applies to $\tilde{\Gamma}(\theta_0)$ where $\theta_0$ is the true value of $\theta$. Consider the statistic we intend to use, $\tilde{\Gamma}(\hat{\theta})$:

$$\begin{aligned} \tilde{\Gamma}(\hat{\theta}) &= \tilde{\Gamma}(\theta_0) + \tilde{\Gamma}'(\theta_0)(\hat{\theta} - \theta_0) \\ &+ \tfrac{1}{2}\tilde{\Gamma}''(\theta_0)(\hat{\theta} - \theta_0, \hat{\theta} - \theta_0) \\ &+ \tfrac{1}{6}\tilde{\Gamma}'''(\tilde{\theta})(\hat{\theta} - \theta_0, \hat{\theta} - \theta_0, \hat{\theta} - \theta_0) \end{aligned} \tag{A.32}$$

where $\tilde{\Gamma}'$ etc. are differentials with respect to $\theta$, and $\tilde{\theta}$ is between $\theta_0$ and $\hat{\theta}$. Now, if without loss of generality, $w(x, \theta_0) = w$, $v(x, \theta_0) = v$ we can write $\tilde{\Gamma}(\theta_0) = V\left(\hat{f}, \hat{\mathrm{f}}, \hat{f}', \hat{\mathrm{f}}'\right)$ where

$$V(f_1, f_2, f_1', f_2') = 2 \int a(x)(m(x, f_1) - M(w, f_2))(\dot{M}(w, x, f_1, f_1') - \dot{m}(w, x, f_2, f_2'))d\tilde{F}(x)$$

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

$$\begin{aligned} \dot{m}(w, x, f_2, f_2') &= \textstyle\int y f_2'(y, x) dy \dot{x}(\theta_0) \;/\; \int f_2(y, x) dy \\ &- \left(\textstyle\int y f_2(y, x) dy \int f_2'(y, x) dy\right) \dot{x}(\theta_0) \;/ (\textstyle\int f_2(x, y) dy)^2 \end{aligned}$$

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

$$V\left(\hat{f}, \hat{\mathrm{f}}, \hat{f}', \hat{\mathrm{f}}'\right) - V\left(\hat{f}, \hat{\mathrm{f}}, f', \mathrm{f}'\right) = O_p(N^{-1} h^{-(p+q)/2}) \tag{A.33}$$

by analogy with Lemma 5 since

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

but $(\hat{\mathrm{f}}' - f') = O_p(N^{-1/2}\, h^{-3(p+q)/2})$. On the other hand it is to be expected that

$$V\left(\hat{f}, \hat{\mathrm{f}}, f', \mathrm{f}'\right) = O_p(N^{-1/2}). \tag{A.34}$$

Combining (A.33) and (A.34) and $\hat{\theta} - \theta_0 = O_p(N^{-1/2})$ we conclude that the first term in (A.32) is $O_p(N^{-1} + N^{-3/2}h^{-3/2(p+q)})$. By similar heuristics $\tilde{\Gamma}''(\theta_0) = O_p(1)$ and

$$\sup\{|\tilde{\Gamma}'''(\tilde{\theta})| : |\tilde{\theta} - \theta_0| \le \epsilon\} = O_p(h^{-7(p+q)/2}).$$

Thus we expect that the conclusion of Theorem 1 will continue to hold if $\delta > 6(p+q)$.

**Proof of Corollary 2:** Note that

$$\tilde{\Gamma} = \Gamma(\hat{F}, F_{\hat{\theta}}, \tilde{F})$$

where $F_\theta$ has $Y \mid X = x$ have the parametric model distribution $F(\cdot \mid x, \theta)$. The argument of Theorem 1 with $\hat{F}$ replaced by $F_{\theta_0}$ where $\theta_0$ is the true value of $\theta$ yields that $\Gamma(\hat{F}, F_{\theta_0}, \tilde{F})$ obeys the conclusion of Theorem 1 (and subsequent propositions) with $\gamma_{12} = \gamma_{32} = 0$ (and no restrictions on $\Delta$ which doesn't appear since $p = 0$). But, by Lemma 1,

$$\begin{aligned} \Gamma(\hat{F}, F_{\hat{\theta}}, \tilde{F}) &= \Gamma(\hat{F}, F_{\theta_0}, \tilde{F}) \\ &+ \int_w (\int_y A((y,w)(f_{\hat{\theta}}(y,w) - f_{\theta_0}(y,w)))^2 dy a^2(w) d\tilde{F}(w) \\ &+ O_p(\|f_{\hat{\theta}} - f_\theta\|^3). \end{aligned} \tag{A.35}$$

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.

**Heuristics for Corollary 3:** Note that,

$$\tilde{\Gamma}^* = \Gamma(\hat{F}^*_{\hat{\theta}}, \hat{\mathrm{F}}^*_{\hat{\theta}}, \tilde{F})$$

where $\hat{F}_\theta$ is the smoothed empirical distribution (by $K_h^{(p+q+1)}$) of $(y_i(\theta), x_i)$ where $y_i(\theta) \equiv y_i - m_\theta(v_i)$ and similarly for $\hat{\mathrm{F}}_\theta$. Thus,

$$\begin{aligned} \tilde{\Gamma}^* &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \left\{\hat{m}^*(w_i, v_i, \theta_0) - \hat{M}^*(w_i, \theta_0)\right\}^2 \\ &- \tfrac{2}{N}\textstyle\sum_{i=1}^N \{(\hat{m}^*(w_i, v_i, \theta_0) - \hat{M}^*(w_i, \theta_0)(\hat{A}_N(v_i) - \hat{B}_N(v_i))\} \\ &+ \tfrac{1}{N}\textstyle\sum_{i=1}^N \left\{\hat{A}_N(v_i) - \hat{B}_N(v_i)\right\}^2 \end{aligned} \tag{A.36}$$

where

$$\begin{aligned} \hat{A}_N(v_i) &= \textstyle\sum_{j=1}^N (m_{\hat{\theta}}(v_i) - m_{\theta_0}(v_i)) K_h^{(p+q)}(x_i - x_j) / \sum_{j=1}^N K_h^{(p+q)}(x_i - x_j) \\ \hat{B}_N(v_i) &= \textstyle\sum_{j=1}^N (m_{\hat{\theta}}(v_i) - m_{\theta_0}(v_i)) K_H^{(p)}(w_i - w_j) / \sum_{j=1}^N K_H^{(p)}(w_i - w_j). \end{aligned}$$

Heuristically, with $\theta$ one dimensional for simplicity

$$\hat{A}_N(v_i) - \hat{B}_N(v_i) = (\hat{\theta} - \theta)(\frac{\partial m_\theta}{\partial \theta}(v_i) - E(\frac{\partial m_\theta}{\partial \theta}(v_i) \mid w_i)\ (1 + o_p(1)).$$

Thus, we expect the third term in (A.36) to be $O_p(N^{-1})$. The second is

$$O_p\left(2\left\{\frac{1}{N}\sum_{i=1}^{N}\left\{\hat{m}^*(w_i,v_i,\theta_0)-\hat{M}^*(w_i,\theta_0)\right\}^2\right\}^{1/2}\right.$$
$$\left.\times\left\{\frac{1}{N}\sum_{i=1}^{N}\left\{\hat{A}(v_i)-\hat{B}(v_i)\right\}^2\right\}^{1/2}\right)$$
$$=O_p(N^{-1/2}h^{-(p+q)/4}N^{-1/2})=o_p(N^{-1}h^{-(p+q)/2})$$

since the first obeys Theorem 4.4 under (4.4).

## References


Aït-Sahalia, Y. (1996a), "Nonparametric Pricing of Interest Rate Derivative Securities," *Econometrica,* 64, 527-560.

Aït-Sahalia, Y. (1996b), "Testing Continuous Time Models of the Term Structure of Interest Rates," *Review of Financial Studies*, 9, 385-426.

Aït-Sahalia, Y. (1995), "The Delta Method for Nonparametric Kernel Functionals," working paper, Graduate School of Business, University of Chicago.

Aït-Sahalia, Y. and A.W. Lo (1997), "Nonparametric Estimation of State-Price-Densities Implicit in Financial Asset Prices," forthcoming in the *Journal of Finance.*

Andrews, D.W.K. (1993), "An Introduction to Econometric Applications of Functional Limit Theory for Dependent Random Variables," *Econometric Reviews,* 12, 183-216.

Bickel, P.J. and M. Rosenblatt (1973), "On Some Global Measures of the Deviations of Density Function Estimates," *Annals of Statistics*, 1, 1071-1096.

Bierens, H.J. (1990), "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443-1458.

Bierens, H.J. and W. Ploberger (1994), "Asymptotic Theory of Integrated Conditional Moment Tests," working paper, Department of Economics, Southern Methodist University.

Black, F. and M. Scholes (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, 637-654.

Doksum, K. and A. Samarov (1995), "Global Functionals and a Measure of the Explanatory Power of Covariates in Nonparametric Regression," *Annals of Statistics,* 23, 1443-1473.

Ellison, G. and S. Fisher-Ellison (1992), "A Simple Framework for Nonparametric Specification Testing," working paper, National Bureau of Economic Research.

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

Fan, J. (1994), "Testing the Goodness-of-Fit of a Parametric Density Function by Kernel Method," *Econometric Theory,* 10, 316-356.

Fan, J. and Q. Li (1996), "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica,* 64, 865-890.

Fernholz, L. (1983), *Von Mises Calculus for Statistical Functionals,* Lecture Notes in Statistics, Vol. 19, Springer-Verlag, 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.

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. and Vieu, P. (1989), *Nonparametric Curve Estimation from Time Series*, Lecture Notes in Statistics, Vol. 60, Springer-Verlag, Berlin.

Hall, P. (1984), "Central Limit Theorem for Integrated Squared Error of Multivariate Nonparametric Density Estimators," *Journal of Multivariate Analysis*, 14, 1-16.

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

Hidalgo, J. (1992), "A General Non-Parametric Misspecification Test," working paper, Department of Economics, London School of Economics.

Heckman, J., I. Ichimura, J. Smith and P. Todd (1995), "Non-Parametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA, Part II: Econometric Theory and Monte Carlo Evidence," working paper, Department of Economics, University of Chicago.

Hong, Y. and H. White (1995), "Consistent Specification Testing via Nonparametric Series Regression," *Econometrica*, 63, 1133-1160.

Horowitz, J.L. and W. Härdle (1994), "Testing a Parametric Model Against a Semiparametric Alternative," *Econometric Theory*, 10, 821-848.

de Jong, R.M. and H.J. Bierens (1994), "On the Limit Behavior of a Chi-Squared Type Test if the Number of Conditional Moments Tested Approaches Infinity," *Econometric Theory*, 9, 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. and Q.H. Vuong (1996), "Nonparametric Selection of Regressors: The Nonnested Case," *Econometrica*, 64, 207-219.

Lee, B.J. (1988), "Nonparametric Tests Using a Kernel Estimation Method," Doctoral 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.

Merton, R.C. (1990), *Continuous-Time Finance*, Blackwell, Oxford.

von Mises, R. (1947) "On the Asymptotic Distribution of Differentiable Statistical Functions," *Annals of Mathematical Statistics*, 18, 309-348.

Monfort, A. (1996), "A Reappraisal of Misspecified Econometric Models," *Econometric Theory*, 12, 597-619.

Nadaraya, E.A. (1964), "On Estimating Regression," *Theory of Probability and Its Applications,* 9, 141-142.

Ossiander, M. (1987), "A Central Limit Theorem Under Metric Entropy with $L_2$ Bracketing", *Annals of Probability*, 15, 1897-1919.

Nadaraya, E.A. (1983), *Nonparametric Estimation of Probability Densities and Regression Curves*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

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. and T.M. Stoker (1992), "A Regression Test of Semiparametric Index Model Specification," working paper, Sloan School of Management, MIT.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

Staniswalis, J.G. and T.A. Severini (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*, 393-406, Academic Press, New York.

Vieu, P. (1994), "Choice of Regressors in Nonparametric Estimation," *Computational Statistics and Data Analysis,* 17, 575-594.

Watson, G.S. (1964), "Smooth Regression Analysis," *Sankhya Series A,* 26, 359-372.

Whang, Y.-J. and D.W.K. Andrews (1991), "Tests of Specification for Parametric and Semiparametric Models," Discussion Paper 968, Cowles Foundation for Economic Research, Yale University.

White, H. and Y. Hong (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 Test Based on Least Squares," *Econometric Theory*, 8, 452-475.

Yoshihara, K. (1976), "Limiting Behavior of U-Statistics for Stationary, Absolutely Regular Processes," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 35, 237-252.

Zhang, P. (1991), "Variable Selection in Nonparametric Regression with Continuous Covariates," *Annals of Statistics*, 19, 1869-1882.

Zheng, J. Xu (1996), "A Consistent Test of Functional Form via Nonparametric Estimation Function," *Journal of Econometrics*, 75, 263-290.

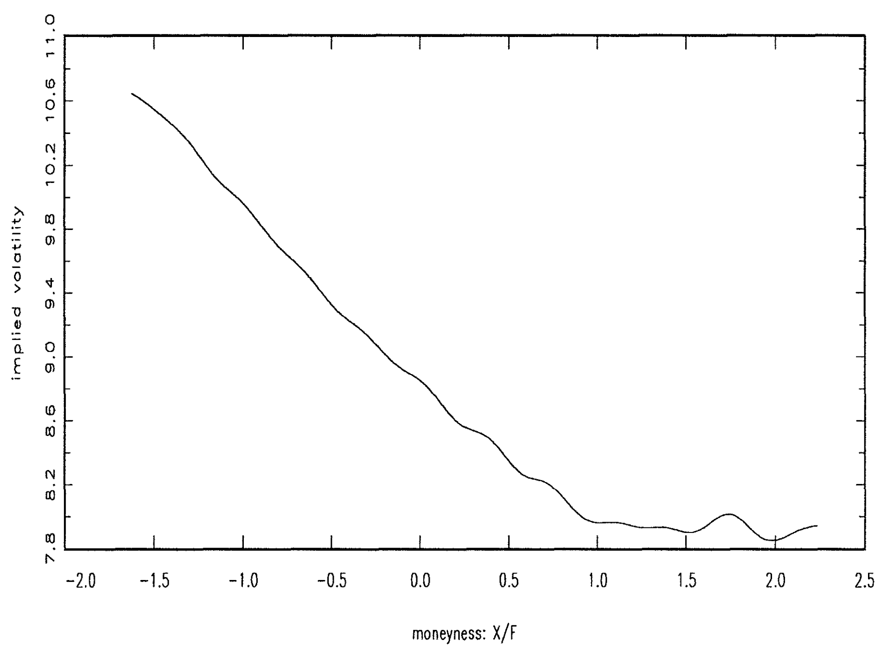


Figure 1: Implied Volatility $g(X/F)$: Semiparametric Model