SPECIFICATION TESTING IN NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION

by

Joel L. Horowitz
Department of Economics
Northwestern University
Evanston, IL 60208
USA

October 2009

ABSTRACT

In nonparametric instrumental variables estimation, the function being estimated is the solution to an integral equation. A solution may not exist if, for example, the instrument is not valid. This paper discusses the problem of testing the null hypothesis that a solution exists against the alternative that there is no solution. We give necessary and sufficient conditions for existence of a solution and show that uniformly consistent testing of an unrestricted null hypothesis is not possible. Uniformly consistent testing is possible, however, if the null-hypothesis is restricted by assuming that any solution to the integral equation is smooth. Many functions of interest in applied econometrics, including demand functions and Engel curves, are expected to be smooth. The paper presents a statistic for testing the null hypothesis that a smooth solution exists. The test is consistent uniformly over a large class of probability distributions of the observable random variables for which the integral equation has no smooth solution. The finite-sample performance of the test is illustrated through Monte Carlo experiments.

Key Words: Inverse problem, instrumental variable, series estimator, linear operator

JEL Listing: C12, C14

I thank Whitney Newey for asking a question that led to this paper. Xiaohong Chen, Hidehiko Ichimura, Sokbae Lee, and Whitney Newey provided helpful comments. This research was supported in part by NSF grants SES-0352675 and SES-0817552.

SPECIFICATION TESTING IN NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION

1. Introduction

Nonparametric instrumental variables (IV) estimation consists of estimating the unknown function g that is identified by the relation

(1.1) $E[Y - g(X)|W = w] = 0$

for almost every w in the support of the random variable W. Equivalently, g satisfies

(1.2) $Y = g(X) + U;\ E(U|W = w) = 0$

for almost every w. In this model, Y is a scalar dependent variable, X is a continuously distributed explanatory variable that may be endogenous (that is, $E(U|X = x)$ may not be zero), W is an instrument for X, and U is an unobserved random variable. The function g is assumed to satisfy mild regularity conditions but does not belong to a known, finite-dimensional parametric family. The data are an independent random sample from the distribution of (Y, X, W).

Methods for estimating g in (1.1) have been developed by Newey and Powell (2003); Darolles, Florens, and Renault (2006); Hall and Horowitz (2005); and Blundell, Chen, and Kristensen (2007). Newey, Powell, and Vella (1999) developed a nonparametric estimator for g in a different setting in which a control function is available. Horowitz and Lee (2007); Chen and Pouzo (2008); and Chernozhukov, Gagliardini, and Scaillet (2008) have developed methods for estimating quantile-regression versions of model (1.1).

All methods for estimating g in model (1.1) assume the existence of a function that satisfies (1.1). However, as is explained in Section 2 of this paper, a solution need not exist if, for example, the instrument W is not valid, that is if $E(U|W = w) \neq 0$ on a set of w values that has non-zero probability. This raises the question whether it is possible to test for existence of a solution to (1.1). This paper provides an answer to the question. The null hypothesis is that a solution to (1.1) exists. The set of alternative hypotheses consists of the distributions of (Y, X, W) for which there is no solution to (1.1). We consider tests that are consistent uniformly over this set. Uniform consistency is important because it ensures that there are not alternatives against which a test has low power even with large samples. If a test is not uniformly consistent over a specified set, then that set contains alternatives against which the test has low power. Some such alternatives may depart from the null hypothesis in extreme ways, as is illustrated by example in Section 3. We show that the null hypothesis cannot be tested consistently uniformly

over the set of alternative hypotheses without placing restrictions on g beyond those needed to estimate it when (1.1) has a solution. Specifically, there is always a distribution of (Y, X, W) such that no solution to (1.1) exists but any α-level test accepts the null hypothesis with a probability that is arbitrarily close to $1-\alpha$.

We also show that it is possible to test the hypothesis that a "smooth" solution to (1.1) exists. The test is consistent uniformly over a large class of non-smooth alternatives. The paper presents such a test. Non-existence of a solution to (1.1) is an extreme form of non-smoothness, so in sufficiently large samples, the test presented here rejects the null hypothesis that g is smooth if no solution to (1.1) exists.

We define g to be smooth if it has sufficiently many square-integrable derivatives. With a sufficiently large sample, the test presented here rejects the null hypothesis that (1.1) has a smooth solution if no solution exists or if one exists but is not smooth. The possibility of rejecting a non-smooth solution is desirable in many applications. For example, a demand function or Engel curve is unlikely to be discontinuous or wiggly. Thus, rejection of the hypothesis that a demand function or Engel curve is smooth implies misspecification of the model that identifies the curve or function (e.g., that W is not a valid instrument for X). Accordingly, the test described here is likely to be useful in many applications.

The test presented here is related to Horowitz's (2006) test of the hypothesis that g belongs to a specified, finite-dimensional parametric family. A smooth function can be approximated accurately by a finite-dimensional parametric model consisting of a truncated series expansion with suitable basis functions. See Section 4 for details. The approximation to a non-smooth function is less accurate. Therefore, one can test for existence of a smooth solution to (1.1) by testing the adequacy of a truncated series approximation to g. The test statistic is similar to Horowitz's (2006) statistic for testing a finite-dimensional parametric model, but its asymptotic behavior is different. In Horowitz (2006), the dimension of the parametric model is fixed. In the present setting, the dimension (or length) of the series approximation increases as the sample size increases. This is necessary to ensure that the truncation error remains smaller than the smallest deviation from the null hypothesis that the test can detect. The increasing dimension of the null hypothesis model changes the asymptotic behavior of the test statistic in ways that are explained in Section 4.

Section 2 of this paper gives a necessary and sufficient condition for existence of a function g that solves (1.1). It also explains why a solution may not exist. Section 3 presents an example showing that it is not possible to construct a uniformly consistent test of the hypothesis

that (1.1) has a solution. No matter how large the sample is, there are always alternatives against which any test has low power. Section 4 describes the statistic for testing the hypothesis that (1.1) has a smooth solution. This section also explains the test's asymptotic behavior under the null and alternative hypotheses. Section 5 presents the results of a Monte Carlo investigation of the test's finite-sample behavior, and Section 6 presents conclusions. The proofs of theorems are in the appendix, which is Section 7.

2. Necessary and Sufficient Conditions for a Solution to (1.1)

Necessary and sufficient conditions for existence of a solution to (1.1) are given by Picard's theorem (e.g., Kress 1999, Theorem 15.18). Before stating the theorem, we define notation that will be used throughout the paper.

Assume that X and W are real-valued random variables. Assume, also, that the support of (X,W) is contained in $[0,1]^2$. This assumption entails no loss of generality as it can always be satisfied by, if necessary, carrying out monotone transformations of X and W. Let f_{XW} and f_W, respectively, denote the probability density functions (with respect to Lebesgue measure) of (X,W) and W. For $x, z \in [0,1]$, define

$$t(x,z) = \int_0^1 f_{XW}(x,w) f_{XW}(z,w) dw .$$

Define the operator $T : L_2[0,1] \to L_2[0,1]$ by

$$(T\nu)(z) = \int_0^1 t(x,z)\nu(x)dx ,$$

where ν is any function in $L_2[0,1]$. Assume that T is non-singular. Denote its eigenvalues and eigenvectors by $\{\lambda_j, \phi_j : j = 1,2,...\}$. Sort these so that $\lambda_1 \geq \lambda_2 \geq ... > 0$. Under the assumptions stated in Section 4, T is a compact operator. Therefore, its eigenvectors form a complete, orthonormal basis for $L_2[0,1]$. Moreover, the eigenvalues are strictly positive and have 0 as their only limit point.

Now, for $z \in [0,1]$ define

$$q(z) = E_{YW}[Yf_{XW}(z,W)]$$

where E_{YW} denotes the expectation with respect to (Y,W). Hall and Horowitz (2005) show that (1.1) is equivalent to the operator equation

(2.1) $\quad q = Tg$.

Therefore, the conditions for existence of a solution to (1.1) are the same as the conditions for existence of a function g that satisfies (2.1).

Let $\langle \cdot, \cdot \rangle$ denote the inner product in $L_2[0,1]$. The following theorem gives necessary and sufficient conditions for existence of a function g that solves (1.1) and (2.1).

Theorem 2.1 (Picard): Let T be a compact, non-singular operator, and assume that $\|q\| \neq 0$. Then (2.1) has a solution if and only if

$$\sum_{j=1}^{\infty} \frac{\langle q, \phi_j \rangle^2}{\lambda_j^2} < \infty .$$

If a solution exists, it is

$$g(x) = \sum_{j=1}^{\infty} b_j \phi_j(x) ,$$

where

$$b_j = \frac{\langle q, \phi_j \rangle}{\lambda_j} . \quad \blacksquare$$

Testing the hypothesis that (1.1) has a solution is equivalent to testing the hypothesis that $\sum_{j=1}^{\infty} b_j^2 < \infty$ against the alternative $\sum_{j=1}^{\infty} b_j^2 = \infty$. The quantities $\langle q, \phi_j \rangle$ are the generalized Fourier coefficients of q using the basis functions $\{\phi_j\}$. That is,

$$q(z) = \sum_{j=1}^{\infty} \langle q, \phi_j \rangle \phi_j(z) .$$

Therefore a solution to (1.1) exists if and only if the Fourier coefficients of q converge sufficiently rapidly relative to the eigenvalues of T. It is easy to construct examples in which the generalized Fourier coefficients converge more slowly than the eigenvalues so that $\sum_{j=1}^{\infty} b_j^2 = \infty$.

In applied econometric research g may be an Engel curve, demand function, or some other economically meaningful function whose existence is not in question. In this case, (1.1) may not have a solution if W is not a valid instrument. Specifically, suppose that $E(U \mid W = w) \neq 0$ on a set of w values with positive probability. Then arguments like those used to obtain (2.1) show that g solves not (1.1) or (2.1) but

(2.2) $\quad (Tg)(z) = q(z) - E_{UW}[U f_{XW}(z, W)] .$

The misspecified models (1.1) and (2.1) need not have solutions when W is an invalid instrument and (2.2) is the correct specification.

3 The Impossibility of Uniformly Consistent Testing with Unrestricted Null and Alternative Hypotheses

This section presents an example in which uniformly consistent testing of the hypothesis that (1.1) has a solution is not possible. The distributions used in the example are nested in any reasonable class of probability distributions for (Y, X, W) in (1.1), so the impossibility result obtained with the example holds generally.

The example consists of a simple null-hypothesis and a simple alternative-hypothesis. "Simple" in this context means that there are no unknown population parameters in either the null or alternative hypotheses. The null hypothesis is that a specific function g solves (1.1). Under the alternative hypothesis, (1.1) has no solution. It follows from the Neyman-Pearson lemma that the likelihood ratio test is the most powerful test of the null hypothesis against the alternative. We show that regardless of the sample size, it is always possible to choose an alternative hypothesis against which the power of the likelihood ratio test is arbitrarily close to its size. Therefore, the likelihood ratio test is not uniformly consistent. It follows that no other test is uniformly consistent because no other test is more powerful than the likelihood ratio test.

To construct the example, write (1.1) in the equivalent form

(3.1) $Y = E[g(X)|W] + V;\ E(V|W) = 0$,

where $V = Y - E(Y|W)$. Assume that f_{XW} is known and is

$$f_{XW}(x,w) = 1 + 2\sum_{j=1}^{\infty} \lambda_j^{1/2} \cos(j\pi x)\cos(j\pi w),$$

where the λ_j's are constants satisfying $\lambda_1 \geq \lambda_2 \geq ... > 0$ and $\sum_{j=1}^{\infty} \lambda_j^{1/2} < \infty$. With this density function, the eigenvalues of T are $\{1, \lambda_1, \lambda_2, ...\}$. The eigenvectors are $\phi_1(x) = 1$ and $\phi_j(x) = 2^{1/2}\cos[(j-1)\pi x]$ for $j \geq 2$.

Assume that V is known to be distributed as $N(0,1)$ and is independent of X and W. Let $\{b_j : j = 0,1,2,...\}$ denote the Fourier coefficients of g with the cosine basis. That is,

(3.2) $$g(x) = b_0 + 2^{1/2}\sum_{j=1}^{\infty} b_j \cos(j\pi x).$$

Then

$$E[g(X)\,|\,W] = b_0 + 2^{1/2}\sum_{j=1}^{\infty} b_j \lambda_j^{1/2} \cos(j\pi W),$$

and model (1.1) becomes

(3.3) $$Y = b_0 + 2^{1/2}\sum_{j=1}^{\infty} b_j \lambda_j^{1/2} \cos(j\pi W) + V; \quad V \sim N(0,1).$$

Now let $J > 0$ be an integer. Consider testing the simple null hypothesis

$$H_0 : \begin{cases} b_0 = 1 \\ b_j = j^{-2} \text{ if } 1 \le j \le J \\ b_j = 0 \text{ if } j > J \end{cases}$$

against the simple alternative hypothesis

$$H_1 : \begin{cases} b_0 = 1 \\ b_j = j^{-2} \text{ if } 1 \le j \le J \\ b_j = 1 \text{ if } j > J \end{cases}.$$

Under H_0, g in (3.2) is an ordinary function on $[0,1]$, and g solves (1.1). Under H_1, g is a linear combination of an ordinary function and a delta function, so g is not a function on $[0,1]$ in the usual sense and (1.1) has no solution.

Let the data be the independent random sample $\{Y_i, X_i, W_i : i = 1,...,n\}$. We show that for any fixed n, no matter how large, it is possible to choose J so that the power of the likelihood ratio test of H_0 against H_1 is arbitrarily close to its size.

The likelihood ratio statistic for testing H_0 against H_1 is

(3.4) $$LR = (1/2)\sum_{i=1}^{n} \left\{ \left[Y_i - 1 - 2^{1/2}\sum_{j=1}^{J} j^{-2} \lambda_j^{1/2} \cos(j\pi W_i) \right]^2 \right.$$

$$\left. - \left[Y_i - 1 - 2^{1/2}\sum_{j=1}^{J} j^{-2} \lambda_j^{1/2} \cos(j\pi W_i) - 2^{1/2} \sum_{j=J+1}^{\infty} \lambda_j^{1/2} \cos(j\pi W_i) \right]^2 \right\}.$$

Substituting (3.3) into (3.4) shows that under H_0, the likelihood ratio statistic is

$$LR_0 = 2^{1/2}\sum_{i=1}^{n} \sum_{j=J+1}^{\infty} \lambda_j^{1/2} \cos(j\pi W_i) V_i - \sum_{i=1}^{n} \left[\sum_{j=J+1}^{\infty} \lambda_j^{1/2} \cos(j\pi W_i) \right]^2.$$

Under H_1, the likelihood ratio statistic is

$$LR_1 = 2^{1/2}\sum_{i=1}^{n}\sum_{j=J+1}^{\infty}\lambda_j^{1/2}\cos(j\pi W_i)V_i + \sum_{i=1}^{n}\left[\sum_{j=J+1}^{\infty}\lambda_j^{1/2}\cos(j\pi W_i)\right]^2.$$

Therefore,

$$LR_1 - LR_0 = 2\sum_{i=1}^{n}\left[\sum_{j=J+1}^{\infty}\lambda_j^{1/2}\cos(j\pi W_i)\right]^2$$

$$\leq 2n\left(\sum_{j=J+1}^{\infty}\lambda_j^{1/2}\right)^2.$$

Because $\sum_{j=1}^{\infty}\lambda_j^{1/2} < \infty$, $LR_1 - LR_0$ can be made arbitrarily small by making J sufficiently large. Therefore, we obtain the following result, which is proved in the Appendix.

Proposition 3.1: Let $c_{n\alpha}$ denote the α-level critical value of LR when the sample size is n. That is, $P(LR > c_{n\alpha} \mid H_0) = \alpha$. Let n be fixed. For each $\varepsilon > 0$ there is a J_0 such that the power of the α-level likelihood ratio test of H_0 against H_1 is less than or equal to $\alpha + \varepsilon$ whenever $J \geq J_0$. That is, $P(LR > c_{n\alpha} \mid H_1) \leq \alpha + \varepsilon$ whenever $J \geq J_0$. ■

Now consider the class of alternative hypotheses consisting of distributions of (Y, X, W) for which H_1 is true for some $J < \infty$. Because no test is more powerful than the likelihood ratio test, it follows from Proposition 3.1 that no test of H_0 is consistent uniformly over this class. Regardless of the sample size n, there are always distributions in H_1 for some finite J against which the power of any test is arbitrarily close to the test's level. The intuitive reason is that Fourier components of g corresponding to eigenvectors of T with small eigenvalues have little effect on Y and, therefore, are hard to detect empirically. This is illustrated by (3.3), where Y is insensitive to changes in Fourier coefficients b_j that are associated with very small eigenvalues λ_j. This problem can be overcome by restricting the null and alternative hypotheses so as to avoid the need for estimating or testing Fourier coefficients associated with very small eigenvalues of T. Section 4 presents a way of doing this.

4 A Uniformly Consistent Test of the Hypothesis That (1.1) Has a Smooth Solution

In this section, we restrict the null hypothesis by requiring g to be smooth in the sense that it has s square-integrable derivatives, where s is a sufficiently large integer. Under the alternative hypothesis, (1.1) has no smooth solution. In sufficiently large samples, the resulting

test rejects the null hypothesis if (1.1) has no solution or if (1.1) has a non-smooth solution. As was explained in the introduction, a non-smooth solution to (1.1) is an indicator of misspecification (possibly due to an invalid instrument) in many applications. Therefore, the ability to reject non-smooth solutions to (1.1) can be a desirable property of a test.

4.1 Motivation

We begin with an informal discussion that provides intuition for the test that is developed here. Let $\{\psi_j : j = 1,2,...\}$ be a complete, orthonormal basis for $L_2[0,1]$. Suppose for the moment that under H_0, the solution to (1.1) has the finite-dimensional representation

$$g(x) = \sum_{j=1}^{J} b_j \psi_j(x), \tag{4.1}$$

for some fixed $J < \infty$ and (generalized) Fourier coefficients $\{b_j : j = 1,..., J\}$. Equation (4.1) restricts g to a finite-dimensional parametric family. The null hypothesis that (4.1) is a solution to (1.1) for some set of b_j's can be tested against the alternative that it is not by using the test of Horowitz (2006). The test statistic is

$$\tau_{Pn} = \int_0^1 \tilde{S}_n(z)^2 dz, \tag{4.2}$$

where

$$\tilde{S}_n(z) = n^{-1/2} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{J} \hat{b}_j \psi_j(X_i) \right] \hat{f}_{XW}^{(-i)}(z, W_i),$$

$\hat{b}_j$ is an estimator of b_j that is $n^{-1/2}$-consistent under the null hypothesis, and $\hat{f}_{XW}^{(-i)}$ is a leave-observation-i-out nonparametric kernel estimator of f_{XW}. Specifically,

$$\hat{f}_{XW}^{(-i)}(x, w) = \frac{1}{(n-1)h^2} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{x - X_j}{h}, \frac{w - W_j}{h} \right), \tag{4.3}$$

where h is a bandwidth and K is a kernel function of a 2-dimensional argument. Horowitz shows that if (4.1) is a solution to (1.1), then τ_{Pn} is asymptotically distributed as a weighted sum of independent chi-square random variables with one degree of freedom. Horowitz (2006) also shows that τ_{Pn} is consistent uniformly over a class of nonparametric alternative hypotheses whose distance from (4.1) is proportional to $n^{-1/2}$.

Now let g be nonparametric, but suppose that its derivatives through order s are square integrable on $[0,1]$. Then g has the infinite-dimensional Fourier representation

$$g(x) = \sum_{j=1}^{\infty} b_j \psi_j(x)$$

but can be approximated accurately by the finite-dimensional model that is obtained by truncating this series. Specifically, let $\{J_n : n = 1,2,...\}$ be a sequence of positive integers with $J_n \to \infty$ as $n \to \infty$. Define

$$g_n(x) = \sum_{j=1}^{J_n} b_j \psi_j(x). \tag{4.4}$$

Then for a wide variety of basis functions $\{\psi_j\}$ that includes trigonometric functions and orthogonal polynomials, the error of g_n as an approximation to g satisfies

$$\|g - g_n\| = O(J_n^{-s}),$$

where $\|\cdot\|$ denotes the norm in $L_2[0,1]$. Thus, a smooth function g can be approximated accurately by a finite-dimensional parametric function. This suggests testing the null hypothesis that (1.1) has a smooth solution by using Horowitz's (2006) procedure to test the hypothesis that (4.4) is the solution. If J_n is sufficiently large, the approximation error will be small compared to the minimum deviation from (4.4) that the test can detect. On the other hand, if the solution to (1.1) is non-smooth or does not exist, then (4.4) will be a poor approximation to the solution to (1.1), and the test will reject the null hypothesis if the sample is sufficiently large.

The main difference between this version of Horowitz's (2006) test and the test based on τ_{Pn} in (4.2) is that when g is nonparametric, J_n must increase as n increases to ensure that the approximation error remains too small to be detected by the test. This changes the asymptotic distributional properties of the test statistic. Among other things, the test statistic is asymptotically degenerate (its asymptotic distribution is concentrated at a single point) under the null hypothesis. We deal with this problem here by splitting the sample into halves. One half is used to estimate the b_j's and the other half is used to construct the test statistic. The sample splitting procedure is explained in more detail in Section 4.3. The degeneracy problem is well-known in nonparametric testing, and a variety of solutions are possible. Sample-splitting leads to a relatively simple test. Other potential solutions that may have some advantages but are much more complicated analytically are discussed in Section 4.8.

4.2 The Null and Alternative Hypotheses

This section provides formal statements of the null and alternative hypotheses of the test that is developed in this paper. The test statistic is presented in Section 4.3.

We use the following notation. For a function $\nu:[0,1]\to\mathbb{R}$ and integer $\ell\geq 0$, define

$$D_\ell\nu(x)=\frac{\partial^\ell\nu(x)}{\partial x^\ell}$$

whenever the derivative exists. Define $D_0\nu(x)=\nu(x)$. Given an integer $s>0$, define the Sobolev norm

$$\|\nu\|_s=\left\{\sum_{\ell=0}^{s}\int_0^1[D_\ell\nu(x)]^2\,dx\right\}^{1/2} \tag{4.5}$$

and the function space

$$\mathcal{H}_s=\{\nu:[0,1]\to\mathbb{R}:\ \|\nu\|_s\leq C_0\},$$

where $C_0<\infty$ is a constant.

The null hypothesis in the remainder of this paper is

H_0: Equation (1.1) has a solution $g\in\mathcal{H}_s$ for an integer $s\geq 2$.

The alternative hypothesis is

H_1: Equation (1.1) does not have a solution in $\mathcal{H}_s$ for any $s\geq 2$.

As is explained in Section 2, H_0 can be false and H_1 true if the instrument W is not valid (that is, $E(U|W=w)\neq 0$ on some set of w values whose probability exceeds 0). H_0 can also be false if measurement errors or omitted variables cause an instrument that is valid in a correctly specified model to be invalid in the model with measurement errors or omitted variables.

4.3 The Test Statistic

This section presents the statistic for testing H_0. The data are the independent random sample $\{Y_i,X_i,W_i:i=1,...,n\}$. To avoid unimportant notational complexities, we assume that n is even. If n is odd, drop one randomly selected observation. This has a negligible effect on the power of the test. Define $\hat{f}_{XW}^{(-i)}$ as in (4.3). Define $\mathcal{S}_1=\{Y_i,X_i,W_i:i=1,...,n/2\}$ and $\mathcal{S}_2=\{Y_i,X_i,W_i:i=n/2+1,...,n\}$. Let $\hat{b}_j$ ($j=1,...,J_n$) be consistent estimators of the Fourier coefficients b_j in (4.4) that are obtained from the data in $\mathcal{S}_2$. The $\hat{b}_j$'s can be obtained by using

the method of Blundell, Chen, and Kristensen (2007), but the derivation of the asymptotic distribution of the test statistic is simpler if the method explained in Section 4.4 is used. Define

$$\hat{g}_n(x) = \sum_{j=1}^{J_n} \hat{b}_j \psi_j(x).$$

Now define

$$S_n(z) = (2/n)^{1/2} \sum_{i \in \mathcal{S}_1} [Y_i - \hat{g}_n(X_i)] \hat{f}_{XW}^{(-i)}(z, W_i).$$

The test statistic is

$$\tau_n = \int_0^1 S_n(z)^2 dz.$$

H_0 is rejected if τ_n is large.[1]

The test works because J_n can be chosen so that truncation error in the finite-series approximation to g is negligibly small. Therefore, under H_0, $S_n(z)$ estimates

$$(4.6) \qquad (2/n)^{1/2} \sum_{i \in \mathcal{S}_1} U_i f_{XW}(z, W_i),$$

where $U_i = Y_i - g(X_i)$. Under H_0, the quantity in (4.6) is a random variable with mean 0 and finite variance, so τ_n is bounded in probability. Under H_1, the truncation error is non-negligible, and τ_n diverges as $n \to \infty$.

4.4 A Sieve Estimator of g

This section describes a modified version of the estimator of Blundell, Chen, and Kristensen (2007). The derivation of the asymptotic distribution of τ_n under H_0 is simpler with the modified estimator than with the original one.

For $w \in [0,1]$, define

$$m(w) = E(Y \mid W = w) f_W(w).$$

Define the operator $A: L_2[0,1] \to L_2[0,1]$ by

$$(A\nu)(w) = \int_0^1 \nu(x) f_{XW}(x, w) dx.$$

---

[1] It may seem odd to combine the series estimator $\hat{g}_n$ and the kernel estimator $\hat{f}_{XW}^{(-i)}$ in the same statistic. The reason for doing this is that my methods of proof require a density estimator that converges uniformly over $(x, w) \in [0,1]^2$ at a sufficiently fast rate. Kernel and log-spline series estimators satisfy these requirements, but the kernel estimator is much easier to compute. Therefore, I use the kernel estimator.

Then (1.1) is equivalent to the operator equation

(4.7) $Ag = m$.

The estimator of g is defined in terms of series expansions of g, m, and A. As before, let $\{\psi_j\}$ denote a complete, orthonormal basis for $L_2[0,1]$. The expansions are

$$g(x) = \sum_{j=1}^{\infty} b_j \psi_j(x),$$

$$m(w) = \sum_{k=1}^{\infty} a_k \psi_k(w),$$

and

$$f_{XW}(x,w) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} c_{jk}\psi_j(x)\psi_k(w),$$

where

$$b_j = \int_0^1 g(x)\psi_j(x)dx,$$

$$a_k = \int_0^1 m(w)\psi_k(w)dw,$$

and

$$c_{jk} = \int_{[0,1]^2} f_{XW}(x,w)\psi_j(x)\psi_k(w)dwdx.$$

We need estimators of a_k and c_{jk} ($j,k = 1,...,J_n$). These are

$$\hat{a}_k = (2/n)\sum_{i\in\mathcal{S}_2} Y_i\psi_k(W_i)$$

and

$$\hat{c}_{jk} = (2/n)\sum_{i\in\mathcal{S}_2} \psi_j(X_i)\psi_k(W_i).$$

In addition, define the operator $\hat{A}$ that estimates A by

$$(\hat{A}\nu)(w) = \int_0^1 \nu(x)\,\hat{f}_{XW}(x,w)dx,$$

where

(4.8) $$\hat{f}_{XW}(x,w) = \sum_{j=1}^{J_n}\sum_{k=1}^{J_n} \hat{c}_{jk}\psi_j(x)\psi_k(w).$$

Also define the estimator

$$\hat{m}(w) = \sum_{k=1}^{J_n} \hat{a}_k \psi_k(w).$$

Finally, define the set of functions on $L_2[0,1]$

$$\mathcal{H}_{ns} = \left\{ \nu = \sum_{j=1}^{J_n} \nu_j \psi_j : \|\nu\|_s \leq C_0 \right\}.$$

The estimator of g is

$$\hat{g} = \arg\min_{\nu \in \mathcal{H}_{ns}} \left\| \hat{A}\nu - \hat{m} \right\|, \tag{4.9}$$

where $\|\cdot\|$ denotes the norm in $L_2[0,1]$. The constraint $\nu \in \mathcal{H}_{ns}$ is not binding in sufficiently large samples, so the asymptotic properties of the estimator are unchanged if (4.9) is solved as an unconstrained optimization problem. However, the unconstrained solution to (4.9) can be unstable if n is small. Blundell, Chen, and Kristensen (2007) describe an easily computed penalization method for overcoming this problem.

The following result is used to obtain the asymptotic distribution of τ_n under H_0 and establish uniform consistency of τ_n under the alternative hypothesis defined in Section 4.2.

Theorem 4.1: Let f_{XW} have r continuous derivatives with respect to any combination of its arguments. Let assumptions 1(i)-1(iv) and 2-6 of Section 4.5 hold. If $r < \infty$, then

$$\|\hat{g} - g\| = O_p\left[J_n^{-s} + J_n^r (J_n / n)^{1/2} \right]$$

as $n \to \infty$. If $r = \infty$ and J_n satisfies assumption 8(ii) for some constant α_0 satisfying $0 < \alpha_0 < 1$, then as $n \to \infty$,

$$\|\hat{g} - g\| = O_p[(\log n)^{-s\alpha_0}]. \quad \blacksquare$$

4.5 The Asymptotic Distribution of τ_n under H_0

This section gives the asymptotic distribution of τ_n under H_0. As in the previous sections, we assume that X and W are scalars. A multivariate extension is outlined in Section 7.5 of the appendix.

We begin by defining additional notation and stating the assumptions that are used. Let $\|(x_1, w_1) - (x_2, w_2)\|_E$ denote the Euclidean distance between (x_1, w_1) and (x_2, w_2). Let $D_j f_{XW}$ denote any j'th partial or mixed partial derivative of f_{XW}. Let $D_0 f_{XW}(x, w) = f_{XW}(x, w)$. Let A^* denote the adjoint operator of A. Define

$$\rho_n = \sup_{\nu \in \mathcal{H}_{ns}} \frac{\|\nu\|}{\left\|(A^*A)^{1/2}\nu\right\|}.$$

Blundell, Chen, and Kristensen (2007) call ρ_n the sieve measure of ill-posedness and discuss its relation to the eigenvalues of $T = A^*A$. Let A_n be the operator on $L_2[0,1]$ whose kernel is

$$a_n(x,w) = \sum_{j=1}^{J_n}\sum_{k=1}^{J_n} c_{jk}\psi_j(x)\psi_k(w).$$

Finally, let $\overline{f}_{XW}(t_1,t_2)$ denote the characteristic function of f_{XW}. Define $\|t\|_E = (t_1^2 + t_2^2)^{1/2}$.

The assumptions are as follows.

Assumption 1: (i) The support of (X,W) is contained in $[0,1]^2$. (ii) (X,W) has a probability density function f_{XW} with respect to Lebesgue measure. (iii) There are an integer $r \geq 2$ and a constant $C_f < \infty$ such that $|D_j f_{XW}(x,w)| \leq C_f$ for all $(x,w) \in [0,1]^2$ and $j = 0,1,...,r$. (iv) $|D_r f_{XW}(x_1,w_1) - D_r f_{XW}(x_2,w_2)| \leq C_f \|(x_1,w_1) - (x_2,w_2)\|_E$ for any order r derivative and any (x_1,w_1) and (x_2,w_2) in $[0,1]^2$. (v) If $r = \infty$, then $\overline{f}_{XW}(t_1,t_2) = O[\exp(-d\|t\|_E^{\beta})]$ for finite constants $d > 0$ and $\beta > 1$.

Assumption 2: $E(Y^2 | W = w) \leq C_Y$ for each $w \in [0,1]$ and some constant $C_Y < \infty$.

Assumption 3: (i) (1.1) has a solution $g \in \mathcal{H}_s$ with $\|g\|_s < C_0$ and $s \geq 2$. (ii) The estimator $\hat{g}$ is as defined in (4.9). (iii) The function m has $r + s$ square-integrable derivatives.

Assumption 4: (i) The basis functions $\{\psi_j\}$ are orthonormal, complete on $L_2[0,1]$, and bounded uniformly over j. (ii) $\|A_n - A\| = O(J_n^{-r})$ if $r < \infty$ and $O(e^{-cJ_n})$ for some $c > 0$ if $r = \infty$. (iii) For any $\nu \in L_2[0,1]$ with ℓ square integrable derivatives, there are coefficients ν_j ($j = 1,2,...$) such that

$$\left\|\nu - \sum_{j=1}^{J}\nu_j\psi_j\right\| = O(J^{-\ell}).$$

Assumption 5: (i) The operator A is nonsingular. (ii) As $n \to \infty$,

$$\rho_n \sup_{\nu \in \mathcal{H}_{ns}} \frac{\|(A_n - A)\nu\|}{\|\nu\|} = O(J_n^{-s}).$$

Assumption 6: Either $r < \infty$ and $\rho_n = O(J_n^r)$ (mildly ill-posed case) or $r = \infty$ and $\rho_n = O(e^{cJ_n})$ for some finite $c > 0$ (severely ill-posed case).

Assumption 7: The kernel function K used to estimate f_{XW} has the form $K(\xi) = \kappa(\xi^{(1)})\kappa(\xi^{(2)})$, where $\xi^{(j)}$ is the j'th component of the vector ξ, κ is a symmetrical, twice continuously differentiable function on $[-1,1]$, and

$$\int_{-1}^{1} v^j \kappa(v)dv = \begin{cases} 1 \text{ if } j = 0 \\ 0 \text{ if } j \le r-1. \end{cases}$$

Assumption 8: (i) If $r < \infty$, the bandwidth, h, satisfies $h = c_h n^{-1/(2r+2)}$, where c_h is a constant, $0 < c_h < \infty$. The truncation parameter J_n satisfies $J_n = C_J n^{\gamma}$ for constants $C_J < \infty$ and γ such that $1/(2r+2s+1) < \gamma < r/[(2r+1)(r+1)]$. (ii) If $r = \infty$, then

$$J_n = \frac{\log n}{2c} - \frac{2s\alpha_0 + 1}{2c}\log\log n$$

for some α_0 satisfying $0 < \alpha_0 < 1$ and c is as in assumption 6. The bandwidth satisfies $h = c_h(\log n)^{-\gamma}$, where c_h and γ are constants, $0 < c_h < \infty$, and $1/\beta < \gamma < \alpha_0 c$.

Assumptions 1 and 2 are smoothness and boundedness conditions. Assumption 3 defines the null hypothesis and the estimator of g. It also ensures that the function m is sufficiently smooth. The assumption requires $\|g\|_s < C_0$ (strict inequality) to avoid complications that arise when g is on the boundary of $\mathcal{H}_s$. Deriving the asymptotic distribution of τ_n when $\|g\|_s = C_0$ is a difficult task that is beyond the scope of this paper. Assumption 4 is satisfied by trigonometric bases and B-splines that have been orthogonalized by, say, the Gram-Schmidt procedure. Orthogonal polynomials do not satisfy the boundedness requirement. However, this does not prevent the use of orthogonal polynomials in applications because, for any fixed integer J, the basis can consist of the first J orthogonal polynomials plus a rotation of B-splines or trigonometric functions that is orthogonal to the polynomials. Assumption 5(ii) ensures that A_n is a "sufficiently accurate" approximation to A on $\mathcal{H}_{ns}$. This assumption complements 4(ii), which specifies the accuracy of A_n as an approximation to A on the larger set $\mathcal{H}_s$. Assumption 5(ii) can be interpreted as a smoothness restriction on f_{XW}. For example, 5(ii) is satisfied if assumptions 4 and 6 hold and A maps $\mathcal{H}_s$ to $\mathcal{H}_{r+s}$. Assumption 5(ii) also can be interpreted as a restriction on the sizes of the values of c_{jk} for $j \neq k$. Hall and Horowitz (2005) used a similar

diagonality restriction. Assumption 6 is a simplified version of assumption 6 of Blundell, Chen, and Kristensen (2007). Blundell, Chen, and Kristensen (2007) and Chen and Reiss (2007) give conditions under which this assumption holds. Assumption 7 requires K to be a higher-order kernel if f_{XW} is sufficiently smooth. K can be replaced by a boundary kernel (Gasser and Müller 1979; Gasser, Müller, and Mammitzsch 1985) if f_{XW} does not approach 0 smoothly on the boundary of its support. The sinc kernel, among other infinite-order kernels, can be used if $r = \infty$. The rate of convergence of h in Assumption 8(i) is asymptotically optimal for estimating f_{XW}. In applications, h can be chosen by cross-validation or any of a variety of other bandwidth selection methods. Assumption 8 requires J_n to increase more rapidly than the asymptotically optimal rate for estimating g. This undersmoothing ensures that the truncation bias in the series approximation to g is $o(n^{-1/2})$ as $n \to \infty$. Rapid convergence of the truncation bias is needed because $S_n(z) = O_p(n^{-1/2})$ under H_0 for each $z \in [0,1]$, so the τ_n test will reject H_0 due to truncation bias unless this bias converges more rapidly than $n^{-1/2}$. The upper bounds on J_n in assumption 8 prevent J_n from increasing so rapidly that the variance of $\|\hat{g} - g\|$ does not converge to 0.

Now define $\sigma_U^2(w) = E(U^2 \mid W = w)$. For $z_1, z_2 \in [0,1]$ define

$$V(z_1, z_2) = E[\sigma_U^2(W) f_{XW}(z_1, W) f_{XW}(z_2, W)].$$

Define the operator Ω on $L_2[0,1]$ by

$$(\Omega v)(z_2) = \int_0^1 V(z_1, z_2) v(z_1) dz_1 . \tag{4.10}$$

Let $\{\omega_j : j = 1, 2, ...\}$ denote the eigenvalues of Ω sorted so that $\omega_1 \geq \omega_2 \geq ... \geq 0$. Let $\{\chi_{1j}^2 : j = 1, 2, ...\}$ denote independent random variables that are distributed as chi-square with one degree of freedom. The following theorem gives the asymptotic distribution of τ_n under H_0.

Theorem 4.2: If H_0 is true and assumptions 1-8 hold, then

$$\tau_n \to^d \sum_{j=1}^{\infty} \omega_j \chi_{1j}^2 . \quad \blacksquare$$

4.6 Obtaining the Critical Value

The statistic τ_n is not asymptotically pivotal, so its asymptotic distribution cannot be tabulated. This section presents a method, similar to that of Horowitz (2006), for obtaining an

approximate asymptotic critical value. The method is based on replacing the asymptotic distribution of τ_n with an approximate distribution. The difference between the true and approximate distributions can be made arbitrarily small, and the quantiles of the approximate distribution can be estimated consistently. The approximate $1-\alpha$ critical value of the τ_n test is a consistent estimator of the $1-\alpha$ quantile of the approximate distribution.

We now describe the approximate asymptotic distribution of τ_n. Under H_0, τ_n is asymptotically distributed as

$$\tilde{\tau} \equiv \sum_{j=1}^{\infty} \omega_j \chi_{1j}^2 .$$

Given any $\varepsilon > 0$, there is an integer $K_\varepsilon < \infty$ such that

$$0 < P\left(\sum_{j=1}^{K_\varepsilon} \omega_j \chi_{1j}^2 \le t \right) - P(\tilde{\tau} \le t) < \varepsilon .$$

uniformly over t. Define

$$\tilde{\tau}_\varepsilon = \sum_{j=1}^{K_\varepsilon} \omega_j \chi_{1j}^2 .$$

Let $z_{\varepsilon\alpha}$ denote the $1-\alpha$ quantile of the distribution of $\tilde{\tau}_\varepsilon$. Then $0 < P(\tilde{\tau} > z_{\varepsilon\alpha}) - \alpha < \varepsilon$. Thus, using $z_{\varepsilon\alpha}$ to approximate the asymptotic $1-\alpha$ critical value of τ_n creates an arbitrarily small error in the probability that a correct null hypothesis is rejected. Similarly, use of the approximation creates an arbitrarily small change in the power of the τ_n test when the null hypothesis is false. The approximate $1-\alpha$ critical value for the τ_n test is a consistent estimator of the $1-\alpha$ quantile of the distribution of $\tilde{\tau}_\varepsilon$. Specifically, let $\hat{\omega}_j$ $(j = 1,2,...,K_\varepsilon)$ be a consistent estimator of ω_j under H_0. Then the estimator of the approximate critical value of τ_n is the $1-\alpha$ quantile of the distribution of

$$\hat{\tau}_n = \sum_{j=1}^{K_\varepsilon} \hat{\omega}_j \chi_{1j}^2 .$$

This quantile, which will be denoted by $\hat{z}_{\varepsilon\alpha}$, can be estimated with arbitrary accuracy by simulation.

In applications, K_ε can be chosen informally by sorting the $\hat{\omega}_j$'s in decreasing order and plotting them as a function of j. They typically plot as random noise near $\hat{\omega}_j = 0$ when j is

sufficiently large. One can choose K_ε to be a value of j that is near the lower end of the "random noise" range. The rejection probability of the τ_n test is not highly sensitive to K_ε, so it is not necessary to attempt precision in making the choice.

The estimated eigenvalues $\hat{\omega}_j$ are those of the estimate of Ω that is defined by

$$(\hat{\Omega}\nu)(z_2) = \int_0^1 \hat{V}(z_1, z_2)\nu(z_1)dz_1 ,$$

where

$$\text{(4.11)} \quad \hat{V}(z_1, z_2) = n^{-1}\sum_{i=1}^{n} \hat{U}_i^2 \hat{f}_{XW}^{(-i)}(z_1, W_i)\hat{f}_{XW}^{(-i)}(z_2, W_i)$$

and $\hat{U}_i = Y_i - \hat{g}(X_i)$. The $\hat{\omega}_j$'s can be computed easily by using a finite-dimensional series estimator, like (4.8), for $\hat{f}_{XW}$. The $\hat{\omega}_j$'s are then the eigenvalues of the finite-dimensional matrix whose (j,k) element is

$$n^{-1}\sum_{i=1}^{n}\sum_{\ell=1}^{L_n}\sum_{m=1}^{L_n} \hat{U}_i^2 \hat{c}_{j\ell}\hat{c}_{km}\psi_\ell(W_i)\psi_m(W_i) ,$$

where L_n is the length of the series used in (4.11) to estimate f_{XW}.

To state the properties of the estimated eigenvalues, define

$$\tilde{g} = \arg\min_{\nu \in \mathcal{H}_{ns}} \|A\nu - m\|$$

and $\tilde{U} = Y - \tilde{g}(X)$. Let $\{\tilde{\omega}_j\}$ be the eigenvalues of the operator that is obtained by replacing $V(z_1, z_2)$ in (4.10) with $E[\tilde{U}^2 f_{XW}(z_1, W) f_{XW}(z_2, W)]$. Then $\tilde{\omega}_j \to \omega_j$ as $n \to \infty$ if H_0 is true. Let $z_{\varepsilon\alpha}$ denote the $1-\alpha$ quantile of the distribution of the random variable

$$\tilde{\tau}_\varepsilon \equiv \sum_{j=1}^{K_\varepsilon} \tilde{\omega}_j \chi_{1j}^2 ,$$

and let $z_{\varepsilon\alpha}$ denote the $1-\alpha$ quantile of the distribution of $\hat{\tau}_n$. Note that $\tilde{\tau}_\varepsilon = \tau_\varepsilon$ if H_0 is true. The following theorem shows that $\hat{z}_{\varepsilon\alpha}$ estimates $z_{\varepsilon\alpha}$ consistently if $L_n \to \infty$ as $n \to \infty$.

Theorem 4.3: Let assumptions 1-8 hold. As $n \to \infty$, (i) $\sup_{1 \le j \le K_\varepsilon} |\hat{\omega}_j - \tilde{\omega}_j| = O[(\log n)/(nh^2)^{1/2}]$ almost surely, (ii) $\sup_{1 \le j \le K_\varepsilon} |\tilde{\omega}_j - \omega_j| = O(J_n^{-s} + L_n^{-r})$, and (ii) $\hat{z}_{\varepsilon\alpha} \to^p z_{\varepsilon\alpha}$.

■

4.7 Consistency of the τ_n Test

This section presents a theorem establishing the consistency of the τ_n test against a fixed alternative hypothesis. The section also shows that for any $\varepsilon > 0$, the τ_n test rejects H_0 with probability exceeding $1-\varepsilon$ uniformly over a large class of alternative hypotheses.

Consistency against a fixed alternative is given by the following theorem. Let $\tilde{z}_\alpha$ denote the $1-\alpha$ quantile of the asymptotic distribution of τ_n under sampling from the model $Y = \tilde{g}(X) + \tilde{U}$.

Theorem 4.4: Let Assumptions 1, 2, and 4-8 hold. Then under H_1,

$$\lim_{n\to\infty} P(\tau_n > \tilde{z}_\alpha) = 1$$

for any α such that $0 < \alpha < 1$. ■

The conclusion of the theorem also holds if $\tilde{z}_\alpha$ is replaced by the estimated approximate critical value $\hat{z}_{\varepsilon\alpha}$.

We now consider uniform consistency. Let A^* denote the adjoint of A. Define

$$\bar{g} = \arg\min_{v\in\mathcal{H}_s} \|Av - m\|.$$

For each $n = 1,2,...$ and $C > 0$, define $\mathcal{F}_{nc}$ as the set of distributions of Y conditional on (X,W) satisfying (i) $E(Y^2 \mid W = w) \le C_Y$ for some constant $C_Y < \infty$ and all $w \in [0,1]$, (ii) $\left\|T\bar{g} - A^*m\right\| \ge Cn^{-1/2}$, and (iii) $h^r \|A\bar{g} - m\| / \left\|T\bar{g} - A^*m\right\| = o(1)$ as $n \to \infty$. Condition (ii) rules out alternatives that depend on x only through sequences of eigenvectors of T whose eigenvalues converge to 0 too rapidly. As the example of Section 3 shows, it is not possible to achieve consistency uniformly over these alternatives. Condition (iii) ensures that random sampling errors in $\hat{f}_{XW}^{(-i)}$ are asymptotically negligible relative to the effects of misspecification.

The following theorem states the uniform consistency result.

Theorem 4.5: Let assumptions 1, 2, and 4-8 hold. Then given any $\delta > 0$, α such that $0 < \alpha < 1$, and any sufficiently large but finite constant C,

(4.11) $$\liminf_{n\to\infty} \inf_{\mathcal{F}_{nc}} \boldsymbol{P}(\tau_n > \tilde{z}_\alpha) \ge 1 - \delta.$$

and

(4.12) $$\liminf_{n\to\infty} \inf_{\mathcal{F}_{nc}} \boldsymbol{P}(\tau_n > \hat{z}_{\varepsilon\alpha}) \ge 1 - 2\delta.$$ ■

4.8 Alternative Approaches to Testing H_0

This section describes some alternative approaches to testing H_0 that do not require sample splitting. These approaches may have certain advantages over τ_n (possibly in terms of power or weaker smoothness assumptions) but are more complicated analytically. Their investigation is left to future research.

The degeneracy problem that is solved by sample-splitting in Section 4.2 is also present in the econometrics literature of the 1990s on testing a parametric or semiparametric model of a conditional mean function against a nonparametric alternative. In its simplest form, the hypothesis tested in that literature is that

(4.13) $E[Y - G(X, \theta) \mid X = x] = 0$

for almost every x, some known function G, and an unknown finite-dimensional parameter θ that is estimated from the data. The alternative hypothesis is that there is no θ satisfying (4.13). Fan and Li (1996) review tests that encounter the degeneracy problem and use sample-splitting to overcome it. Degeneracy and the need for sample-splitting can be avoided by using a test statistic that measures the distance from 0 of an empirical analog of $E[Y - G(X, \theta) \mid X = x]$. Tests that avoid degeneracy this way include, among others, Härdle and Mammen (1993) and Horowitz and Spokoiny (2001) for testing (4.13) and Fan and Li (1996) for testing a semiparametric model of a conditional mean function.

A test of the null hypothesis of this paper that is analogous to the Härdle-Mammen and Horowitz-Spokoiny tests of (4.13) can be based on an empirical analog of the conditional moment $E[Y - g(X) \mid W = w] f_W(w)$. The analog is

$$S_n^*(w) = (nh)^{-1/2} \sum_{i=1}^{n} [Y_i - \hat{g}(X_i)] K\left(\frac{w - W_i}{h}\right),$$

where K is a kernel function of a scalar argument. Define

$$\tau_n^* = \int_0^1 S_n^*(w)^2 \, dw.$$

Under H_0, one can expect that τ_n^* differs from 0 only by random sampling error, whereas τ_n^* is large if H_0 is false. Accordingly, one might use τ_n^* to test H_0. However, deriving the asymptotic distribution of τ_n^* under H_0 requires solving a difficult problem in the theory of empirical U processes and, consequently, is beyond the scope of this paper.

Another possibility is to base a test on the optimized objective function in (4.9), $\left\|\hat{A}\hat{g} - \hat{m}\right\|$. One can also let the series approximation for $\hat{m}$ be longer than that for $\hat{g}$, thereby achieving a form of overidentification. However, results obtained with tests based on $\left\|\hat{A}\hat{g} - \hat{m}\right\|$ are highly sensitive to the value of C_0 and the lengths of the series used to estimate g and m. One can obtain virtually any result one wants by choosing these regularization parameters appropriately. Thus, a method for choosing regularization parameters is crucial to the development of any test based on $\left\|\hat{A}\hat{g} - \hat{m}\right\|$. The choice of regularization parameters is a major unsolved problem in nonparametric IV estimation. It, too, is beyond the scope of this paper.

The τ_n test described in Section 4.2 requires choosing the regularization parameter J_n, but the results of the Monte Carlo experiments discussed in Section 5 suggest that this can be done satisfactorily by using a simple heuristic procedure that is described in that section. Therefore, the need to choose J_n does not present an obstacle to implementation of the τ_n test.

Finally, suppose there are several instruments, say $W^{(1)},...,W^{(L)}$ for some $L \geq 2$, and these are believed to satisfy the moment conditions $E[Y - g(X) | W^{(\ell)} = w^{(\ell)}] = 0$ ($\ell = 1,...,L$). Suppose, further, that each moment condition (or, possibly, a subset of more than one but fewer than L conditions) identifies g when such a function exists. Then one can consider testing the hypothesis all of the moment conditions hold, thereby obtaining a version of the GMM test of overidentifying restrictions. However, such a test asks whether the same g satisfies all the moment conditions, whereas the question being addressed in this paper is whether any g satisfies at least one of the moment conditions. The availability of multiple instruments does not alter the issues that have been discussed concerning tests of the hypothesis that there is a g satisfying at least one of the moment conditions.

5. Monte Carlo Experiments

This section reports the results of a Monte Carlo investigation of the finite-sample performance of the τ_n test. The experiments use a sample size of 1000. The nominal level of the test is 0.05, and there are 1000 Monte Carlo replications in each experiment.

Realizations of (X,W) were generated by $X = \Phi(\xi)$ and $W = \Phi(\zeta)$, where Φ is the cumulative normal distribution function, $\zeta \sim N(0,1)$, $\xi = \rho\zeta + (1-\rho^2)^{1/2}\varepsilon$, $\varepsilon \sim N(0,1)$, and $\rho = 0.7$. Realizations of Y were generated from

(5.1) $Y = g(X) + \sigma_U U$,

where $U = \eta\varepsilon + (1-\eta^2)^{1/2}\nu$, $\nu \sim N(0,1)$, $\sigma_U = 0.1$, and $\eta = 0.4$. In experiments where H_0 is true, the function g is either

$$g_1(x) = 0.5 + \sum_{j=1}^{\infty} j^{-4}\cos(j\pi x)$$

or

$$g_2(x) = \sum_{j=1}^{\infty}(-1)^{j+1} j^{-2}\sin(j\pi x) .$$

The series were truncated at $j = 100$ for computational purposes. The resulting functions are displayed in Figure 1.

In experiments where H_0 is false, $E(U|W) \neq 0$ so W is not a valid instrument. Realizations of Y were generated from

(5.2) $Y = g_k(X) + \tilde{U}; \quad k = 1, 2,$

where

$$\tilde{U} = E[\Delta(X)|W] + \sigma_U U ,$$

U and σ_U are as in (5.1),

$$\Delta(x) = \frac{1}{2d} I(0.5 - d < x \leq 0.5 + d) ,$$

and $d = 0.02$ or $d = 0.005$, depending on the experiment. With model (5.2), the solution to (1.1) is

$$g(x) = g_k(x) + \Delta(x) .$$

The function g has a rectangular spike that is centered at $x = 0.5$. The spike has width $2d$ and height $1/(2d)$. Thus, the function g identified by the invalid instrument is highly non-smooth.

The computation of τ_n is easiest if the kernel estimator of f_{XW} is approximated by a truncated series expansion. With basis functions $\{\nu_j\}$, this gives

(5.1) $$\hat{f}_{XW}^{(-i)}(x,w) = \sum_{j=1}^{J}\sum_{k=1}^{J}\tilde{c}_{jk}\nu_j(x)\nu_k(w) ,$$

where J is the point at which the series is truncated and

$$\tilde{c}_{jk} = \frac{1}{(n-1)h^2}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n}\int_{[0,1]^2} K\left(\frac{x - X_\ell}{h}, \frac{w - W_\ell}{h}\right)\nu_j(x)\nu_k(w)dxdw .$$

Sample splitting is not needed for $\hat{f}_{XW}^{(-i)}$, because $\hat{f}_{XW}^{(-i)} - f_{XW}$ is asymptotically negligible in τ_n. See the proofs of lemmas 1 and 2 in Section 7. In preliminary Monte Carlo experiments the numerical performance of τ_n was unaffected by replacing $\tilde{c}_{jk}$ with its limit as $h \to 0$. This gives coefficients

$$\tilde{c}_{jk} = \frac{1}{n-1} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} \nu_j(X_\ell)\nu_k(W_\ell) . \tag{5.2}$$

Accordingly, the Monte Carlo experiments reported here use (5.1) and (5.2) to estimate f_{XW}. The basis functions are $\nu_1 = 1$ and $\nu_j(x) = \sqrt{2}\cos[(j-1)\pi x]$ for $j \geq 2$. The series was truncated at $J = 25$. Similarly, the asymptotic critical value of τ_n was estimated by setting $K_\varepsilon = 25$. The results of the experiments are not sensitive to the choice of J and K_ε. The estimated eigenvalues $\hat{\omega}_j$ are very close to 0 when $j > 25$.

We now discuss the choice of the basis functions, $\{\psi_j\}$, and truncation parameter, J_n, for the series approximation to g. Consider, first, the choice of basis functions. Estimation of g presents an ill-posed inverse problem because T^{-1} is a discontinuous operator. One consequence of this is that with samples of moderate size, it is usually possible to estimate only low-order Fourier coefficients b_j with reasonable precision. Therefore, it is desirable to choose basis functions that provide a good low-order approximation to g. Demand functions, Engel curves, and earnings functions, among other functions of interest in applied econometrics, are likely to have few inflection points. Functions with few inflection points are often well-approximated by low-degree polynomials. In preliminary Monte Carlo experiments, we found that approximating these functions accurately with trignonometric or spline bases requires longer series and leads to much noisier estimates. Accordingly, the experiments reported here use Legendre polynomials (centered and scaled to be orthonormal on [0,1]) for the basis functions.

Now consider the choice of J_n. If J_n is too small, the τ_n test will tend to reject a true H_0 because the truncated series, $\sum_{j=1}^{J_n} b_j\psi_j$, is a poor approximation to g. If J_n is too large, $\hat{g}$ will be a very noisy estimate of g and this, too, will tend to cause rejection of a correct H_0. The integrated variance of $\hat{g}$ is $E\|\hat{g} - E\hat{g}\|^2 = \sum_{j=1}^{J_n} \sigma_j^2$, where $\sigma_j^2 = Var(\hat{b}_j)$ and $\hat{b}_j$ is the estimator of b_j. The variance components σ_j^2 can be estimated by using the standard formulae

of GMM estimation. We have found through Monte Carlo experiments that as J_n increases from 1, $E\|\hat{g} - E\hat{g}\|^2$ changes little at first but increases by a factor of 10 or more when J_n crosses a "critical value." This suggests the following heuristic procedure for choosing J_n in applications: choose the largest J_n that does not produce a very large increase in the estimated value of $E\|\hat{g} - E\hat{g}\|^2$. In the experiments reported here, the large increase in $E\|\hat{g} - E\hat{g}\|^2$ occurs when the degree of the approximating polynomial increases from 3 to 4. Accordingly, the experiments reported here approximate g with a cubic polynomial.

As in Blundell, Chen, and Kristensen (2007), it is convenient for computational purposes to replace the constrained estimation problem (4.9) with a penalized estimator. However, penalization has little effect on the results of the experiments. Accordingly, the results reported here are based on solving (4.9) without imposing the constraint $\nu \in \mathcal{H}_{ns}$.

The results of the experiments are shown in Table 1. When H_0 is true and $g = g_1$, the difference between the empirical and nominal levels of the τ_n test is very small. The difference is somewhat larger when $g = g_2$ because the error in the cubic polynomial approximation to g_2 is larger than the error in the cubic approximation to g_1. The use of a quartic or higher-degree polynomial reduces the approximation error when $g = g_2$ but increases $E\|\hat{g} - E\hat{g}\|^2$ by a factor of over 100. This illustrates the importance of choosing a basis that provides a good low-order approximation to g. The probability of rejecting H_0 when it is false is high in all cases. It is higher when $d = 0.02$ than when $d = 0.005$ because there are more data points in the interval containing the spike when $d = 0.02$.

6. Conclusions

This paper has been concerned with uniformly consistent testing of the null hypothesis that the identifying equation of nonparametric IV estimation has a solution. The paper has shown that no test can be uniformly consistent over all probability distributions of (Y, X, W) for which the identifying equation has no solution. No matter how large the sample is, there are alternative distributions that depart from the null hypothesis in extreme ways but against which any test has low power. Uniformly consistent testing is possible if the null and alternative hypotheses are restricted in appropriate ways. In this paper, the null hypothesis is restricted by assuming that any solution to the identifying equation is smooth. The paper has presented a test of the hypothesis that a smooth solution exists. The test is uniformly consistent against a large class of

distributions of (Y, X, W) for which no smooth solution exists. Monte Carlo experiments have illustrated the test's finite sample performance as well as certain limitations of the test. The paper has also outlined several other testing approaches that are more complicated analytically than the one developed here but may have some advantages. The investigation of these tests is left to future research.

Several extensions of the test described in this paper may be possible. One is to a quantile regression model with endogeneity. Chen and Pouzo (2009); Chernozhukov, Gagliardini, and Scaillet (2009); and Horowitz and Lee (2007) have developed nonparametric estimators for this model. Horowitz and Lee (2009) showed how to test a parametric quantile regression model with endogeneity against a nonparametric alternative. In a quantile regression model with endogeneity, the structural function is the solution of a nonlinear integral equation. It is likely that a test of the hypothesis that this equation has a solution can be constructed from the test of Horowitz and Lee (2009) by replacing their parametric model with a sieve approximation to a smooth function. Another possible extension is to semiparametric models with endogeneity. It is likely that the tests of Horowitz (2006) and Horowitz and Lee (2009) can be adapted to provide specification tests for semiparametric mean and quantile regressions with endogeneity, such as the partially linear model of Ai and Chen (2003).

7. Mathematical Appendix

7.1 Proof of Proposition 3.1

LR_0 is continuously distributed, so for each $\varepsilon > 0$, there is a $\delta > 0$ such that $P(LR_0 \le c_{n\alpha} - \delta) \ge P(LR_0 \le c_{n\alpha}) - \varepsilon = 1 - \alpha - \varepsilon$. Choose J_0 so that

$$2n\left(\sum_{j=J_0+1}^{\infty} \lambda_j^{1/2}\right)^2 < \delta .$$

Then

$$P(LR \le c_{n\alpha} \mid H_1) = P(LR_1 \le c_{n\alpha})$$

$$= P[LR_0 + (LR_1 - LR_0) \le c_{n\alpha}]$$

$$= P[LR_0 \le c_{n\alpha} - (LR_1 - LR_0)]$$

$$\geq P(LR_0 \leq c_{n\alpha} - \delta)$$

$$\geq 1 - \alpha - \varepsilon.$$

Therefore,

$$P(LR > c_{n\alpha} \mid H_1) = 1 - P(LR \leq c_{n\alpha} \mid H_1)$$

$$\leq \alpha + \varepsilon.$$

Q.E.D.

7.2 Proof of Theorem 4.1

Define

$$g_n = \sum_{j=1}^{J_n} b_j \psi_j .$$

Let A_n be the operator whose kernel is

$$a_n(x,w) = \sum_{j=1}^{J_n}\sum_{k=1}^{J_n} c_{jk}\psi_j(x)\psi_k(w).$$

Also define

$$\hat{a}_n(x,w) = \sum_{j=1}^{J_n}\sum_{k=1}^{J_n} \hat{c}_{jk}\psi_j(x)\psi_k(w),$$

$$a(x,w) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} c_{jk}\psi_j(x)\psi_k(w),$$

and $m_n = A_n g_n$.

Proof of Theorem 4.1: The theorem is proved for the mildly ill-posed case. The proof for the severely ill-posed case is similar. By the triangle inequality,

(7.1) $$\|\hat{g} - g\| \leq \|\hat{g} - g_n\| + \|g_n - g\| .$$

By assumption 4,

(7.2) $$\|g_n - g\| = O(J_n^{-s})$$

Now consider $\|\hat{g} - g_n\|$. We have

(7.3) $$\|\hat{g} - g_n\| \leq \rho_n \|A(\hat{g} - g_n)\| .$$

But $P(\hat{A}\hat{g} = \hat{m}) \to 1$ as $n \to \infty$. Therefore, with probability approaching 1,

$$A(\hat{g} - g_n) = A\hat{g} - Ag_n$$

.

$$= (A - \hat{A})\hat{g} + \hat{m} - m - A(g_n - g).$$

The triangle inequality gives

$$\left\|A(\hat{g} - g_n)\right\| \le \left\|(\hat{A} - A)\hat{g}\right\| + \left\|\hat{m} - m\right\| + \left\|A(g_n - g)\right\|.$$

Straightforward calculations show that $\left\|\hat{m} - E\hat{m}\right\| = O_p[(J_n / n)^{1/2}]$. It follows from assumptions 3-4 that $\left\|E\hat{m} - m\right\| = O(J_n^{-r-s})$. In addition, $A_n(g_n - g) = 0$, so $\left\|A(g_n - g)\right\| = \left\|(A_n - A)(g_n - g)\right\|$. But

$$\left\|(A_n - A)(g_n - g)\right\| = \frac{\left\|(A_n - A)(g_n - g)\right\|}{\left\|g_n - g\right\|}\left\|g_n - g\right\|$$

$$= O(J_n^{-r-s})$$

by assumptions 3-4. Therefore

(7.4) $$\left\|A(\hat{g} - g_n)\right\| \le O_p[J_n^{-r-s} + (J_n / n)^{1/2}] + \left\|(\hat{A} - A)\hat{g}\right\|.$$

Now consider $\left\|(\hat{A} - A)\hat{g}\right\|$. By the triangle inequality and assumption 5

$$\left\|(\hat{A} - A)\hat{g}\right\| \le \left\|(\hat{A} - A_n)\hat{g}\right\| + \left\|(A_n - A)\hat{g}\right\|$$

$$= \left\|(\hat{A} - A_n)\hat{g}\right\| + O(\rho_n^{-1} J_n^{-s}).$$

Now

$$\left\|(\hat{A} - A_n)\hat{g}\right\| \le \sup_{\nu \in \mathcal{H}_{ns}} \left\|(\hat{A} - A_n)\nu\right\|.$$

Write ν in the form

$$\nu = \sum_{j=1}^{J_n} \nu_j \psi_j ,$$

where

$$\nu_j = \int \nu(x)\psi_j(x)dx .$$

Then

(7.5) $$\left\|(\hat{A} - A_n)\nu\right\|^2 = \sum_{k=1}^{J_n}\left[\sum_{j=1}^{J_n}(\hat{c}_{jk} - c_{jk})\nu_j\right]^2 .$$

But $\sum_{j=1}^{J_n} |\nu_j|$ is bounded uniformly over $\nu \in \mathcal{H}_{ns}$ and n. Moreover,

$$\sum_{j=1}^{J_n} \nu_j \hat{c}_{jk} = \sum_{j=1}^{J_n} \nu_j n^{-1} \sum_{i=1}^{n} \psi_j(X_i)\psi_k(W_i).$$

Therefore, it follows from Hoeffding's inequality that

$$\sum_{j=1}^{J_n} \nu_j (\hat{c}_{jk} - c_{jk}) = O_p(n^{-1/2})$$

uniformly over $\nu \in \mathcal{H}_{ns}$. Combining this result with (7.5) yields

$$\sup_{\nu \in \mathcal{H}_{ns}} \left\| (\hat{A} - A_n)\nu \right\| = O_p[(J_n / n)^{1/2}]. \tag{7.6}$$

Combining (7.4) and (7.6) yields

$$\left\| A(\hat{g} - g_n) \right\| = O_p[J_n^{-r-s} + (J_n / n)^{1/2} + \rho_n^{-1} J_n^{-s}].$$

This result and assumption 6 imply that

$$\rho_n \left\| A(\hat{g} - g_n) \right\| = O_p[J_n^{-s} + \rho_n (J_n / n)^{1/2}]. \tag{7.7}$$

The theorem follows by combining (7.1)-(7.3) and (7.7). Q.E.D.

7.3 Proofs of Theorems 4.2 and 4.3

Let H_0 be true. Define

$$S_{n1}(z) = (2/n)^{1/2} \sum_{i \in \mathcal{S}_1} U_i f_{XW}(z, W_i),$$

$$S_{n2}(z) = (2/n)^{1/2} \sum_{i \in \mathcal{S}_1} [g(X_i) - \hat{g}(X_i)] f_{XW}(x, W_i),$$

$$S_{n3}(z) = (2/n)^{1/2} \sum_{i \in \mathcal{S}_1} U_i [\hat{f}_{XW}^{(-i)}(z, W_i) - f_{XW}(z, W_i)],$$

and

$$S_{n4}(z) = (2/n)^{1/2} \sum_{i \in \mathcal{S}_1} [g(X_i) - \hat{g}(X_i)][\hat{f}_{XW}^{(-i)}(z, W_i) - f_{XW}(z, W_i)].$$

Then $S_n(z) = \sum_{j=1}^{4} S_{nj}(z)$.

Lemma 1: As $n \to \infty$, $S_{n3}(z) = o_p(1)$ uniformly over $z \in [0,1]$.

Proof: See Lemma 3 of Horowitz (2006). Q.E.D.

Lemma 2: As $n \to \infty$, $S_{n4}(z) = o_p(1)$ uniformly over $z \in [0,1]$.

Proof: Standard properties of kernel estimators give the result that

$$\max_{1\le i\le n}\sup_{x,w\in[0,1]^2}|\hat{f}_{XW}^{(i)}(x,w)-f_{XW}(x,w)|=O\left[\frac{\log n}{(nh^2)^{1/2}}\right]$$

almost surely. Therefore,

$$\sup_{z\in[0,1]}|S_{n4}(z)|\le R_n\sum_{i\in\mathcal{S}_1}|\hat{g}(X_i)-g(X_i)|,$$

where $R_n=O[(\log n)/(nh)]$ almost surely. For $i\in\mathcal{S}_1$, X_i is independent of $\hat{g}$. By Theorem 2.4 of van de Geer (2000), the class of functions $\{|H-g|:H\in\mathcal{H}_r\}$ for fixed g satisfies the conditions of the uniform law of large numbers (van de Geer 2000, Lemma 3.1). Therefore,

$$n^{-1}\sum_{i\in\mathcal{S}_1}|\hat{g}(X_i)-g(X_i)|\to\int_0^1|\hat{g}(x)-g(x)|\,f_X(x)dx$$

almost surely. But

$$\int_0^1|\hat{g}(x)-g(x)|\,f_X(x)dx\le C_4\|\hat{g}-g\|$$

for some constant $C_4<\infty$ by the Cauchy-Schwarz inequality. Therefore,

$$\sup_{z\in[0,1]}|S_{n4}(z)|=O_p\left[\frac{\log n}{h}\|\hat{g}-g\|\right]$$

$$=o_p(1).$$

Q.E.D.

Lemma 3: As $n\to\infty$,

$$S_{n2}(z)=-(2/n)^{1/2}\sum_{i\in\mathcal{S}_2}U_i\,f_{XW}(z,W_i)+r_n,$$

where $\|r_n\|=o_p(1)$.

Proof: For $z\in[0,1]$ and $\nu\in\mathcal{H}_s$, define the empirical process

$$H_n(z,\nu)=(2/n)^{1/2}\sum_{i\in\mathcal{S}_2}[f_{XW}(z,W_i)\nu(X_i)-E_{XW}f_{XW}(z,W)\nu(X)]$$

$$=(2/n)^{1/2}\sum_{i\in\mathcal{S}_2}[f_{XW}(z,W_i)\nu(X_i)-(T\nu)(z)].$$

H_n is stochastically equicontinuous by Theorem 5.12 of van de Geer (2000). Therefore, because $\hat{g}\in\mathcal{H}_s$, for each $\varepsilon>0$ and $\eta>0$ there is a $\delta>0$ such that

$$P\left[\sup_{z\in[0,1]^p}|H_n(z,\hat{g})-H_n(z,g)|>\eta\right]<\varepsilon$$

whenever $\|\hat{g} - g\| < \delta$. Because $\|\hat{g} - g\| = o_p(1)$, it follows that $H_n(z, \hat{g}) - H_n(z, g) = o_p(1)$ uniformly over $z \in [0,1]$ and

(7.8) $$S_{n2}(z) = -(n/2)^{1/2}[T(\hat{g} - g)](z) + o_p(1)$$

uniformly over $z \in [0,1]$.

Now $\hat{A}\hat{g} = \hat{m}$ with probability approaching 1 as $n \to \infty$. Some algebra now shows that

$$A(\hat{g} - g) = -\hat{A}g + \hat{m} + (A - \hat{A})(\hat{g} - g)$$

$$= \hat{m} - \hat{A}g + r_{n1},$$

where $\|r_{n1}\| = o_p(n^{-1/2})$. But

$$\hat{m} - \hat{A}g = \sum_{k=1}^{J_n}(\hat{a}_k - \sum_{j=1}^{J_n} b_j \hat{c}_{jk})\psi_k .$$

By the definitions of $\hat{a}_k$ and $\hat{c}_{jk}$,

$$\hat{a}_k - \sum_{j=1}^{J_n} b_j \hat{c}_{jk} = (2/n)\sum_{i \in \mathcal{S}_2}[Y_i - \sum_{j=1}^{J_n} b_j \psi_j(X_i)]\psi_k(W_i) .$$

Define $\Delta g = \sum_{j=1}^{J_n} b_j \psi_j - g$. Then because $Y_i = g(X_i) + U_i$,

(7.9) $$(\hat{m} - \hat{A}g)(w) = (2/n)\sum_{i \in \mathcal{S}_2}\sum_{k=1}^{J_n} U_i \psi_k(W_i)\psi_k(w) - (2/n)\sum_{i \in \mathcal{S}_2}\sum_{k=1}^{J_n} \Delta g(X_i)\psi_k(W_i)\psi_k(w).$$

Define

$$D_n(w) = (2/n)\sum_{i \in \mathcal{S}_2}\sum_{k=1}^{J_n} \Delta g(X_i)\psi_k(W_i)\psi_k(w).$$

Then

$$\|D_n\|^2 = (2/n)^2 \sum_{k=1}^{J_n}\left[\sum_{i \in \mathcal{S}_2} \Delta g(X_i)\psi_k(W_i)\right]^2$$

$$= (2/n)^2 \sum_{k=1}^{J_n}\sum_{i \in \mathcal{S}_2} \Delta g(X_i)^2 \psi_k(W_i)^2 + (2/n)^2 \sum_{k=1}^{J_n}\sum_{i \in \mathcal{S}_2}\sum_{\substack{j \in \mathcal{S}_2 \\ j \neq i}} [\Delta g(X_i)\psi_k(W_i)][\Delta g(X_j)\psi_k(W_j)]$$

$$\equiv D_{n1} + D_{n2}.$$

Now

$$ED_{n1} = (2/n)\sum_{k=1}^{J_n}\int_{[0,1]^2}\Delta g(x)^2\psi_k(w)^2\, f_{XW}(x,w)dxdw$$

$$\leq (2/n)C_f\sum_{k=1}^{J_n}\int_0^1\Delta g(x)^2\,dx$$

$$= (2/n)C_f J_n \|\Delta g\|^2$$

$$= o(n^{-1}).$$

Therefore, Markov's inequality gives

(7.10) $D_{n1} = o_p(n^{-1})$.

Moreover,

$$ED_{n2} = [1-(2/n)]\sum_{k=1}^{J_n}[E\Delta g(X)\psi_k(W)]^2\ .$$

But

$$E\Delta g(X)\psi_k(W) = \int_{[0,1]^2}\Delta g(x)\psi_k(w)\, f_{XW}(x,w)dxdw$$

.

$$= \int_0^1\Delta g(x)\delta_k(x)dx,$$

where

$$\delta_k(x) = \int_0^1\psi_k(w)\, f_{XW}(x,w)dw\,.$$

Therefore, by the Cauchy-Schwarz inequality

$$[E\Delta g(X)\psi_k(W)]^2 \leq \|\Delta g\|^2\int_0^1\delta_k(x)^2\,dx\,,$$

and

$$ED_{n2} \leq [1-(2/n)]\|\Delta g\|^2\int_0^1\sum_{k=1}^{J_n}\delta_k(x)^2\,dx\,.$$

But $\delta_k(x)$ is the k 'th Fourier coefficient of $f_{XW}(x,\cdot)$, and $|\, f_{XW}(x,w)\,| \leq C_f$. Therefore,

$$ED_{n2} \leq [1-(2/n)]C_f^2\|\Delta g\|^2$$

$$= O(J_n^{-2s})$$

$$= o(n^{-1}).$$

Markov's inequality now gives

(7.11) $D_{n2} = o_p(n^{-1})$.

Combining (7.9)-(7.11) yields

(7.12) $$(\hat{m} - \hat{A}g)(w) = (2/n)\sum_{i \in \mathcal{S}_2}\sum_{k=1}^{J_n} U_i \psi_k(W_i)\psi_k(w) + r_{n2},$$

where $\|r_{n2}\| = o_p(n^{-1/2})$.

Now

$$T(\hat{g} - g) = A^* A(\hat{g} - g)$$

so

(7.13) $$T(\hat{g} - g) = A^*(\hat{m} - \hat{A}g) + r_{n3},$$

where $\|r_{n3}\| = o_p(n^{-1/2})$. By (7.12)

$$A^*(\hat{m} - \hat{A}g)(z) = (2/n)\sum_{i \in \mathcal{S}_2}\sum_{k=1}^{J_n} U_i \psi_k(W_i)(A^*\psi_k)(z) + r_{n4},$$

where $\|r_{n4}\| = o_p(n^{-1/2})$. Now

$$(A^*\psi_k)(z) = \int_0^1 f_{XW}(z,w)\psi_k(w)dw$$

$$= \sum_{j=1}^{\infty} c_{jk}\psi_j(z).$$

Therefore,

$$A^*(\hat{m} - \hat{A}g)(z) = (2/n)\sum_{i \in \mathcal{S}_2}\sum_{j=1}^{\infty}\sum_{k=1}^{J_n} U_i c_{jk}\psi_j(z)\psi_k(W_i) + r_{n4}.$$

Define $\Delta f_{XW}(x,w) = \sum_{j=1}^{\infty}\sum_{k=1}^{J_n} c_{jk}\psi_j(x)\psi_k(w) - f_{XW}(x,w)$. Then

$$A^*(\hat{m} - \hat{A}g)(z) = (2/n)\sum_{i \in \mathcal{S}_2} U_i f_{XW}(z, W_i) + (2/n)\sum_{i \in \mathcal{S}_2} U_i \Delta f_{XW}(z, W_i) + r_{n4}$$

$$\equiv (2/n)\sum_{i \in \mathcal{S}_2} U_i f_{XW}(z, W_i) + r_{n5}(z) + r_{n4}.$$

Note that

$$\int_{[0,1]^2} [\Delta f_{XW}(x,w)]^2 \, dxdw = O(J_n^{-2r}) .$$

Therefore,

$$\|r_{n5}\|^2 = (4/n)\left\{\int_{[0,1]^2} \sigma^2(w)[\Delta f_{XW}(x,w)]^2 \, f_W(w)dxdw\right\}$$

$$= o(n^{-1}J_n^{-2r}),$$

and

$$(7.14) \quad A^*(\hat{m} - \hat{A}g)(z) = (2/n)\sum_{i\in\mathcal{S}_2} U_i \, f_{XW}(z,W_i) + r_{n6},$$

where $\|r_{n6}\| = o_p(n^{-1/2})$. Combining (7.8), (7.13), and (7.14) yields the lemma. Q.E.D.

Proof of Theorem 4.2: Define $U_i^* = U_i$ if $i \in \mathcal{S}_1$ and $U_i^* = -U_i$ if $i \in \mathcal{S}_2$. Define

$$B_n(z) = (2/n)^{1/2}\sum_{i=1}^{n} U_i^* \, f_{XW}(z,W_i) .$$

It follows from lemmas 1-3 that τ_n is asymptotically distributed as $\|B_n\|^2$, which is a degenerate U-statistic of order 2. The theorem follows from the asymptotic distribution of such a statistic. See, for example, Serfling (1980, pp. 193-194). Q.E.D.

Proof of Theorem 4.3: $|\hat{\omega}_j - \tilde{\omega}_j| = O\left(\left\|\hat{\Omega} - \tilde{\Omega}\right\|\right)$ by Theorem 5.1a of Bhatia, Davis, and McIntosh (1983). Moreover, standard calculations for kernel density estimators show that $\left\|\hat{\Omega} - \tilde{\Omega}\right\| = O[(\log n)/(nh^2)^{1/2}]$. Part (i) of the theorem follows by combining these two results. Part (ii) is a consequence of Assumption 4(ii). Part (iii) follows by combining parts (i) and (ii). Q.E.D.

7.4 Proofs of Theorems 4.4 and 4.5

Redefine

$$S_{n1}(z) = (2/n)^{1/2}\sum_{i\in\mathcal{S}_1} [Y_i - \bar{g}(X_i)] \, f_{XW}(z,W_i)$$

$$S_{n2}(z) = (2/n)^{1/2}\sum_{i\in\mathcal{S}_1} [\bar{g}(X_i) - \hat{g}(X_i)] \, f_{XW}(x,W_i) ,$$

$$S_{n3}(z) = (2/n)^{1/2}\sum_{i\in\mathcal{S}_1} [Y_i - \bar{g}(X_i)][\, \hat{f}_{XW}^{(-i)}(z,W_i) - f_{XW}(z,W_i)] ,$$

and

$$S_{n4}(z) = (2/n)^{1/2} \sum_{i \in \mathcal{S}_1} [\overline{g}(X_i) - \hat{g}(X_i)][\hat{f}_{XW}^{(-i)}(z, W_i) - f_{XW}(z, W_i)]$$

Proof of Theorem 4.4: It suffices to show that under H_1,

$$\operatorname{plim}_{n \to \infty} n^{-1} \tau_n \geq 0 .$$

Arguments like those leading to (7.8) show that $n^{-1} \|S_{n2}\|^2 = o_p(1)$. It is clear that $n^{-1} \|S_{n3}\|^2 = o_p(1)$ and $n^{-1} \|S_{n4}\|^2 = o_p(1)$. Therefore, $n^{-1} \tau_n = n^{-1} \|S_{n1}\|^2 + o_p(1)$. But $n^{-1/2} S_{n1}(z) \to (A^* m - T\tilde{g})(z)$ almost surely uniformly in $z \in [0,1]$ by Jennrich's (1969) uniform strong law of large numbers. Therefore, $n^{-1} \tau_n \to^p \left\| A^* m - T\overline{g} \right\|^2$. The theorem follows from the fact that $\left\| A^* m - T\overline{g} \right\|^2 > 0$ under H_1. Q.E.D.

Proof of Theorem 4.5: We prove (4.11). The proof of (4.12) is similar. Define $D_n = S_{n2} + S_{n4} + E(S_{n1} + S_{n3})$ and $\tilde{S}_n = S_n - D_n$. Then $\tau_n = \left\| \tilde{S}_n + D_n \right\|^2$. Use the inequality

(7.15) $a^2 \geq 0.5 b^2 - (b - a)^2$

with $a = S_n$ and $b = D_n$ to obtain

$$P(\tau_n > \tilde{z}_\alpha) \geq P\left(0.5 \|D_n\|^2 - \left\| \tilde{S}_n \right\|^2 > \tilde{z}_\alpha \right).$$

For any finite $M > 0$,

$$P\left(0.5 \|D_n\|^2 - \left\| \tilde{S}_n \right\|^2 \leq \tilde{z}_\alpha \right) = P\left(0.5 \|D_n\|^2 \leq \tilde{z}_\alpha + \left\| \tilde{S}_n \right\|^2, \left\| \tilde{S}_n \right\|^2 \leq M \right)$$

$$+ P\left(0.5 \|D_n\|^2 \leq \tilde{z}_\alpha + \left\| \tilde{S}_n \right\|^2, \left\| \tilde{S}_n \right\|^2 > M \right)$$

$$\leq P\left(0.5 \|D_n\|^2 \leq \tilde{z}_\alpha + M \right) + P\left(\left\| \tilde{S}_n \right\|^2 > M \right).$$

$\left\| \tilde{S}_n \right\|$ is bounded in probability uniformly over $\mathcal{F}_{nc}$. Therefore, for each $\varepsilon > 0$ there is $M_\varepsilon < \infty$ such that for all $M > M_\varepsilon$

$$P\left(0.5 \|D_n\|^2 - \left\| \tilde{S}_n \right\|^2 \leq \tilde{z}_\alpha \right) \leq P\left(.5 \|D_n\|^2 \leq \tilde{z}_\alpha + M \right) + \varepsilon .$$

Equivalently,

$$P\left(0.5\|D_n\|^2 - \left\|\tilde{S}_n\right\|^2 > \tilde{z}_\alpha\right) \geq P\left(.5\|D_n\|^2 > \tilde{z}_\alpha + M\right) - \varepsilon$$

and

(7.16) $$P(\tau_n > \tilde{z}_\alpha) \geq P\left(.5\|D_n\|^2 > \tilde{z}_\alpha + M\right) - \varepsilon .$$

Now a further application of (7.15) with $a = D_n$ and $b = E(S_{n1} + S_{n3})$ gives

$$\|D_n\|^2 \geq 0.5\|E(S_{n1} + S_{n3})\|^2 - \|S_{n2} + S_{n4}\|^2 .$$

Some algebra shows that $\|S_{n2} + S_{n4}\|^2 = O_p(1)$ as $n \to \infty$, $ES_{n1}(z) = (n/2)^{1/2}(T\bar{g} - A^* m)(z)$, and $\|ES_{n3}\| = O\left(n^{1/2} h^r \|A\bar{g} - m\|\right)$. Therefore,

(7.17) $$\|D_n\|^2 \geq .125n\left\|T\bar{g} - A^* m\right\|^2 + O_p(1)$$

uniformly over $\mathcal{F}_{nc}$ as $n \to \infty$. Inequality (4.11) follows by substituting (7.17) into (7.16) and choosing C to be sufficiently large. Q.E.D.

7.5 Multivariate Extension

Let X and W be $p \geq 2$-dimensional random vectors. Assume that $(X, W) \in [0,1]^{2p}$. Let $g : [0,1]^p \to \mathbb{R}$ be identified by (1.1) if a solution to this equation exists. Let $\{\psi_j\}$ be a uniformly bounded orthonormal basis for $L_2[0,1]^p$. A sieve estimator of g can be obtained by substituting the multivariate basis functions into the formulae of Section 4.4.

To describe the properties of this estimator and state the multivariate versions of H_0 and H_1, it is necessary to define the p-dimensional version of $\mathcal{H}_s$. Let $j = (j_1, ..., j_p)$, where $j_1, ..., j_p \geq 0$ are integers, be a multi-index. Define

$$| j | = \sum_{k=1}^{p} j_k .$$

For any function $\nu(x_1, ..., x_p) : [0,1]^p \to \mathbb{R}$ define

$$D_j \nu(x_1, ..., x_p) = \frac{\partial^{|j|} \nu(x_1, ..., x_p)}{\partial x_1^{j_1} ... \partial x_p^{j_p}}$$

whenever the derivative exists. Define $D_0 \nu = \nu$. Define the norm

$$\|\nu\|_s^2 = \sum_{|j| \leq s} \int_{[0,1]^p} [D_j \nu(x_1, ..., x_p)]^2 \, dx_1 ... dx_p$$

The multivariate version of $\mathcal{H}_s$ is

$$\mathcal{H}_s = \{\nu : [0,1]^p \to \mathbb{R} : \|\nu\|_s \le C_0\}.$$

The results of Theorem 4.1 hold for the multivariate model with the multivariate versions of the basis functions and $\mathcal{H}_s$ in place of the univariate versions in the assumptions.

The multivariate versions of H_0 and H_1 are the same as in Section 4.2 but with the multivariate version of $\mathcal{H}_s$. The test statistic is formed as in Section 4.3 but with a kernel function of a $2p$-dimensional argument in (4.3). The obvious multivariate generalizations of Theorems 4.2-4.5 hold after modifying assumption 8(i) as follows:

Assumption 8 (i)′ If $r < \infty$, the bandwidth, h, satisfies $h = c_h n^{-1/(2r+2p)}$, where c_h is a constant, $0 < c_h < \infty$. The truncation parameter J_n satisfies $J_n = C_J n^{\gamma}$ for constants $C_J < \infty$ and $1/(2r + 2s + 1) < \gamma < r/[(2r + 1)(r + p)]$.

Assumption 8(i)′ requires $s > p[1 + 1/(2r)]$, so g must be smoother when p is large than when p is small. Thus, the multivariate test has a form of the curse of dimensionality.

REFERENCES

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica*, 71, 1795-1844.

Bhatia, R., C. Davis, and A. McIntosh (1983). Perturbation of spectral subspaces and solution of linear operator equations, *Linear Algebra and Its Applications*, 52/53, 45-67.

Blundell, R., X. Chen, and D. Kristensen (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves, *Econometrica*, 75, 1613-1669.

Chen, X. and D. Pouzo (2008). Estimation of nonparametric conditional moment models with possibly nonsmooth moments, working paper, Department of Economics, Yale University, New Haven, CT.

Chen, X. and M. Reiss (2007). On rate optimality for ill-posed inverse problems in econometrics, *Econometric Theory*, forthcoming.

Chernozhukov, V., P. Gagliardini, and O. Scaillet (2008). Nonparametric instrumental variable estimation of quantile structural effects, working paper, Department of Economics, Massachusetts Institute of Technology, Cambridge, MA.

Darolles, S., J.-P. Florens, and E. Renault (2006): Nonparametric instrumental regression, Working paper, GREMAQ, University of Social Science, Toulouse, France.

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

Gasser, T. and H.G. Müller (1979). Kernel Estimation of Regression Functions, in *Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics*, 757, 23-68. New York: Springer.

Gasser, T. and H.G. Müller, and V. Mammitzsch (1985). Kernels and Nonparametric Curve Estimation, *Journal of the Royal Statistical Society Series B*, 47, 238-252.

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

Hall, P. and J.L. Horowitz (2005): Nonparametric methods for inference in the presence of instrumental variables, *Annals of Statistics*, 33, 2904-2929.

Horowitz, J.L. (2006). Testing a parametric model against a nonparametric alternative with identification through instrumental variables, *Econometrica*, 521-538.

Horowitz, J.L. and S. Lee (2007). Nonparametric instrumental variables estimation of a quantile regression model, *Econometrica*,75, 1191-1208.

Horowitz, J.L. and V.G. Spokoiny (2001). An adaptive rate-optimal test of a parametric mean-regression model against a nonparametric alternative, *Econometrica*, 69, 599-631.

Jennrich, R.I. (1969). Asymptotic properties of non-linear least squares estimators, *Annals of Mathematical Statistics*, 40, 633-643.

Kress, R. (1999). *Linear Integral Equations*, 2nd edition, New York: Springer-Verlag.

Newey, W.K. and J.L. Powell (2003): Instrumental variable estimation of nonparametric models, *Econometrica*, 71, 1565-1578.

Newey, W.K., J.L. Powell, and F. Vella (1999): Nonparametric estimation of triangular simultaneous equations models, *Econometrica*, 67, 565-603.

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

van de Geer, S. (2000). *Empirical Processes in M-Estimation*, Cambridge, U.K.: Cambridge University Press.

TABLE 1: RESULTS OF MONTE CARLO EXPERIMENTS

g under H_0	d	Empirical Rejection Probability
	H_0 True	
g_1	n.a	0.053
g_2	n.a	0.070
	H_0 False	
g_1	0.02	0.895
g_1	0.005	0.795
g_2	0.02	0.896
g_2	0.005	0.799

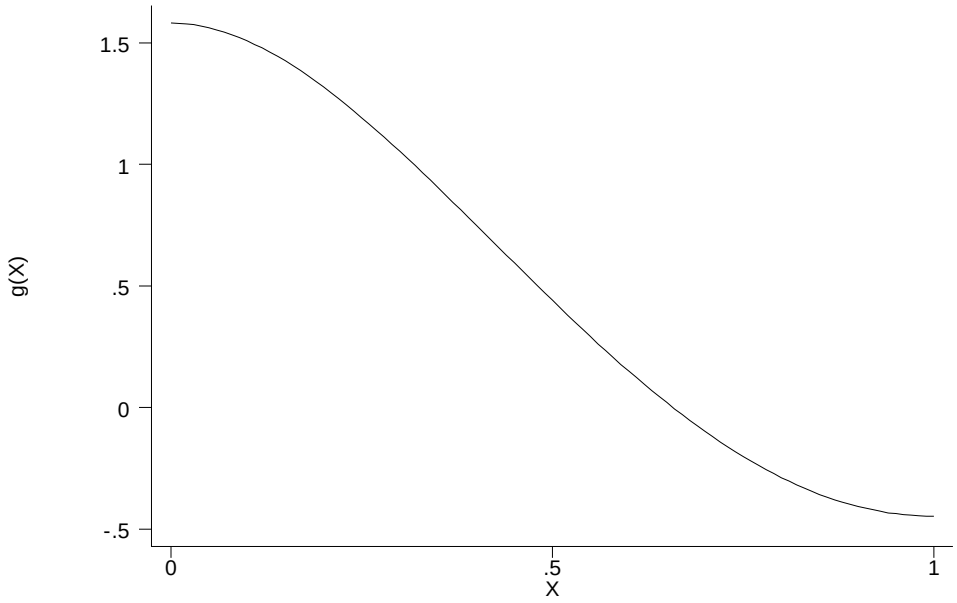

Figure 1a: Graph of $g_1(x)$

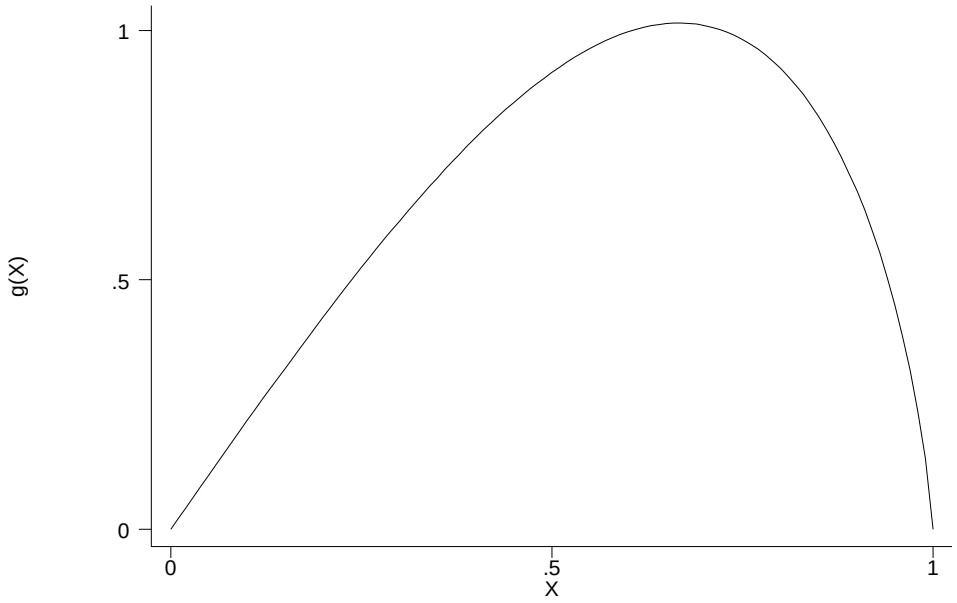

Figure 1b: Graph of $g_2(x)$