APPLIED NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION

by

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

March 2010

ABSTRACT

Instrumental variables are widely used in applied econometrics to achieve identification and carry out estimation and inference in models that contain endogenous explanatory variables. In most applications, the function of interest (e.g., an Engel curve or demand function) is assumed to be known up to finitely many parameters (e.g., a linear model), and instrumental variables are used identify and estimate these parameters. However, linear and other finite-dimensional parametric models make strong assumptions about the population being modeled that are rarely if ever justified by economic theory or other *a priori* reasoning and can lead to seriously erroneous conclusions if they are incorrect. This paper explores what can be learned when the function of interest is identified through an instrumental variable but is not assumed to be known up to finitely many parameters. The paper explains the differences between parametric and nonparametric estimators that are important for applied research, describes an easily implemented nonparametric instrumental variables estimator, and presents empirical examples in which nonparametric methods lead to substantive conclusions that are quite different from those obtained using standard, parametric estimators.

Key words: Nonparametric estimation, instrumental variable, ill-posed inverse problem, endogenous variable, eigenvalues, linear operator

JEL Listing: C12, C13, C14

This article is based on the Fisher-Schultz Lecture that I presented at the 2008 Econometric Society European Meeting. I thank Richard Blundell for providing data from the Family Expenditure Survey, Xiaohong Chen and Charles F. Manski for comments and suggestions, and Brendan Kline for research assistance. This research was supported in part by NSF grants SES-0352675 and SES-0817552.

APPLIED NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION

1. INTRODUCTION

Instrumental variables are widely used in applied econometrics to achieve identification and carry out estimation and inference in models that contain endogenous explanatory variables. In most applications, the function of interest (e.g., an Engel curve or demand function) is assumed to be known up to finitely many parameters (e.g., a linear model), and instrumental variables are used identify and estimate these parameters. However, linear and other finite-dimensional parametric models make strong assumptions about the population being modeled that are rarely if ever justified by economic theory or other *a priori* reasoning and can lead to seriously erroneous conclusions if they are incorrect. This paper explores what can be learned when the function of interest is identified through an instrumental variable but is not assumed to be known up to finitely many parameters.

Specifically, this paper is about estimating the unknown function g in the model

(1.1) $$Y = g(X) + U;\ E(U \mid W = w) = 0$$

for all w or, equivalently,

(1.2) $$E[Y - g(X) \mid W = w] = 0.$$

In this model, g is a function that satisfies regularity conditions but is otherwise unknown, Y is a scalar dependent variable, X is an explanatory variable or vector that may be correlated with U (that is, X may be endogenous), W is an instrument for X, and U is an unobserved random variable. For example, if Y is a household's expenditure share on a good or service and X is the household's total expenditure, then g is an Engel curve. If income from wages and salaries is not influenced by household budgeting decisions, then the household head's total earnings from wages and salaries can be used as an instrument, W, for X (Blundell, Chen, and Kristensen 2007; Blundell and Horowitz 2007). The data used to estimate g are an independent random sample of (Y, X, W).

If some explanatory variables are exogenous, it is convenient to use notation that distinguishes between endogenous and exogenous explanatory variables. We write the model as

(1.3) $$Y = g(X, Z) + U;\ E(U \mid W = w, Z = z) = 0$$

or

(1.4) $$E[Y - g(X, Z) \mid W = w, Z = z] = 0$$

for all w and z. In this model, X denotes the explanatory variables that may be endogenous, Z denotes the exogenous explanatory variables, and W is an instrument for X. The data are an independent random sample of (Y, X, W, Z).

Methods for estimating g in (1.1)-(1.2) and, to a lesser extent, (1.3)-(1.4) have become available recently but have not yet been used much in applied research. This paper explores the usefulness of nonparametric instrumental variables (IV) estimators for applied econometric research. Among other things, the paper:

1. Explains that nonparametric and parametric estimators differ in ways that are important for applied research. Nonparametric estimation is not just a flexible form of parametric estimation.

2. Presents an estimator of g in (1.1)-(1.2) that is as easy to compute as an IV estimator of a linear model. Thus, computational complexity is not a barrier to the use of nonparametric IV estimators in applications.

3. Presents empirical examples in which nonparametric methods lead to substantive conclusions that are quite different from those obtained using standard, parametric estimators.

Some characteristics of nonparametric IV methods may be unattractive to applied researchers. One of these is that nonparametric IV estimators can be very imprecise. This is not a defect of the estimators. Rather, it reflects the fact that the data often contain little information about g when it is identified through instrumental variables. When this happens, applied researchers may prefer to add "information" in the form of *a priori* assumptions about the functional form of g in order to increase the precision of the estimates. For example, g may be assumed to be a linear or quadratic function. However, the improvement in apparent precision obtained from a parametric model carries the risk of misleading inference if the model is misspecified. There is no assurance that a parametric model that is chosen for analytic or computational convenience or because of frequent use in the literature contains the true g or even a good approximation to it. Moreover, neither economic theory nor econometric procedures can lead one reliably to a correct parametric specification. Depending on the substantive meaning of g (e.g., a demand function), economic theory may provide information about its shape (e.g., convex, concave, monotonic) or smoothness, but theory rarely if ever provides a parametric model. The risk of specification error cannot be eliminated through specification testing. Failure to reject a parametric model in a specification test does not necessarily imply that the model is correctly specified. In fact, a specification test may accept several parametric models that yield different substantive conclusions. Nonparametric estimation reveals the information that is

available from the data as opposed to functional form assumptions. It enables one to assess the importance of functional form assumptions in drawing substantive conclusions from a parametric model. Even if an applied researcher ultimately decides to use a parametric model, he or she should be aware of the conclusions that are justified under the weak assumptions of nonparametric estimation and of how these conclusions may differ from those obtained from the parametric model.

Another possible obstacle to the use of nonparametric IV in applications is that certain methodological problems are not yet solved. Some of these problems are outlined later in this paper. It is likely that the problems will be solved in the near future and will not present serious long-run obstacles to applied nonparametric IV estimation.

1.1 Summary of Recent Literature

Nonparametric estimation of g in (1.1)-(1.2) when X and W are continuously distributed has been the object of much recent research. Several estimators are now available, and much is known about the properties of some of them. The available estimators include kernel-based estimators (Darolles, Florens and Renault 2006; Hall and Horowitz 2005) and series or sieve estimators (Newey and Powell 2003; Blundell, Chen, and Kristensen 2007). The estimator of Hall and Horowitz (2005) also applies to model (1.3)-(1.4). The estimators Hall and Horowitz (2005) and Blundell, Chen, and Kristensen (2007) converge in probability at the fastest possible rates under their assumptions (Hall and Horowitz 2005; Chen and Reiss 2007), so these estimators are the best possible in that sense. Horowitz (2007) has given conditions under which the Hall-Horowitz (2005) estimator is asymptotically normally distributed. Horowitz and Lee (2009) show how to obtain uniform confidence bands for g in (1.1)-(1.2). Horowitz (2006) shows how to test a parametric specification for g (e.g., the hypothesis that g is a linear function) against a nonparametric alternative, and Blundell and Horowitz (2007) show how to test the hypothesis that X is exogenous. Horowitz (2009a) shows how to test the hypothesis that a function g satisfying (1.1)-(1.2) exists.

There are also estimators for a quantile version of (1.1)-(1.2) with continuously distributed X and W (Chen and Pouzo 2008; Chernozhukov, Imbens, and Newey 2007; Horowitz and Lee 2007). In the quantile model, the conditional moment restriction $E(U \mid W = w) = 0$ is replaced by a conditional quantile restriction. The resulting model is

$$Y = g(X) + U; \; P(U \le 0 \mid W = w) = q \tag{1.5}$$

for some q such that $0 < q < 1$. Horowitz and Lee (2007) show that this model subsumes the non-separable model

(1.6) $Y = g(X, U)$,

where U is independent of the instrument, W. Chernozhukov and Hansen (2005) and Chernozhukov, Imbens, and Newey (2007) give conditions under which g is identified in (1.5) or (1.6).

When X and W are discretely distributed, as happens in many applications, g is not identified except in special cases. However, informative bounds on g may be identified, even if g is not identified. Manski and Pepper (2000) and Chesher (2004, 2005) give conditions under which informative identified bounds are available.

1.2 The Control Function Model

The control function model is an alternative formulation of the nonparametric IV estimation problem that is non-nested with the formulation of Sections 1.1-1.2. In the control function model,

(1.7) $Y = g(X) + U$

and

(1.8) $X = h(W) + V$,

where g and h are unknown functions,

(1.9) $E(V \mid W = w) = 0$

for all w, and

(1.10) $E(U \mid X = x, V = v) = E(U \mid V = v)$

for all x and v. Assuming that the mean of X conditional on W exists, (1.8) and (1.9) can always be made to hold by setting $h(w) = E(X \mid W = w)$. Identification in the control function approach comes from (1.10). It follows from (1.7) and (1.10) that

(1.11) $E(Y \mid X = x, V = v) = g(x) + k(v)$,

where g and k are unknown functions. If V were observable, g could be estimated by using any of a variety of estimators for nonparametric additive models. See, for example, Horowitz (2009b, Ch. 3). Although V is not observable, it can be estimated consistently by the residuals from nonparametric estimation of h in (1.8). The estimated V can be used in place of the true one for purposes of estimating g from (1.11). Newey, Powell, and Vella (1999) present an estimator and give conditions under which it is consistent and achieves the optimal nonparametric

rate of convergence. Further discussion of the control function approach is available in Pinkse (2000) and Blundell and Powell (2003).

Models (1.1)-(1.2) and (1.7)-(1.10) are non-nested. It is possible for (1.2) to be satisfied but not (1.10) and for (1.10) to be satisfied but not (1.2). Therefore, neither model is more general than the other. Blundell and Powell (2003) and Heckman and Vytlacil (2007) discuss the relative merits of the two models in various settings. At present, there is no statistical procedure for distinguishing empirically between the two models. This paper is concerned mainly with estimation of g in models (1.1)-(1.2) and (1.3)-(1.4). A version of the control function approach will be discussed in Section 6.1 in connection with models in which X and W are discrete. In other respects, the control function approach will not be discussed further.

The remainder of the paper is organized as follows. Section 2 deals with the question of whether there is any important difference between a nonparametric estimator of g and a sufficiently flexible parametric one. Section 3 summarizes the theory of nonparametric estimation of g when X and W are continuous random variables. Section 4 presents a nonparametric estimator that is easy to compute. Section 5 presents empirical examples that illustrate the methods and conclusions of Sections 2-4. Section 6 discusses identification and, when possible, estimation of g when X and W are discrete random variables. Section 7 presents concluding comments. The exposition in this paper is informal. The emphasis is on conveying ideas and important results, not on technical details. Proofs and other details of mathematical rigor are available in the cited reference material.

2. THE DIFFERENCE BETWEEN PARAMETRIC AND NONPARAMETRIC METHODS

If g in (1.1) were known up to a finite-dimensional parameter θ, (that is, $g(x) = G(x,\theta)$ for all x, some known function G and some finite-dimensional θ), then $n^{-1/2}$-consistent, asymptotically normal estimators of θ and g could be obtained by using the generalized method of moments (GMM) (Hansen 1982). When g is unknown, one can consider approximating it by a finite-dimensional parametric model, $G(x,\theta)$, for some suitable G.

It is easy to find functions G that yield good approximations. Engel curves, demand functions, and many other functions that are important in economics are likely to be smooth. They are not likely to be wiggly or discontinuous. A smooth function on a compact interval can be approximated arbitrarily well by a polynomial of sufficiently high degree. Thus, for example, if X is a scalar random variable with compact support, we can write

(2.1) $$g(x) \approx \theta_0 + \theta_1 x + ... + \theta_K x^K$$

$$\equiv G_1(x, \theta),$$

where $K > 0$ is an integer, $\theta_0, \ldots, \theta_K$ are constants, and $\theta = (\theta_0, ..., \theta_K)'$. The approximation error can be made arbitrarily small by making K sufficiently large. Alternatively, one can use a set of basis functions $\{\psi_j : j = 1, 2, ...\}$, such as trigonometric functions, orthogonal polynomials, or splines in place of powers of x. In this case,

(2.2) $$g(x) \approx \theta_1 \psi_1(x) + ... + \theta_K \psi_K(x)$$

$$= G_2(x, \theta).$$

Again, the approximation error can be made arbitrarily small by making K sufficiently large. The parameter vector θ in either (2.1) or (2.2) can be estimated by GMM based on the approximate moment condition $E[G(X, \theta) | W = w)] = 0$. The parameter estimates are $n^{-1/2}$-consistent and asymptotically normal. As will be discussed further in Section 3, nonparametric series estimators of g are based on estimating θ in G_2 for some set of basis functions $\{\psi_j\}$. Therefore, it is possible for parametric and nonparametric estimates to be identical. This makes it reasonable to ask whether there is any practical difference between a nonparametric estimator and a sufficiently flexible parametric one.

The answer is that parametric and nonparametric estimators lead to different inference (confidence intervals and hypothesis tests). This is because inference based on a parametric model treats the model as if it were exact, whereas nonparametric estimation treats it as an approximation that depends on the size of the sample. Specifically, in nonparametric estimation, the "size" of the model (e.g., K in (2.2)) is larger with large samples than with small ones. Consequently, the approximation error is smaller with large samples than with small ones. In contrast, the size (or dimension) of a parametric model is fixed and independent of the sample. Although it is possible to find a parametric model that coincides with a nonparametric model, a given parametric model coincides with a nonparametric model only for a narrow range of sample sizes. This makes inference based on parametric and nonparametric models different because the two models are different except when the sample size is in a small range that depends on the details of the estimation problem. As an analogy, it may be useful to consider the difference between estimates based on random and arbitrary samples. One can always find an arbitrary sample that coincides with a random sample, but a given arbitrary sample is unlikely to coincide with a random one. Therefore, estimates obtained from a given arbitrary sample and a random

sample are different except in the unlikely event that the two coincide. The dependence of a nonparametric model on the sample size makes it important to have a good way of choosing the size of the model (e.g., K in (2.2)). Section 4.2 outlines a method for doing this in nonparametric IV estimation based on a series approximation.

Because parametric estimation ignores approximation error and the dependence of the approximating model on the sample size, a parametric estimate tends to give a misleading indication of estimation precision unless the parametric model is really correct. Parametric methods typically indicate that the estimates are more precise than they really are. Often the assumptions of a highly restrictive parametric model are much more "informative" than the data are. Consequently, conclusions that are supported by the parametric model may not be supported by nonparametric methods. This is illustrated by empirical examples that are presented in Sections 5 and 6.

3. NONPARAMETRIC IV ESTIMATION WHEN X AND W ARE CONTINUOUSLY DISTRIBUTED

This section summarizes the theory of nonparametric IV estimation and explains why nonparametric IV estimation presents problems that are not present in parametric IV estimation. The discussion is concerned with estimating g in model (1.1)-(1.2) when X and W are continuously distributed scalars. Allowing X and W to be vectors complicates the notation but does not change the essential ideas or results, though it may reduce estimation precision owing to curse-of-dimensionality effects. It is assumed that the support of (X,W) is contained in $[0,1]^2$. This assumption can always be satisfied by, if necessary, carrying out monotone increasing transformations of X and W. For example, one can replace X and W by $\Phi(X)$ and $\Phi(W)$, where Φ is the normal distribution function.

3.1 Identification

We begin by deriving a mapping from the population distribution of (Y,X,W) to g. This mapping identifies g and provides the starting point for estimation of g.

Let $f_{X|W}$ denote the probability density function of X conditional on W. Let f_{XW} and f_W, respectively, denote the probability density functions of (X,W) and W. Note that $f_{X|W} = f_{XW} / f_W$. Model (1.1)-(1.2) can be written

$$E(Y \mid W = w) = E[g(X) \mid W = w]$$

$$= \int_0^1 g(x) f_{X|W}(x, w) dx$$

$$= \int_0^1 g(x) \frac{f_{XW}(x, w)}{f_W(w)} dx. \tag{3.1}$$

Therefore,

$$E(Y \mid W = w) f_W(w) = \int_0^1 g(x) f_{XW}(x, w) dx \tag{3.2}$$

and

$$E(Y \mid W = w) f_{XW}(z, w) f_W(w) = \int_0^1 g(x) f_{XW}(x, w) f_{XW}(z, w) dx \tag{3.3}$$

for any $z \in [0,1]$. Define

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

Then integrating with respect to w on both sides of (3.3) yields

$$E[Y f_{XW}(z, W)] = \int_0^1 g(x) t(x, z) dx \tag{3.4}$$

for any $z \in [0,1]$, where the expectation on the left-hand side is over the distribution of (Y, W). Equation (3.4) shows that g is the solution to an integral equation. The integral equation is called a Fredholm equation of the first kind in honor of the Swedish mathematician Erik Ivar Fredholm.

Now define the operator (that is, mapping from one set of functions to another) T by

$$(Th)(z) = \int_0^1 h(x) t(x, z) dx.$$

Define $r(z) = E[Y f_{XW}(z, W)]$. Then (3.4) is equivalent to the operator equation

$$r(z) = (Tg)(z). \tag{3.5}$$

It may be useful to think of T as the infinite-dimensional generalization of a matrix and (3.5) as the infinite-dimensional generalization of a system of simultaneous equations. Assume that T is non-singular or one-to-one.[1] That is, if $Th = 0$, then $h = 0$ almost everywhere. Then T has an inverse, and the solution to (3.5) is

[1] Blundell, Chen, and Kristensen (2007) give examples of distributions that satisfy the non-singularity condition. There has been little research on what can be learned about g when X

(3.6) $g(x) = (T^{-1}r)(x)$.

Equation (3.6) is the desired mapping from the population distribution of (Y, X, W) to g. Equation (3.6) identifies g and can be used to form estimators of g.

3.2 Background from Functional Analysis

The properties of estimators of g depend on those of T.[2] Stating the relevant properties of T requires the use of concepts and results from functional analysis. These are infinite-dimensional analogs of similar concepts and results for finite-dimensional vectors and matrices and will be stated briefly here. Mathematical details can be found in textbooks on functional analysis, such as Conway (1990) and Liusternik and Sobolev (1961).

Define the function space $L_2[0,1]$ as the set of functions that are square integrable on $[0,1]$. That is,

$$L_2[0,1] = \left\{ h : \int_0^1 h(x)^2 dx < \infty \right\} .$$

Define the norm, $\|h\|$ of any function $h \in L_2[0,1]$ by

$$\|h\| = \left[\int_0^1 h(x)^2 dx \right]^{1/2} .$$

For any functions $h_1, h_2 \in L_2[0,1]$, define the inner product

$$\langle h_1, h_2 \rangle = \int_0^1 h_1(x) h_2(x) dx .$$

Let $\{\lambda_j, \phi_j : j = 1, 2, ...\}$ denote the eigenvalues and eigenvectors of T. These are the solutions to the equation

$$T\phi_j = \lambda_j \phi_j ; \quad j = 1, 2, ...$$

and are analogous to the eigenvalues and eigenvectors of a real, symmetric matrix. T is always positive semidefinite or definite and is assumed to be non-singular, so $\lambda_j > 0$ for all $j = 1, 2, ...$ Sort the eigenvalues and eigenvectors so that $\lambda_1 \geq \lambda_2 \geq ... > 0$.

---

and W are continuously distributed and T is singular. Section 6 reviews research on what can be learned about g when X and W are discrete and the discrete analog of T is singular.

[2] The investigation of properties of estimators of g can also be based on (3.1) or (3.2). The conclusions are the same as those obtained using (3.4)-(3.6), and the necessary mathematical tools are simpler with (3.4)-(3.6). If X is exogenous and $W = X$, then $f_{XW}(x, w) = f_W(w)\delta(x - w)$, where δ is the Dirac delta function. The delta function in f_{XW} changes the properties of T, and the results of Sections 3-4 of this paper no longer apply.

Assume that

$$\int_0^1\int_0^1 f_{XW}(x,w)^2\,dxdw < \infty .$$

Then eigenvalues and eigenvectors of T have the following properties:

1. Zero is a limit point of the eigenvalues. Therefore, there are infinitely many λ_j's within any neighborhood of zero. Zero is the only limit point of the eigenvalues.

2. The eigenvectors are orthonormal. That is $\langle \phi_j, \phi_k \rangle = 1$ if $j = k$ and 0 otherwise.

3. The eigenvectors are a basis for $L_2[0,1]$. That is, any function $h \in L_2[0,1]$ has the series representation

$$h(x) = \sum_{j=1}^{\infty} h_j \phi_j(x),$$

where $h_j = \langle h, \phi_j \rangle$. Moreover,

$$\|h\|^2 = \sum_{j=1}^{\infty} h_j^2 .$$

4. For any $h \in L_2[0,1]$,

$$(Th)(x) = \sum_{j=1}^{\infty} \lambda_j h_j \phi_j(x).$$

In addition, if

$$\sum_{j=1}^{\infty}\left(\frac{h_j}{\lambda_j}\right)^2 < \infty,$$

then

$$(T^{-1}h)(x) = \sum_{j=1}^{\infty} \frac{h_j}{\lambda_j}\phi_j(x).$$

Because of property 3, we can write

$$r(z) = \sum_{j=1}^{\infty} r_j \phi_j(z)$$

and

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

where $r_j = \langle r, \phi_j \rangle$ and $g_j = \langle g, \phi_j \rangle$ for each j. The coefficients r_j and g_j are called generalized Fourier coefficients of r and g, respectively. Because of property 4,

$$(3.7) \qquad (T^{-1}r)(x) = \sum_{j=1}^{\infty} \frac{r_j}{\lambda_j} \phi_j(x)$$

if

$$(3.8) \qquad \sum_{j=1}^{\infty} \left(\frac{r_j}{\lambda_j} \right)^2 < \infty .$$

Combining (3.6) and (3.7) yields the result that

$$(3.9) \qquad g(x) = \sum_{j=1}^{\infty} \frac{r_j}{\lambda_j} \phi_j(x)$$

if (3.8) holds. Equation (3.9) provides a representation of g that can be used to investigate the properties of estimators.

3.3 The Ill-Posed Inverse Problem

The key fact about (3.9) that makes nonparametric IV different from parametric IV is that because $\lambda_j \to 0$ as $j \to \infty$, g is not a continuous functional of r. To see this, let r_1 and r_2 be functions in $L_2[0,1]$ with the representations

$$r_1(x) = \sum_{j=1}^{\infty} r_{1j} \phi_j(x)$$

and

$$r_2(x) = \sum_{j=1}^{\infty} r_{2j} \phi_j(x) .$$

Define

$$g_1(x) = \sum_{j=1}^{\infty} \frac{r_{1j}}{\lambda_j} \phi_j(x)$$

and

$$g_2(x) = \sum_{j=1}^{\infty} \frac{r_{2j}}{\lambda_j} \phi_j(x) .$$

Then

$$\|r_2 - r_1\| = \left[\sum_{j=1}^{\infty} (r_{1j}^2 - r_{2j}^2) \right]^{1/2},$$

and

$$\|g_2 - g_1\| = \left[\sum_{j=1}^{\infty} \left(\frac{r_{1j}^2 - r_{2j}^2}{\lambda_j^2} \right) \right]^{1/2}.$$

Given any $\varepsilon > 0$, no matter how small, and any $M > 0$, no matter how large, it is possible to choose the r_{1j}'s and r_{2j}'s such that $\|r_1 - r_2\| < \varepsilon$ and $\|g_1 - g_2\| > M$. Therefore, an arbitrarily small change in r in (3.5) can produce an arbitrarily large change in g. This phenomenon is called the *ill-posed inverse problem*. The ill-posed inverse problem also arises in deconvolution and nonparametric density estimation (Härdle and Linton 1994, Horowitz 2009b).

The ill-posed inverse problem has important consequences for how much information the data contain about g and how accurately g can be estimated. To see why, denote the data by $\{Y_i, X_i, W_i : i = 1,...,n\}$, where n is the sample size. Suppose that f_{XW} and, therefore, T and the λ_j's, are known. Then the r_j's are the only unknown quantities on the right-hand side of (3.8). It follows from (3.4) and $r_j = \langle r, \phi_j \rangle$ that

$$r_j = E\left[Y \int_0^1 f_{XW}(z, W) \phi_j(z) dz \right]; \quad j = 1, 2, \ldots.$$

Therefore, r_j is a population moment and can be estimated $n^{-1/2}$ consistently by the sample analog

$$\hat{r}_j = n^{-1} \sum_{i=1}^{n} Y_i \int_0^1 f_{XW}(z, W_i) \phi_j(z) dz; \quad j = 1, 2, \ldots$$

The generalized Fourier coefficients of g are estimated consistently and without bias by

$$\hat{g}_j = \frac{\hat{r}_j}{\lambda_j}.$$

Because $\lambda_j \to 0$ as $n \to \infty$, random sampling errors in $\hat{r}_j$ can have large effects on $\hat{g}_j$ when j is large. Indeed, $Var(\hat{g}_j) = Var(\hat{r}_j)/\lambda_j^2 \to \infty$ as $j \to \infty$, except in special cases. As a consequence, except in special cases, only low-order generalized Fourier coefficients of g can be estimated with useful precision with samples of practical size. Thus, the ill-posed inverse problem limits what can be learned about g.

The following example illustrates the problem.

Example 3.1: The ill-posed inverse problem.

Let $g(x) = x$. Let

$$(3.10) \quad f_{XW}(x,w) = \sum_{j=1}^{\infty} \lambda_j^{1/2} \phi_j(x)\phi_j(w);\ 0 \le x, w \le 1,$$

where $\phi_1(z) = 1$, $\phi_j(z) = \sqrt{2}\cos[(j-1)\pi z]$ for $j \ge 2$, $\lambda_1 = 1$, and $\lambda_j = 0.2(j-1)^{-4}$ for $j \ge 2$. With this f_{XW}, the marginal distributions of X and W are uniform on $[0,1]$, but X and W are not independent of one another. The generalized Fourier coefficients of g are $g_1 = 0.5$ and

$$g_j = \sqrt{2}[(-1)^{j-1} - 1][\pi(j-1)]^{-4};\ j \ge 2.$$

The reduced form model is

$$Y = E[g(X)|W] + V$$

$$= \sum_{j=1}^{\infty} g_j [E\phi_j(X)|W] + V,$$

where V is a random variable satisfying $E(V|W=w) = 0$. Now

$$E[\phi_j(X)|W] = \int_0^1 \phi_j(x) \frac{f_{XW}(x,W)}{f_W(W)} dx$$

$$= \int_0^1 \phi_j(x) f_{XW}(x,W) dx,$$

where the last line uses the fact that the marginal distribution of W is $U[0,1]$. By (3.10),

$$\int_0^1 \phi_j(x) f_{XW}(x,W) dx = \lambda_j^{1/2} \phi_j(W).$$

Therefore, the reduced-form model can be written

$$Y = \sum_{j=1}^{\infty} c_j \phi_j(W) + V,$$

where $c_j = g_j \lambda_j^{1/2}$.

Now let $V \sim N(0, 0.01)$ independently of W. With data $\{Y_i, X_i, W_i : i = 1,...,n\}$, the maximum likelihood (and asymptotically efficient) estimator of the c_j's can be obtained by applying ordinary least squares to

$$Y_i = \sum_{j=1}^{\infty} c_j \phi_j(W_i) + V_i;\ i = 1, ..., n .$$

Let $\hat{c}_j$ ($j = 1, 2, ...$) denote the resulting estimates. The maximum likelihood estimator of g_j is $\hat{c}_j / \lambda_j^{1/2}$.

Figure 1 shows a graph of $|g_j|$ and the standard deviation of $\hat{g}_j$ for $n = 1000$. Only the first 4 generalized Fourier coefficients are estimated with useful precision. The standard deviation of $\hat{g}_j$ is much larger than g_j when $j > 4$. ■

The result of Example 3.1 is very general. Except in special cases, only low-order generalized Fourier coefficients of g can be estimated with useful precision with samples of practical size. This is a consequence of the ill-posed inverse problem and is a characteristic of the estimation problem, not a defect of the estimation method. When identification is through the moment condition (1.2), the data contain little information about the higher-order generalized Fourier coefficients of g. Therefore, to obtain a useful estimator of g, one must find a way to avoid the need for estimating higher-order coefficients. Procedures for doing this are called "regularization." They amount to modifying T in a suitable way. The amount of modification is controlled by a parameter (the regularization parameter) and decreases as $n \to \infty$ to ensure consistent estimation. Several regularization methods are available. See Engl, Hanke, and Neubauer (1996); Kress (1999); and Carrasco, Florens, and Renault (2007). In this paper, regularization will be carried out by replacing T with a finite-dimensional approximation. The method for doing this is described in Section 4. Section 3.4 provides the mathematical rationale for the method.

3.4 Avoiding Estimation of Higher-Order Generalized Fourier Coefficients: The Role of Smoothness

One way of avoiding the need to estimate higher-order generalized Fourier coefficients is to specify a low-dimensional parametric model for g. That is, $g(x) = G(x, \theta)$ for some known function G and low-dimensional θ. A parametric model, in effect, specifies high-order coefficients in terms of a few low-order ones, so only a few low-order ones have to be estimated. But the assumption that g has a known parametric form is strong and leads to incorrect inference unless the parametric model is exact or a good approximation to the true g. The parametric model provides no information about the accuracy of the approximation or the effect of

approximation error on inference. Therefore, it is useful to ask whether we can make an assumption that is weaker than parametric modeling but provides asymptotically correct inference.

The assumption that g is smooth in the sense of having one or more derivatives achieves this goal. Assuming smoothness is usually weaker than assuming that g belongs to a known parametric family, because most parametric families used in applied research are subsets of the class of smooth functions. The smoothness assumption is likely to be satisfied by many functions that are important in applied econometrics, including demand functions and Engel curves, so smoothness is not excessively restrictive in a wide variety of applications. Moreover, as will be explained, smoothness provides enough information about higher-order generalized Fourier coefficients to make consistent estimation of g and asymptotically correct inference possible.

We first provide a formal definition of the smoothness concept that will be used for estimating g. Let $D^k g(x) = d^k g(x)/dx^k$ for $k = 0,1,2,...$ with $D^0 g(x) = g(x)$. Define g to have smoothness s if

$$\|g\|_s^2 \equiv \sum_{j=0}^{s} \left\| D^j g \right\|^2 \leq C_0^2$$

for some finite, positive constant C_0. In other words, g has smoothness s if it has s square-integrable derivatives.

To see why smoothness is useful for estimating g, let $\{\psi_j\}$ be a basis for $L_2[0,1]$. The ψ_j's need not be eigenfunctions of T. If g has smoothness $s > 0$ and $\{\psi_j\}$ is any of a variety of bases that includes trigonometric functions, orthogonal polynomials, and splines (see, e.g., Chen 2007), then there are coefficients $\{g_j\}$ and a constant $C < \infty$ not depending on g such that

$$\left\| g - \sum_{j=1}^{J} g_j \psi_j \right\| \leq C J^{-s} \tag{3.11}$$

for each $J = 1,2,...$ Therefore, smoothness provides an upper bound on the error of a truncated series approximation to g. This bound is sufficient to permit consistent estimation of g and asymptotically correct inference. In other words, smoothness makes nonparametric estimation and inference possible.

Although smoothness makes nonparametric estimation of g possible, it does not eliminate the need for judgment in estimation. Depending on the details of g and the basis functions, many generalized Fourier coefficients g_j may be needed to achieve a good

approximation to g. This is a concern because, due to the ill-posed inverse problem, it is possible to estimate only low-order g_j's with useful precision. Therefore, it is desirable to choose basis functions that provide a good low-dimensional approximation to g. This is not the same as parametric modeling because we do not assume that the truncated series approximation is exact and, consequently, the length of the series approximation depends on the sample size. Theoretically justified methods for choosing basis functions in applications are not yet available.

4. NONPARAMETRIC ESTIMATION AND TESTING OF A SMOOTH FUNCTION

Section 4.1 presents an estimator of g in model (1.1)-(1.2). The estimator is extended to model (1.3)-(1.4) in Section 4.2. Section 4.3 describes two specification tests that will be used in the empirical illustrations of Section 5. It is assumed that X, W, and Z are scalar random variables. The extension to random vectors complicates the notation but does not affect the main ideas and results. See Hall and Horowitz (2005); Horowitz (2006, 2009a); Blundell, Chen, and Kristensen (2007); and Blundell and Horowitz (2007).

4.1 Estimation of g in Model (1.1)-(1.2)

This section presents an estimator of g in model (1.1)-(1.2). The estimator is a simplified version of the estimator of Blundell, Chen, and Kristensen (2007). It is analogous to an IV estimator for a linear model and can be computed the same way. The estimator is also a version of the Petrov-Galerkin method for solving a Fredholm integral equation of the first kind (Kress 1999).

To begin the derivation of the estimator, define

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

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

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

Then (3.2) is equivalent to

(4.1) $$Ag = m.$$

The estimator of this section is obtained by replacing A and m with series estimators and solving the resulting empirical version of (4.1).[3]

[3] Equation (3.5) and the results of Section 3 can be obtained from (4.1) by setting $T = A^*A$ and $r = A^*m$, where A^* is the adjoint of A. The eigenvalues λ_j are the singular values of A. The

To obtain the estimators, let $\{\psi_j\}$ be an orthonormal basis for $L_2[0,1]$ that satisfies (3.11). Then we can write

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

$$m(w) = \sum_{j=1}^{\infty} m_j \psi_j(w),$$

and

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

where $g_j = \langle g, \psi_j \rangle$, $m_j = \langle m, \psi_j \rangle$, and

$$a_{jk} = \int_0^1\int_0^1 f_{XW}(x,w)\psi_j(x)\psi_k(w)dxdw.$$

In addition,

$$(Ag)(w) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} g_j a_{jk}\psi_k(w).$$

The m_j's and a_{jk}'s are estimated $n^{-1/2}$ consistently by

$$\hat{m}_j = n^{-1}\sum_{i=1}^{n} Y_i\psi_j(W_i)$$

and

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

The functions m and operator A are estimated consistently by

$$\hat{m}(w) = \sum_{j=1}^{J_n} \hat{m}_j\psi_j(w)$$

and

$$(\hat{A}h)(x) = \sum_{j=1}^{J_n}\sum_{k=1}^{J_n} h_j\hat{a}_{jk}\psi_k(w),$$

---

formulation of Section 3 is useful for expository purposes because it does not require familiarity with the singular value decomposition of an operator. However, (4.1) yields an estimator that is easier to compute.

where h is any function in $L_2[0,1]$, $h_j = \langle h, \psi_j \rangle$, and the integer J_n is a truncation point that increases at a suitable rate as $n \to \infty$.[4] The empirical version of (4.1) is

(4.2) $\hat{A}\hat{g} = \hat{m}$.

The solution to (4.2) has the form of a conventional linear IV estimator. To obtain it, let $\mathcal{W}_n$ and $\mathcal{X}_n$, respectively, denote the $n \times J_n$ matrices whose (i, j) elements are $\psi_j(W_i)$ and $\psi_j(X_i)$. Define $\mathcal{Y}_n = (Y_1, ..., Y_n)'$. Let $\{\hat{g}_j : j = 1, ..., J_n\}$ denote the generalized Fourier coefficients of $\hat{g}$. That is,

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

Define $\hat{G} = (\hat{g}_1, ..., \hat{g}_{J_n})'$. Then the solution to (4.2) is (4.3) with

(4.4) $\hat{G} = (\mathcal{W}_n' \mathcal{X}_n)^{-1} \mathcal{W}_n' \mathcal{Y}_n$.

$\hat{G}$ has the form of an IV estimator for a linear model in which the matrix of variables is $\mathcal{X}_n$ and the matrix of instruments is $\mathcal{W}_n$.

When n is small, $\hat{g}$ in (4.3)-(4.4) can be highly variable. Blundell, Chen, and Kristensen (2007) propose stabilizing $\hat{g}$ by replacing (4.4) with the solution to a penalized least-squares problem. Blundell, Chen, and Kristensen (2007) provide an analytic, easily computed solution to this problem and present the results of numerical experiments on the penalization method's ability to stabilize $\hat{g}$ in small samples.

Horowitz (2009a) derived the rate of convergence in probability of $\hat{g}$. When f_{XW} has $r < \infty$ continuous derivatives with respect to any combination of its arguments and certain other regularity conditions hold, then

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

If $r = \infty$, the rate of convergence is slower, as is discussed below. When $r < \infty$, the rate of convergence of $\|\hat{g} - g\|$ is fastest when the terms J_n^{-s} and $J_n^r (J_n / n)^{1/2}$ converge to zero at the same rate. This happens when $J_n \propto n^{1/(2r+2s+1)}$, which gives

[4] More generally, the series for $\hat{m}$ and $\hat{A}$ or in the x and w directions can use different basis functions and have different lengths. This extension is not carried out here. The effects on estimation efficiency of using different basis functions and series lengths for different functions or directions are unknown at present.

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

Chen and Reiss (2007) show that $n^{-s/(2r+2s+1)}$ is the fastest possible rate of convergence in probability of $\|\hat{g} - g\|$ that is achievable uniformly over functions g and f_{XW} satisfying Horowitz's (2009a) conditions. The rate of convergence in (4.6) is slower than the $n^{-1/2}$ rate that is usually achieved by finite-dimensional parametric models. It is also slower than the rate of convergence of a nonparametric estimator of a conditional mean or quantile function. For example, if $E(Y \mid X = x)$ and the probability density function of X are twice continuously differentiable, then a nonparametric estimator of $E(Y \mid X = x)$ can achieve the rate of convergence $n^{-2/5}$, whereas the rate in (4.6) with $r = s = 2$ is $n^{-2/9}$. A nonparametric IV estimator converges relatively slowly because the data contain little information about g in model (1.1)-(1.2), not because of any defect of the estimator.

In (4.5), the term J_n^{-s} arises from the bias of $\hat{g}$ that is caused by truncating the series approximation (4.3). The truncation bias decreases as s increases and g becomes smoother (see (3.11)). Therefore, increased smoothness of g accelerates the rate of convergence of $\hat{g}$. The term $J_n^r(J_n/n)^{1/2}$ in (4.5) is caused by random sampling errors in the estimates of the generalized Fourier coefficients $\hat{g}_j$. Specifically, $J_n^r(J_n/n)^{1/2}$ is the rate of convergence of in probability of $\left[\sum_{j=1}^{J_n}(\hat{g}_j - g_j)^2\right]^{1/2}$. Because g_j is inversely proportional to λ_j (see the discussion in Section 3), $\left[\sum_{j=1}^{J_n}(\hat{g}_j - g_j)^2\right]^{1/2}$ converges more slowly when the eigenvalues of T converge rapidly than when they converge slowly. When f_{XW} has smoothness r, the eigenvalues decrease at a rate that is at least as fast as j^{-2r} (Pietsch 1980). Therefore, the fastest possible rates of convergence of $\sum_{j=1}^{J_n}(\hat{g}_j - g_j)^2$ and $\|\hat{g} - g\|$ decrease as f_{XW} becomes smoother. Smoothness of f_{XW} increases the severity of the ill-posed inverse problem and reduces the accuracy with which g can be estimated.

When f_{XW} is the bivariate normal density, $r = \infty$ and the eigenvalues of T decrease at the rate e^{-cj}, where $c > 0$ is a constant. The problem of estimating g is said to be severely ill posed, and the rate of convergence of $\|\hat{g} - g\|$ in (4.3) is $O_p[(\log n)^{-s}]$. This is the fastest

possible rate. Therefore, when f_{XW} is very smooth, the data contain very little information about g in (1.1)-(1.2). Unless g is restricted in other ways, such as assuming that it belongs to a low-dimensional parametric family of functions or is infinitely differentiable, a very large sample is needed to estimate g accurately in the severely ill-posed case.

Now let $x_1, x_2, ..., x_L$ be L points in $[0,1]$. Horowitz and Lee (2009) give conditions under which $[\hat{g}(x_1), ..., \hat{g}(x_L)]$ is asymptotically L-variate normally distributed and the bootstrap can be used to obtain simultaneous confidence intervals for $[g(x_1), ..., g(x_L)]$. Horowitz and Lee (2009) also show how to interpolate the simultaneous confidence intervals to obtain a uniform confidence band for g. The bootstrap procedure of Horowitz and Lee (2009) estimates the joint distribution of the leading terms of the asymptotic expansions of $\hat{g}(x_\ell) - g(x_\ell)$ ($\ell = 1, ..., L$). To describe this procedure, let $s_n^2(x_\ell)$ denote the following consistent estimator of the variance of the asymptotic distribution of $\hat{g}(x)$:

$$s_n^2(x) = n^{-2} \sum_{i=1}^{n} \{\hat{A}^{-1}[\delta_n(\cdot, Y_i, X_i, W_i, \hat{g}) - \bar{\delta}_n(\cdot, \hat{g})](x)\}^2 ,$$

where for any $h \in L_2[0,1]$

$$\delta_n(x, Y, X, W, h) = [Y - h(X)] \sum_{k=1}^{J_n} \psi_k(W) \psi_k(x)$$

and

$$\bar{\delta}_n(x, h) = n^{-1} \sum_{i=1}^{n} \delta_n(x, Y_i, X_i, W_i, h) .$$

Let $\{Y_i^*, X_i^*, W_i^* : i = 1, ..., n\}$ be a bootstrap sample that is obtained by sampling the estimation data $\{Y_i, X_i, W_i : i = 1, ..., n\}$ randomly with replacement. The bootstrap version of the asymptotic form of $\hat{g}(x) - g(x)$ is

$$\Delta_n(x) = n^{-1} \sum_{i=1}^{n} \{\hat{A}^{-1}[\delta_n(\cdot, Y_i^*, X_i^*, W_i^*, \hat{g}) - \bar{\delta}_n(\cdot, \hat{g})]\}(x) .$$

Let A_n^* be the estimator of A that is obtained from the bootstrap sample. Define $[s_n^*(x)]^2$ as the following bootstrap estimator of the variance of $\Delta_n(x)$:

$$[s_n^*(x)]^2 = n^{-2} \sum_{i=1}^{n} \{(A_n^*)^{-1}[\delta_n(\cdot, Y_i^*, X_i^*, W_i^*, g^*) - \bar{\delta}_n^*(\cdot, g^*)](x)\}^2 .$$

where g^* is the estimate of g obtained from the bootstrap sample and $\bar{\delta}_n^*(\cdot, g^*) = n^{-1}\sum_{i=1}^n \delta_n(\cdot, Y_i^*, X_i^*, W_i^*, g^*)$. The bootstrap procedure is as follows.

Step 1: Draw a bootstrap sample $\{Y_i^*, X_i^*, W_i^* : i = 1,...,n\}$ by sampling the estimation data $\{Y_i, X_i, W_i : i = 1,...,n\}$ randomly with replacement. Use this sample to form bootstrap estimates $\Delta_n(x_1),...,\Delta_n(x_L)$ and $s_n^*(x_1),...,s_n^*(x_\ell)$. Compute the statistic

$$t_n^* = \max_{1\le \ell \le L} \frac{|\Delta_n(x_\ell)|}{s_n^*(x_\ell)}.$$

Step 2: Repeat step 1 many times. Let M be the number of repetitions and t_{nm}^* be the value of t_n^* obtained on the m'th repetition. Let $\zeta_{n\alpha}^* = \inf\{\zeta : F_M^*(\zeta) \ge \alpha\}$ for any $\alpha \in (0,1)$, where

$$F_M^*(\tau) = M^{-1}\sum_{m=1}^M I(t_{nm}^* \le \tau)$$

and I is the indicator function. Then $\zeta_{n\alpha}^*$ is a consistent estimator of the $1-\alpha$ quantile of the bootstrap distribution of t_n^*.

Step 3: The simultaneous $1-\alpha$ confidence intervals for $[\hat{g}(x_1),...,\hat{g}(x_L)]$ are

$$\hat{g}(x_\ell) - \zeta_{n\alpha}^* s_n(x_\ell) \le g(x_\ell) \le \hat{g}(x_\ell) + \zeta_{n\alpha}^* s_n(x_\ell);\ \ell = 1,...,L.$$

Implementation of the estimator (4.3) requires choosing the value of J_n. One possible choice is an estimator of the asymptotically optimal J_n. The asymptotically optimal J_n, denoted here by $J_{n,opt}$, minimizes $Q_n(J) \equiv E_A\|\hat{g} - g\|^2$, where E_A denotes the expectation with respect to the asymptotic distribution of $\hat{g} - g$. Note that Q_n depends on J through $\hat{g}$. Define $\hat{J}_{n,opt}$ to be an asymptotically optimal estimator of $J_{n,opt}$ if $Q_n(\hat{J}_{n,opt})/Q_n(J_{n,opt}) \to^p 1$ as $n \to \infty$. Horowitz (2010) gives conditions under which an asymptotically optimal estimator of $J_{n,opt}$ can be obtained by minimizing the quantity

$$\hat{Q}_n(J) = n^{-2}\sum_{i=1}^n \left\{ [Y_i - \hat{g}(X_i)]^2 \sum_{j=1}^n \{(\hat{A}^{-1})^* \psi_j](W_i)\}^2 \right\} - \|\hat{g}\|^2.$$

Horowitz (2010) presents Monte Carlo evidence indicating that this estimator performs well with samples of practical size in both mildly and severely ill-posed estimation problems.

4.2 Extension to Model (1.3)-(1.4)

This section extends the estimator (4.3) to model (1.3)-(1.4), which contains the exogenous explanatory variable Z. Assume that Z is a scalar whose support is $[0,1]$. The data are the independent random sample $\{Y_i, X_i, W_i, Z_i : i = 1,...,n\}$.

If Z is discretely distributed with finitely many mass points, then $g(x,z)$, where z is a mass point, can be estimated by using (4.3) with only observations i for which $Z_i = z$. The results of Section 4.1 hold with n replaced by the number of observations for which $Z_i = z$, which is $n_z = \sum_{i=1}^{n} I(Z_i = z)$.

If Z is continuously distributed, then $g(x,z)$ can be estimated by using (4.3) with observations i for which Z_i is "close" to z. Kernel weights can be used to select the appropriate observations. To this end, let K be a kernel function in the sense of nonparametric density estimation or regression, and let $\{b_n\}$ be a positive sequence of bandwidths that converges to 0 as $n \to \infty$. Define $K_b(v) = K(v/b)$ for any real v and b. Also define

$$\hat{m}_{jz} = \frac{1}{nb_n}\sum_{i=1}^{n} Y_i \psi_j(W_i) K_{b_n}(z - Z_i),$$

$$\hat{a}_{jkz} = \frac{1}{nb_n}\sum_{i=1}^{n} \psi_j(X_i)\psi_k(W_i) K_{b_n}(z - Z_i),$$

$$\hat{m}_z(w) = \sum_{j=1}^{J_n} \hat{m}_{jz}\psi_j(w),$$

and

$$\hat{f}_{XWZ}(x,w,z) = \sum_{j=1}^{J_n}\sum_{k=1}^{J_n} \hat{a}_{jkz}\psi_j(x)\psi_k(w).$$

Define the operator $\hat{A}_z$ by

$$(\hat{A}_z h)(w,z) = \int_0^1 h(x)\hat{f}_{XWZ}(x,w,z)dx$$

for any $h \in L_2[0,1]$. Let f_{XWZ} and f_{WZ} denote the probability density functions of (X,W,Z) and (W,Z), respectively. Estimate $g(x,z)$ for any $z \in (0,1)$ by solving

(4.7) $\hat{A}_z \hat{g} = \hat{m}_z$.

This is a finite-dimensional matrix equation because $\hat{A}_z$ is a $J_n \times J_n$ matrix and $\hat{m}_z$ is a $J_n \times 1$ vector. Equation (4.7) is an empirical analog of the relation

(4.8) $E(Y \mid W = w, Z = z) f_{WZ}(w,z) = (A_z g)(w,z)$,

where the operator A_z is defined by

$$(A_z h)(w,z) = \int_0^1 h(x,z) f_{XWZ}(x,w,z) dw .$$

Equation (4.8) can be derived from (1.3)-(1.4) by using reasoning like that used to obtain (3.6).

Under regularity conditions that are stated in the Section A.1 of the appendix,

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

where $\kappa = 2r/(2r+1)$. The estimator can be extended to $z = 0$ and $z = 1$ by using a boundary kernel (Gasser and Müller 1979; Gasser, Müller, and Mammitzsch 1985) in $\hat{m}_{jz}$ and $\hat{a}_{jkz}$. Boundary kernels are explained in the discussion of the second specification test in Section 4.3.

4.3 Two Specification Tests

This section presents two specification tests that will be used in the empirical illustrations of Section 5. One test is of the hypothesis that $g(x,z) = G(x,z,\theta)$ for all $(x,z) \in [0,1]^2$, where G is a known function and θ is a finite-dimensional parameter whose value must be estimated from the data. Under this hypothesis, the parametric model $G(x,\theta)$ satisfies (1.3)-(1.4) for some θ. A similar test applies to (1.1)-(1.2). In this case, the hypothesis is $g(x) = G(x,\theta)$. The second test presented in this section is of the hypothesis that $g(x,z)$ does not depend on x. The first test was developed by Horowitz (2006). The second test is new.

Testing a Parametric Model against a Nonparametric Alternative: In this test, the null hypothesis, H_0, is that

(4.10) $g(x,z) = G(x,z,\theta)$

for a known function G, some finite-dimensional θ in a parameter set Θ, and almost every $(x,z) \equiv [0,1]^2$. "Almost every (x,z)" means every (x,z) except, possibly, a set of (x,z) values whose probability is 0. The alternative hypothesis, H_1, is that there is no $\theta \in \Theta$ such that (4.10) holds for almost every (x,z). The discussion here applies to model (1.3)-(1.4). A test of $H_0 : g(x) = G(x,\theta)$ for model (1.1)-(1.2) can be obtained by dropping z and setting $\ell(x,z) = 1$ in the discussion below. The test statistic is

$$\tau_n = \int_0^1 \int_0^1 S_n^2(x,z) dx dz ,$$

where

$$S_n(x,z) = n^{-1/2}\sum_{i=1}^{n}[Y_i - G(X_i, Z_i, \hat{\theta})]\hat{f}_{XWZ}^{(-i)}(x, W_i, Z_i)\ell(Z_i, z),$$

$\hat{\theta}$ is a GMM estimator of θ and $\hat{f}_{XWZ}^{(-i)}$ is a leave-observation-i-out kernel estimator of f_{XWZ}. That is

$$\hat{f}_{XWZ}^{(-i)}(x,w,z) = \frac{1}{nb_n^3}\sum_{\substack{j=1 \\ j\neq i}}^{n} K_{b_n}(x - X_j)K_{b_n}(w - W_j)K_{b_n}(z - Z_j)$$

where K is a kernel function and b_n is a bandwidth. In applications, the value of b_n can be chosen by cross-validation. The function ℓ is any function on $[0,1]$ with the property that

$$\int_0^1 \ell(x,z)h(x)dx = 0$$

for almost every $z \in [0,1]$ only if $h(x) = 0$ for almost every $x \in [0,1]$. H_0 is rejected if τ_n is too large. Horowitz (2006) derives the asymptotic distribution of τ_n under H_0 and H_1 and gives a method for computing its critical value. The τ_n test is consistent against any fixed alternative model and against a large class of alternative models whose distance from the null-hypothesis parametric model is $O(n^{-1/2})$ or greater.

The test can be understood intuitively by observing that as $n \to \infty$, $n^{-1/2}S_n(x,z)$ converges in probability to

$$S_\infty(x,z) = E_{XWZ}\{[g(X,Z) - G(X,Z,\theta_\infty)]f_{XWZ}(x,W,Z)\ell(Z,z)\},$$

where E_{XWZ} denotes the expectation with respect to the distribution of (X,W,Z) and θ_∞ is the probability limit of θ_n as $n \to \infty$. If g is identified, then $S_\infty(x,z) = 0$ for almost every $(x,z) \in [0,1]^2$ only if $g(x,z) = G(x,z,\theta_\infty)$ almost every (x,z). Therefore,

$$\tau_\infty = \int_0^1\int_0^1 S_\infty(x,z)^2\,dxdz$$

is a measure of the distance between $g(x,z)$ and $G(x,z,\theta_\infty)$. The test statistic τ_n is an empirical analog of τ_∞.

Testing the Hypothesis that *g*(*x*,*z*) Does Not Depend on *x*: This test is a modification of the exogeneity test of Blundell and Horowitz (2007). The null hypothesis, H_0, is that

(4.11) $g(x,z) = G(z)$

for almost every $(x,z)\in[0,1]^2$ and some unknown function G. The alternative hypothesis, H_1, is that there is no G such that (4.11) holds for almost every $(x,z)\in[0,1]^2$. It follows from (1.3)-(1.4) that $G(z)=E(Y\mid Z=z)$ if H_0 is true. Accordingly, we set $G(z)=E(Y\mid Z=z)$ for the rest of the discussion of the test of H_0.

The test statistic is

$$\tilde{\tau}_n=\int_0^1\int_0^1\tilde{S}_n^2(x,z)dx,dz\ ,$$

where

$$\text{(4.12)}\quad \tilde{S}_n(x,z)=n^{-1/2}\sum_{i=1}^{n}[Y_i-\hat{G}^{(-i)}(Z_i)]\hat{f}_{XWZ}^{(-i)}(x,W_i,Z_i)\ell(Z_i,z)\,.$$

In (4.12), ℓ is defined as in the test of a parametric model. $\hat{G}^{(-i)}$ and $\hat{f}_{XWZ}^{(-i)}$, respectively, are leave-observation-*i*-out "boundary kernel" estimators of the mean of Y conditional on Z and f_{XWZ}. Boundary kernels are defined in the next paragraph. The estimators are

$$\hat{f}_{XWZ}^{(-i)}(x,w,z)=$$

$$\frac{1}{nb_1^3}\sum_{\substack{j=1\\ j\neq i}}^{n}K_{b_1}\left(x-X_j,x\right)K_{b_1}\left(w-W_j,w\right)K_{b_1}\left(z-Z_j,z\right)$$

and

$$\hat{G}^{(-i)}(z)=\frac{1}{nb_2\hat{f}_Z^{(-i)}(z)}\sum_{\substack{j=1\\ j\neq i}}^{n}Y_iK_{b_2}\left(z-Z_j,z\right),$$

where b_1 and b_2 are bandwidths, and

$$\hat{f}_Z^{(-i)}(z)=\frac{1}{nb_2}\sum_{\substack{j=1\\ j\neq i}}^{n}K_{b_2}\left(z-Z_j,z\right).$$

In applications, b_1 can be chosen by cross-validation. The value of b_2 can be set at $n^{-7/40}$ times the value obtained by cross-validation.

The boundary kernel function K_b has the property that for all $\xi\in[0,1]$

$$\text{(4.12)}\quad b^{-(j+1)}\int_{\xi}^{\xi+1}u^jK_b(u,\xi)du=\begin{cases}1 & \text{if } j=0\\ 0 & \text{if } j=1.\end{cases}$$

If b is small and ξ is not close to 0 or 1, then we can set $K_b(u,\xi) = K(u/b)$, where K is an "ordinary" order *s* kernel. If ξ is close to 1, then we can set $K_b(u,\xi) = \bar{K}(u/b)$, where $\bar{K}$ is a bounded, compactly supported function satisfying

$$\int_0^\infty u^j \bar{K}(u)du = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } 1 \le j \le s-1. \end{cases}$$

If ξ is close to 0, we can set $K_b(u,\xi) = \bar{K}(-u/b)$. Gasser and Müller (1979) and Gasser, Müller, and Mammitzsch (1985) give examples of boundary kernels. A boundary kernel is used here instead of an ordinary kernel because, to prevent imprecise estimation of G, the probability density function of Z, f_Z, is assumed to be bounded away from 0. This causes $f_Z(z)$ and $f_{XWZ}(x,w,z)$ to be discontinuous at $z = 0$ and $z = 1$. The boundary kernel overcomes the resulting edge effects.

The $\tilde{\tau}_n$ test rejects H_0 if $\tilde{\tau}_n$ is too large. Section A.2 of the appendix gives the asymptotic properties of the test, including the asymptotic distribution of $\tilde{\tau}_n$ under H_0, a method for computing the critical value of the test, and the test's consistency. The $\tilde{\tau}_n$ test can be understood intuitively by observing that as $n \to \infty$, $n^{-1/2}\tilde{S}_n(x,z)$ converges in probability to

$$\tilde{S}_\infty(x,z) = E_{XWZ}\{[g(X,Z) - G_\infty(Z)] f_{XWZ}(x,W,Z)\ell(Z,z)\},$$

where $G_\infty(z) = E(Y \mid Z = z)$. Therefore, $\tilde{\tau}_n$ is an empirical measure of the distance between $g(x,z)$ and $E(Y \mid Z = z)$.

5. EMPIRICAL EXAMPLES

This section presents two empirical examples that illustrate the usefulness of nonparametric IV estimation and how conclusions drawn from parametric and nonparametric IV estimators may differ. The first example is about estimation of an Engel curve. The second is about estimating the effects of class size on students' performances on standardized tests.

5.1 Estimating an Engel Curve

This section shows the result of using the method of Section 4.1 to estimate an Engel curve for food. The data are 1565 household-level observations from the British Family Expenditure Survey. The households consist of married couples with an employed head-of-household between the ages of 25 and 55 years. The model is (1.1)-(1.2). Y denotes a household's expenditure share on food, X denotes the logarithm of the household's total

expenditures, and W denotes the logarithm of the household's gross earnings. Blundell, Chen, and Kristensen (2007) used the Family Expenditure Survey for nonparametric IV estimation of Engel curves. Blundell, Chen, and Kristensen (2007) also report the results of an investigation of the validity of the logarithm of gross earnings as an instrument for expenditures.

The basis functions used here are B-splines with 4 knots. The estimation results are similar with 5 or 6 knots. The estimated Engel curve is shown in Figure 2. The curve is nonlinear and different from what would be obtained with a simple parametric model such as a linear or quadratic model. The τ_n test of Horowitz (2006) that is described in Section 4.3 rejects the hypothesis that the Engel curve is a linear, quadratic, or cubic function ($p < 0.05$ in all cases). Thus, in this example, nonparametric methods reveal an aspect of data (the shape of the Engel curve) that would be hard to detect using conventional parametric models. Of course, with sufficient effort it may be possible to find a simple parametric model that gives a curve similar to the nonparametric one. Although such a parametric model may be a useful way to represent the curve, it could not be used for valid inference for the reasons explained in Section 2.

5.2 The Effect of Class Size on Students' Performances on Standardized Tests

Angrist and Lavy (1999) studied the effects of class size on test scores of 4th and 5th grade students in Israel. Here, I use one of their models for 4th grade reading comprehension and their data to illustrate differences between parametric and nonparametric IV estimation and the effects that parametric assumptions can have on the conclusions drawn from IV estimation. The data are available at http://econ-www.mit.edu/faculty/angrist/data1/data/anglavy99. Angrist's and Lavy's substantive conclusions are based on several different models and methods. The discussion in this section is about one model and is not an evaluation or critique of Angrist's and Lavy's substantive findings, which are more broadly based.

One of the models that Angrist and Lavy (1999) use is

$$Y_{CS} = \beta_0 + \beta_1 X_{CS} + \beta_2 D_{CS} + \nu_S + U_{CS}. \tag{5.1}$$

In this model, Y_{CS} is the average reading comprehension test score of 4th grade students in class C of school S, X_{CS} is the number of students in class C of school S, D_{CS} is the fraction of disadvantaged students in class C of school S, ν_S is a school-specific random effect, and U_{CS} is an unobserved random variable that is independently distributed across schools and classes. X_{CS} is a potentially endogenous explanatory variable. The instrument for X_{CS} is

$$Z_{CS} = E_S / \text{int}[1 + (E_S - 1)/40],$$

where E_S is enrollment in school S. The data consist of observations of 2049 classes that were tested in 1991. The IV estimate of β_1 in (5.1) is -0.110 with a standard error of 0.040 (Angrist and Lavy 1999, Table V). Thus, according to model (5.1), increasing class size has a negative and statistically significant effect on reading comprehension test scores.

The nonparametric version of (5.1) is

(5.2) $$Y_{CS} = g(X_{CS}, D_{CS}) + \nu_S + U_{CS};\ E(\nu_S + U_{CS} \mid Z_{CS}, D_{CS}) = 0.$$

Figure 3 shows the result of using the method of Section 4.2 to estimate g as a function of X_{CS} for $D_{CS} = 1.5$ percent. The basis functions are orthogonal (Legendre) polynomials, the series length is 3, and the bandwidth is $b_n = 1.5$. The solid line in the figure is the estimate of g, and the dots show a bootstrap-based uniform 95% confidence band obtained using the method of Horowitz and Lee (2009). Unobserved school-specific effects, ν_S, were handled by using schools as the bootstrap sampling units. The nonparametrically estimated relation between test scores and class size is nonlinear and non-monotonic, but the confidence band is very wide. Functions that are monotonically increasing and decreasing can fit easily in the band. Moreover, the $\tilde{\tau}_n$ test of Section 4.3 does not reject the hypothesis that test scores are independent of class size ($p > 0.10$). Thus, the data and the instrumental variable assumption, by themselves, are uninformative about the form of any dependence of test scores on class size. This does not necessarily imply that test scores and class sizes are independent. For example, the $\tilde{\tau}_n$ test may not be sufficiently powerful to detect any dependence, or the effects of class size might be obscured by heterogeneity that is not accounted for by D_{CS}. However, the nonparametric model does not support the conclusion drawn from the linear model that increases in class sizes are associated with decreased test scores.

Average derivatives can be estimated more precisely than functions can, so it is possible that an estimator of $E\partial g(X, D \mid D = 1.5)/\partial X$ is more informative about the effects of class size on test scores than is the function $g(x, 1.5)$. The average here is over the distributrion of X conditional on $D = 1.5$. Ai and Chen (2009) provide asymptotic distributional results for nonparametric IV estimators of unconditional average derivatives, but there is no existing theory on nonparametric IV estimation of conditional average derivatives such as $E\partial g(X, D \mid D = 1.5)/\partial X$. To get some insight into whether an estimate of the conditional derivative can clarify the relation between test scores and class size, $E\partial g(X, D \mid D = 1.5)/\partial X$ was estimated by

$$(5.3)\qquad \hat{E}\frac{\partial \hat{g}(X, D \mid D = 1.5)}{\partial X} = \frac{\sum_{C,S} \frac{\partial \hat{g}(X_{CS}, 1.5)}{\partial X} K_{b_n}(D_{CS} - 1.5)}{\sum_{C,S} K_{b_n}(D_{CS} - 1.5)}.$$

The standard error of the estimate was obtained by applying the bootstrap to the leading term of the asymptotic expansion of the right-hand side of (5.3) with schools as the bootstrap sampling units. The resulting estimate of the conditional average derivative is 0.064 with a standard error of 0.14. Therefore, the nonparametric average derivative estimate does not support the conclusion from the linear model that increases in class size are associated with decreases in test scores.

The conclusions drawn from the linear model might be persuasive, nonetheless, if this model were consistent with the data. However, the τ_n test of Section 4.3 rejects the hypothesis that g is a linear function of X_{CS} and D_{CS} ($p < 0.05$). This does not necessarily imply that the linear model is a poor approximation g in (5.2), but the quality of the approximation is unknown. Therefore, one should be cautious in drawing conclusions from the linear model. In summary, the data are uninformative about the dependence, if any, of g in (5.2) on X_{CS}. The conclusion from (5.1) that increases in class size decrease test scores is a consequence of the linearity assumption, not of information contained in the data *per se*.

6. DISCRETELY DISTRIBUTED EXPLANATORY VARIABLES AND INSTRUMENTS

This section is concerned with identification and estimation of g when, as happens in many applications, X, W, and Z are discretely distributed random variables with finitely many points of support. Because Z is exogenous and discrete, all of the analysis can be carried out conditional on Z being held fixed at one of its points of support. Accordingly, the discussion in this section is concerned with identifying and estimating g as a function of X at a fixed value of Z. The notation displays dependence only on X and W. Section 6.1 discusses identification and estimation of g. Section 6.2 presents empirical illustrations of the results of Section 6.1.

6.1 Identification and Estimation of g

Let the supports of X and W, respectively, be $\{x_1, ..., x_J\}$ and $\{w_1, ..., w_K\}$ for finite, positive integers J and K. For $j = 1, ..., J$ and $k = 1, ..., K$, define $g_j = g(x_j)$,

$m_k = E(Y \mid W = w_k)$, and $\pi_{jk} = P(X = x_j \mid W = w_k\}$. When X and W are discretely distributed, condition (1.2) is equivalent to

(6.1) $\quad m_k = \sum_{j=1}^{J} \pi_{jk} g_j;\ k = 1,...,K$.

Let Π be the $J \times K$ matrix whose (j,k) element is π_{jk}. If $K \geq J$ and $Rank(\Pi) = J$, then (6.1) can be solved to obtain

(6.2) $\quad g = (\Pi\Pi')^{-1}\Pi M$,

where $M = (m_1,...,m_K)'$ and $g = (g_1,...,g_J)'$.

An estimator of g that is $n^{-1/2}$-consistent and asymptotically normal can be obtained by replacing Π and M in (6.2) with estimators. With data $\{Y_i, X_i, W_i : i = 1,...,n\}$, the m_k's and π_{jk}'s are estimated $n^{-1/2}$ consistently by

$$\hat{m}_k = n_k^{-1} \sum_{i=1}^{n} Y_i I(W_i = w_k)$$

and

$$\hat{\pi}_{jk} = n_k^{-1} \sum_{i=1}^{n} I(X_i = x_j) I(W_i = w_k),$$

where

$$n_k = \sum_{i=1}^{n} I(W_i = w_k).$$

The estimator of g is

$$\hat{g} = (\hat{\Pi}\hat{\Pi}')^{-1}\hat{\Pi}\hat{M},$$

where $\hat{\Pi}$ is the $J \times K$ matrix whose (j,k) element is $\hat{\pi}_{jk}$, $\hat{M} = (\hat{m}_1,...,\hat{m}_K)'$, and $\hat{g} = (\hat{g}_1,...,\hat{g}_J)$. There is no ill-posed inverse problem and, under mild regularity conditions, there are no other complications.

There are, however, many applications in which $K < J$. In some applications, W is binary, so $K = 2$. For example, Card (1995) estimates models of earnings as a function of years of schooling and other variables. Years of schooling is an endogenous explanatory variable. The instrument for it is a binary indicator of whether there is an accredited four-year college in an individual's metropolitan area.

When W is binary, g is not identified nonparametrically if $J > 2$. Nor are there informative, nonparametrically identified bounds on g in the absence of further information or assumptions. A linear model for g, such as that used by Card (1995), is identified but not testable. Thus, in contrast to the case in which X and W are continuously distributed, when X and W are discretely distributed and W has too few points of support, the problem is identification, not estimation. The remainder of this section discusses what can be learned about g when it is not point identified.

Chesher (2004) gives conditions under which there are informative, nonparametrically identified bounds on g. Write model (1.1)-(1.2) in the form

(6.3) $Y = g(X) + U;\ E(U \mid W = w_k) = 0;\ k = 1,...,K$

and

(6.4) $X = H(W, \varepsilon);\ \varepsilon \sim U[0,1];\ \varepsilon \perp W$.

Equation (6.4) defines H to be the conditional quantile function of X and is a tautology. Order the points of support of X so that $x_1 < x_2 < ... < x_J$. Assume that

(6.5) $E(U \mid W = w_k, \varepsilon = e) = c(e)$

for all $k = 1,...,K$ and some monotonic function c. This is a version of assumption (1.10) of the control function model that is discussed in Section 1.2. Also assume that there are an $\overline{e} \in (0,1)$ and points w_{j-1}, w_j in the support of W such that

(6.6) $P(X \le x_j \mid W = w_j) \le \overline{e} \le P(X \le x_{j-1} \mid W = w_{j-1})$

for some $j = 1,...,J$. Chesher (2004) shows that if (6.5) and (6.6) hold, then

(6.7) $\min[E(Y \mid X = x_j, W = w_j), E(Y \mid X = x_j, W = w_{j-1})] \le g_j + c(\overline{e})$

$$\le \max[E(Y \mid X = x_j, W = w_j), E(Y \mid X = x_j, W = w_{j-1})].$$

Inequality (6.7) makes it possible to obtain identified bounds on differences $g_j - g_k$ if (6.6) holds for j and k with the same value of $\overline{e}$. Specifically,

(6.8) $g_{j,\min} - g_{k,\max} \le g_j - g_k \le g_{j,\max} - g_{k,\min}$,

where $g_{j,\min}$ and $g_{j,\max}$, respectively, are the lower and upper bounds on g_j in (6.7). The quantities $g_{k,\min}$ and $g_{k,\max}$ are the bounds obtained by replacing j with k in (6.7). The bounds on $g_j - g_k$ can be estimated consistently by replacing the conditional expectations in

(6.7) with sample averages. Specifically, $E(Y \mid X = x, W = w)$ for any (x, w) in the support of (X, W) is estimated by

$$\hat{E}(Y \mid X = x, W = w) = n_{xw}^{-1} \sum_{i=1}^{n} Y_i I(X_i = x, W_i = w),$$

where

$$n_{xw} = \sum_{i=1}^{n} I(X_i = x, W_i = w).$$

Manski and Pepper (2000) give conditions under which there are identified upper and lower bounds on g and an identified upper bound on $g_j - g_k$. The conditions are:

Monotone treatment response (MTR): Let $y^{(1)}$ and $y^{(2)}$ denote the outcomes (e.g., earnings) that an individual would receive with treatment values (that is, values of x) $x^{(1)}$ and $x^{(2)}$, respectively. Then $x^{(2)} \geq x^{(1)}$ implies $y^{(2)} \geq y^{(1)}$.

Monotone treatment selection (MTS): Let X_S denote the treatment (e.g., years of schooling) that an individual selects. Let x denote any possible treatment level. Then $x^{(2)} \geq x^{(1)}$ implies

$$E(Y \mid X_S = x^{(2)}) \geq E(Y \mid X_S = x^{(1)}).$$

Assumption MTR is analogous to Chesher's (2004) monotonicity condition (6.5). Assumption MTS replaces the assumption that a conventional instrument is available. Manski and Pepper (2000) show that under MTR and MTS,

$$\sum_{\ell: x_\ell < x_j} E(Y \mid X = x_\ell) P(X = x_\ell) + E(Y \mid X = x_j) P(X \geq x_j)$$

$$\leq g_j$$

$$\leq \sum_{\ell: x_\ell > x_j} E(Y \mid X = x_\ell) P(X = x_\ell) + E(Y \mid X = x_j) P(X \leq x_j)$$

and

$$(6.9)\quad 0 \le g_j - g_k \le \sum_{\ell=1}^{k} [E(Y \mid X = x_j) - E(Y \mid X = x_\ell)]P(X = x_\ell)$$

$$+ [E(Y \mid X = x_j) - E(Y \mid X = x_k)]P(x_k \le X \le x_j).$$

$$+ \sum_{\ell=j+1}^{J} [E(Y \mid X = x_\ell) - E(Y \mid X = x_k)]P(X = x_\ell).$$

These bounds can be estimated consistently by replacing expectations with sample averages. Confidence intervals for these bounds and for those in (6.8) can be obtained by taking advantage of the asymptotic normality of sample averages. See, for example, Horowitz and Manski (2000), Imbens and Manski (2004), and Stoye (2009).

6.2 An Empirical Example

This section applies the methods of Section 6.1 to nonparametric estimation of the return to a college education, which is defined here as the percentage change in earnings from increasing an individual's years of education from 12 to 16. The data are those used by Card (1995). They are available at http://emlab.berkeley.edu/users/card/data_sets.html and consist of 3010 records taken from the National Longitudinal Survey of Young Men. Card (1995) treats years of education as endogenous. The instrument for years of education is a binary variable equal to 1 if there is an accredited 4-year college in what Card (1995) calls an individual's "local labor market" and 0 otherwise. A binary instrument point identifies returns to education in Card's parametric models, but it does not provide nonparametric point identification. We investigate the possibility of obtaining bounds on returns to a college education by using the methods of Chesher (2004) and Manski and Pepper (2000).

In the notation of Section 6.1, Y is the logarithm of earnings, X is the number of years of education, and W is the binary instrument. To use Chesher's (2004) method for bounding returns to a college education, the monotonicity condition (6.6) must be satisfied. This requires either

$$(6.10)\quad P(X \le J \mid W = 1) \le P(X \le J - 1 \mid W = 0)$$

or

$$(6.11)\quad P(X \le J \mid W = 0) \le P(X \le J - 1 \mid W = 1)$$

for $J = 12$ and $J = 16$. Table 1 shows the relevant empirical probabilities obtained from Card's (1995) data.. Neither (6.10) nor (6.11) is satisfied. Therefore, Chesher's (2004) method with Card's (1995) data and instrument cannot be used to bound returns to a college education.

Manski's and Pepper's (2000) approach does not require an instrument but depends on the MTR and MTS assumptions, which are not testable. If these assumptions hold for the population represented by Card's data, then replacing population expectations in (6.9) with sample averages yields estimated upper bounds on returns to a college education. These are shown in Table 2 for several levels of labor-force experience. Card (1995) estimated returns from linear models with a variety of specifications. He obtained point estimates in the range 36%-78%, depending on the specification, regardless of experience. The estimates of returns at the lower end of Card's range are consistent with the Manski-Pepper bounds in Table 2.

7. CONCLUSIONS

Nonparametric IV estimation is a new econometric method that has much to offer applied research. Nonparametric estimation:

1. Minimizes the likelihood of specification errors.

2. Reveals the information that is available from the data and the assumption of validity of the instrument as opposed to functional form assumptions.

3. Enables one to assess the importance of functional form assumptions in drawing substantive conclusions from a parametric model.

As this paper has illustrated with empirical examples, nonparametric estimates may yield results that are quite different from those reached with a parametric model. Even if one ultimately chooses to rely on a parametric model to draw conclusions, it is important to understand when the restrictions of the parametric model, as opposed to information in the data and the assumption of instrument validity, are driving the results.

There are also unresolved issues in nonparametric IV estimation. These include choosing basis functions for series estimators and choosing instruments if the dimension of W exceeds that of X.

APPENDIX

Section A.1 outlines the proof of (4.9). Section A.2 presents the asymptotic distributional properties of the $\tilde{\tau}_n$ test of the hypothesis that $g(x,z)$ does not depend on x.

A.1 Outline of Proof of (4.9)

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 $D_j f_{XWZ}(x,w,z)$ denote any j'th partial or mixed partial derivative of $f_{XWZ}(x,w,z)$ with

respect to its first two arguments. Let $D_0 f_{XWZ}(x,w,z) = f_{XWZ}(x,w,z)$. For each $z \in [0,1]$, define $m(w,z) = E(Y \mid W = w, Z = z) f_{WZ}(w,z)$. Define the sequence of function spaces

$$\mathcal{H}_{ns} = \left\{ h = \sum_{j=1}^{J_n} h_j \psi_j : \|h\|_s \le C_g \right\}.$$

Let A_z^* denote the adjoint of A_z. For $z \in [0,1]$, define

$$\rho_{nz} = \sup_{h \in \mathcal{H}_{ns}} \frac{\|h\|}{\left\| (A_z^* A_z)^{1/2} h \right\|}.$$

Blundell, Chen, and Kristensen (2007) call ρ_{nz} the sieve measure of ill posedness and discuss its relation to the eigenvalues of $A_z^* A_z$. Define

$$g_{nz}(x) = \sum_{j=1}^{J_n} g_{jz} \psi_j(x).$$

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

$$a_{jkz} = \int f_{XWZ}(x,w,z) \psi_j(x) \psi_k(w) dx dw.$$

Let A_{nz} be the operator whose kernel is

$$a_{nz}(x,w) = \sum_{j=1}^{J_n} \sum_{k=1}^{J_n} a_{jkz} \psi_j(x) \psi_k(w).$$

Also define $m_{nz} = A_{nz} g_{nz}$.

Make the following assumptions.

Assumption 1: (i) The support of (X,W,Z) is contained in $[0,1]^3$. (ii) (X,W,Z) has a probability density function f_{XWZ} with respect to Lebesgue measure. (iii) There are an integer $r \ge 2$ and a constant $C_f < \infty$ such that $|D_j f_{XWZ}(x,w,z)| \le C_f$ for all $(x,w,z) \in [0,1]^3$ and $j = 0,1,...,r$. (iv) $|D_r f_{XWZ}(x_1,w_1,z) - D_r f_{XWZ}(x_2,w_2,z)| \le C_f \left\| (x_1,w_1) - (x_2,w_2) \right\|_E$ for any order r derivative, any (x_1,w_1) and (x_2,w_2) in $[0,1]^2$ and any $z \in [0,1]$.

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

Assumption 3: (i) For each $z \in [0,1]$, (1.3) has a solution $g(\cdot, z)$ with $\|g(\cdot, z)\|_s < C_0$ and $s \geq 2$. (ii) The estimator $\hat{g}$ is as defined in (4.7). (iii) The function $m(w, z)$ has $r + s$ square-integrable derivatives with respect to w and r bounded derivatives with respect to z.

Assumption 4: (i) The basis functions $\{\psi_j\}$ are orthonormal, complete on $L_2[0,1]$, and bounded uniformly over j. (ii) $\|A_{nz} - A_z\| = O(J_n^{-r})$ uniformly over $z \in [0,1]$. (iii) For any $\nu \in L_2[0,1]$ with ℓ square integrable derivatives, there are coefficients ν_j ($j = 1, 2, \ldots$) and a constant $C < \infty$ that does not depend on ν such that

$$\left\| \nu - \sum_{j=1}^{J} \nu_j \psi_j \right\| \leq C J^{-\ell}.$$

Assumption 5: (i) The operator A_z is nonsingular for each $z \in [0,1]$. (ii) $\rho_{nz} = O(J_n^r)$ uniformly over $z \in [0,1]$. (iii) As $n \to \infty$,

$$\rho_{nz} \sup_{\nu \in \mathcal{H}_{ns}} \frac{\|(A_{nz} - A_z)\nu\|}{\|\nu\|} = O(J_n^{-s})$$

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

Assumption 6: The kernel function K is a symmetrical, twice continuously differentiable function on $[-1,1]$, and

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

Assumption 7: (i) The bandwidth, b_n, satisfies $b_n = c_b n^{-1/(2r+1)}$, where c_b is a constant and $0 < c_b < \infty$. (ii) $J_n = C_J n^{\kappa/(2r+2s+1)}$ for some constant $C_J < \infty$.

Assumptions 1 and 2 are smoothness and boundedness conditions. Assumption 3 defines the function being estimated and the estimator. The assumption requires $\|g(\cdot, z)\|_s < C_0$ (strict inequality) to avoid complications that arise when g is on the boundary of $\mathcal{H}_s$. Assumption 3 also ensures that the function m is sufficiently smooth. This function has more derivatives with respect to w than z because $m(w, z) = [A_z g(\cdot, z)](w, z)$, and A_z smooths g along its first argument but not its second. 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) 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 5(iii) ensures that A_{nz} is a "sufficiently accurate" approximation to A_z on $\mathcal{H}_{ns}$. This assumption complements 4(ii), which specifies the accuracy of A_{nz} as an approximation to A_z on the larger set $\mathcal{H}_s$. Assumption 5(iii) can be interpreted as a smoothness restriction on f_{XWZ}. For example, 5(iii) is satisfied if assumptions 4 and 5(ii) hold and A_z maps $\mathcal{H}_s$ to $\mathcal{H}_{r+s}$. Assumption 5(iii) also can be interpreted as a restriction on the sizes of the values of a_{jkz} for $j \neq k$. Hall and Horowitz (2005) used a similar diagonality restriction. Assumption 6 requires K to be a higher-order kernel if f_{XWZ} is sufficiently smooth. K can be replaced by a boundary kernel (Gasser and Müller 1979; Gasser, Müller, and Mammitzsch 1985) if f_{XWZ} does not approach 0 smoothly on the boundary of its support.

Proof of (4.9): Use the notation $g(x,z) = g_z(x)$, $\hat{g}(x,z) = \hat{g}_z(x)$, and $m(w,z) = m_z(w)$. For each $z \in (0,1)$,

(A.1) $$\|\hat{g}_z - g_z\| \leq \|\hat{g}_z - g_{nz}\| + \|g_{nz} - g_z\|.$$

Moreover,

$$\|g_{nz} - g_z\| = O(J^{-s})$$

by assumption 4(iii). Therefore,

(A.2) $$\|\hat{g}_z - g_z\| \leq \|\hat{g}_z - g_{nz}\| + O(J^{-s}).$$

Now consider $\|\hat{g}_z - g_{nz}\|$. By $P(\hat{g}_z \in \mathcal{H}_{ns}) \to 1$ as $n \to \infty$ and the definition of ρ_{nz},

(A.3) $$\|\hat{g}_z - g_{nz}\| \leq \rho_{nz} \|A_z(\hat{g}_z - g_{nz})\|$$

with probability approaching 1 as $n \to \infty$. In addition, $\hat{A}_z \hat{g}_z = \hat{m}_z$ and $A_z g_z = m_z$. Therefore,

$$A_z(\hat{g}_z - g_{nz}) = (A_z - \hat{A}_z)\hat{g}_z + \hat{A}_z \hat{g}_z - A_z(g_{nz} - g_z) - A_z g_z$$

$$= (A_z - \hat{A}_z)\hat{g}_z + \hat{m}_z - m_z - A_z(g_{nz} - g_z).$$

The triangle inequality now gives

$$\|A_z(\hat{g}_z - g_{nz})\| \leq \|(\hat{A}_z - A_z)\hat{g}_z\| + \|\hat{m}_z - m_z\| + \|A_z(g_{nz} - g_z)\|.$$

Standard calculations for kernel estimators show that under assumptions 3(iii), 6, and 7,

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

and

$$\left\|E\hat{m}_z - m_z\right\| = O[J_n^{1/2} n^{-r/(2r+1)} + J_n^{-r-s}].$$

Therefore,

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

In addition, $A_{nz}(g_{nz} - g_z) = 0$, so $\left\|A_z(g_{nz} - g_z)\right\| = \left\|(A_{nz} - A_z)(g_{nz} - g_z)\right\|$. Therefore,

$$\left\|A_z(g_{nz} - g_z)\right\| = \frac{\left\|(A_{nz} - A_z)(g_{nz} - g_z)\right\|}{\left\|g_{nz} - g_z\right\|}\left\|g_{nz} - g_z\right\|$$

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

by assumptions 4 and 5. Therefore, we have

(A.4) $$\left\|A_z(\hat{g}_z - g_{nz})\right\| \le \left\|(\hat{A}_z - A_z)\hat{g}_z\right\| + O_p(J_n^{1/2} n^{-r/(2r+1)} + J_n^{-r-s}).$$

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

$$\left\|(\hat{A}_z - A_z)\hat{g}\right\| \le \left\|(\hat{A}_z - A_{nz})\hat{g}\right\| + \left\|(A_{nz} - A_z)\hat{g}\right\|$$

$$= \left\|(\hat{A}_z - A_{nz})\hat{g}\right\| + O(J_n^{-r-s}).$$

For each $z \in (0,1)$

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

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

(A.5) $$\left\|(\hat{A}_z - A_{nz})\nu\right\| = \sum_{k=1}^{J_n}\left[\sum_{j=1}^{J_n}(\hat{a}_{jkz} - a_{jkz})\nu_j\right]^2.$$

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

$$\sum_{j=1}^{J_n} \nu_j \hat{a}_{jkz} = \sum_{j=1}^{J_n} \nu_j \frac{1}{nb_n} \sum_{i=1}^{n} \psi_j(X_i)\psi_k(W_i)K_{b_n}(z - Z_i).$$

Therefore, it follows from Bernstein's inequality that

$$\sum_{j=1}^{J_n} \nu_j (\hat{a}_{jkz} - E\hat{a}_{jkz}) = O_p[(nb_n)^{-1/2}]$$

uniformly over $\nu \in \mathcal{H}_{ns}$. Therefore,

$$\text{(A.6)} \quad \left\| (\hat{A}_z - E\hat{A}_{nz})\nu \right\| = O[J_n^{1/2}/(nb_n)^{1/2}]$$

uniformly over $\nu \in \mathcal{H}_{ns}$. In addition, $E\hat{a}_{jkz} = a_{jkz} + O(b_n^r)$. Therefore, boundedness of $\sum_{j=1}^{J_n} |\nu_j|$ gives

$$\sum_{j=1}^{J_n} (E\hat{a}_{jkz} - a_{jkz})\nu_j = O(b_n^r)$$

and

$$\text{(A.7)} \quad \left\| (E\hat{A}_z - A_{nz})\nu \right\| = O(J_n^{1/2} b_n^r)$$

uniformly over $\nu \in \mathcal{H}_{ns}$. Combining (A.6) and (A.7) and using assumption 7 gives

$$\sup_{\nu \in \mathcal{H}_{ns}} \left\| (\hat{A}_z - A_{nz})\nu \right\| = O_p[J_n^{1/2} n^{-r/(2r+1)}].$$

Therefore,

$$\text{(A.8)} \quad \sup_{\nu \in \mathcal{H}_{ns}} \left\| (\hat{A}_z - A_z)\nu \right\| = O_p[J_n^{1/2} n^{-r/(2r+1)} + J_n^{-r-s}].$$

Combining (A.4) and (A.8) gives

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

This result and assumption 5(ii) imply that

$$\text{(A.9)} \quad \rho_{nz} \left\| A_z(\hat{g}_z - g_{nz}) \right\| = O_p[J_n^{r+1/2} n^{-r/(2r+1)} + J_n^{-s}].$$

The theorem follows by combining (A.2), (A.3), and (A.9). Q.E.D.

A.2 Asymptotic Properties of the $\tilde{\tau}_n$ Test

Let $\|(x_1, w_1, z_1) - (x_2, w_2, z_2)\|_E$ denote the Euclidean distance between the points (x_1, w_1, z_1) and (x_2, w_2, z_2). Let $D_j f_{XWZ}$ denote any j'th partial or mixed partial derivative of f_{XWZ}. Set

$D_0 f_{XWZ}(x,w,z) = f_{XWZ}(x,w,z)$. Let $s \geq 2$ be an integer. Define $V = Y - G(Z)$, and let f_Z denote the density of Z. Define $T_z = A_z^* A_z$. Make the following assumptions.

1. (i) The support of (X,W,Z) is contained in $[0,1]^3$. (ii) (X,W,Z) has a probability density function f_{XWZ} with respect to Lebesgue measure. (ii) There is a constant $C_Z > 0$ such that $f_Z(z) \geq C_Z$ for all $z \in \text{supp}(Z)$. (iv) There is a constant $C_f < \infty$ such that $|D_j f_{XWZ}(x,w,z)| \leq C_f$ for all $(x,w,z) \in [0,1]^3$ and $j = 0,1,2$, where derivatives at the boundary of $\text{supp}(X,W,Z)$ are defined as one-sided. (iv) $|D_s f_{XWZ}(x_1,w_1,z_1) - D_s f_{XZW}(x_2,w_2,z_2)|$ $\leq C_f \|(x_1,w_1,z_1) - (x_2,w_2,z_2)\|_E$ for any 2nd derivative and any $(x_1,w_1,z_1),(x_2,w_2,z_2) \in [0,1]^3$. (v) T_z is nonsingular for almost every $z \in [0,1]$.

2. (i) $E(U \mid Z = z, W = w) = 0$ and $\boldsymbol{E}(U^2 \mid Z = z, W = w) \leq C_{UV}$ for each $(z,w) \in [0,1]^2$ and some constant $C_{UV} < \infty$. (ii) $|g(x,z)| \leq C_g$ for some constant $C_g < \infty$ and all $(x,z) \in [0,1]^2$.

3. The function G satisfies $|D_j G(z)| \leq C_f$ for all $z \in [0,1]$ and $j = 0,1,2$. (ii) $|D_s G(z_1) - D_s G(z_2)| \leq C_f |z_1 - z_2|$ for any 2nd derivative and any $(z_1,z_2) \in [0,1]^2$. (iii) $\boldsymbol{E}(V^2 \mid Z = z) \leq C_{UV}$ for each $z \in [0,1]$.

4. (i) K_b satisfies (4.12) and $|K_b(u_2,\xi) - K_b(u_1,\xi)| \leq C_K |u_2 - u_1| / b$ for all u_2, u_1, all $\xi \in [0,1]$, and some constant $C_K < \infty$. For each $\xi \in [0,1]$, $K_h(b,\xi)$ is supported on $[(\xi - 1)/b, \xi/b] \cap \mathcal{K}$, where $\mathcal{K}$ is a compact interval not depending on ξ. Moreover,

$$\sup_{b>0,\xi\in[0,1],u\in\mathcal{K}} |K_b(bu,\xi)| < \infty .$$

(ii) The bandwidth b_1 satisfies $b_1 = c_{b1} n^{-1/7}$, where $c_{b1} < \infty$ is a constant. (iii) The bandwidth, b_2, satisfies $b_2 = c_{b2} n^{-\alpha}$, where $c_{b2} < \infty$ is a constant and $1/4 < \alpha < 1/2$.

Assumption 1(ii) is used to avoid imprecise estimation of G in regions where f_Z is close to 0. The assumption can be relaxed by replacing the fixed distribution of (X,Z,W) by a sequence of distributions with densities $\{f_{nXZW}\}$ and $\{f_{nZ}\}$ $(n = 1,2,...)$ that satisfy $f_{nZ}(z) \geq C_n$ for all $(z) \in [0,1]$ and a sequence $\{C_n\}$ of strictly positive constants that converges to 0 sufficiently slowly. Assumption 1(v) combined with the moment condition $E(U \mid X, Z) = 0$ implies that g is identified and the instruments W are valid in the sense of being suitably related to X. Assumption 4(iii) implies that the estimator of G is undersmoothed. Undersmoothing

prevents the asymptotic bias of $\hat{G}^{(-i)}$ from dominating the asymptotic distribution of $\tilde{\tau}_n$. The remaining assumptions are standard in nonparametric estimation.

The $\tilde{\tau}_n$ test is a modification of the exogeneity test of Blundell and Horowitz (2007), and its properties can be derived by using the methods of that paper. Accordingly, the properties of the $\tilde{\tau}_n$ test are stated here without proof. Define $V_i = Y_i - G(Z_i)$ $(i = 1,...,n)$,

$$B_n(x,z) = n^{-1/2}\sum_{i=1}^{n} V_i \left[f_{XZW}(x, Z_i, W_i) - \frac{1}{f_Z(Z_i)} \int_0^1 t_{Z_i}(\xi, x) d\xi \right] \ell(Z_i, z),$$

and

$$R(x_1, z_1; x_2, z_2) = \boldsymbol{E}[B_n(x_1, z_1) B_n(x_2, z_2)].$$

Define the operator Ω on $L_2[0,1]^2$ by

$$(\Omega h)(x,z) = \int_0^1 R(x, z; \xi, \zeta) h(\xi, \zeta) d\xi d\zeta .$$

Let $\{\omega_j : j = 1,2,...\}$ denote the eigenvalues of Ω sorted so that $\omega_1 \geq \omega_2 \geq ... \geq 0$. Let $\{\chi^2_{1j} : j = 1,2,...\}$ denote independent random variables that are distributed as chi-square with one degree of freedom. Define the random variable

$$\tilde{\tau}_\infty = \sum_{j=1}^{\infty} \omega_j \chi^2_{1j} .$$

For any α such that $0 < \alpha < 1$, let ξ_α denote the $1-\alpha$ quantile of the distribution of τ_∞. Then

1. Under H_0, $\tilde{\tau}_n \to^d \tilde{\tau}_\infty$.
2. Under H_1,

$$\lim_{n\to\infty} P(\tilde{\tau}_n > \xi_\alpha) = 1$$

for any α such that $0 < \alpha < 1$. Thus, the $\tilde{\tau}_n$ test is consistent.

The final result shows that for any $\varepsilon > 0$ and as $n \to \infty$, the $\tilde{\tau}_n$ test rejects H_0 with probability exceeding $1-\varepsilon$ uniformly over a set of functions g whose distance from G is $O(n^{-1/2})$. The practical consequence of this result is to define a large class of alternatives against which the $\tilde{\tau}_n$ test has high power in large samples. The following additional notation is used. Let L be the operator on $L_2[0,1]$ that is defined by

$$(Lh)(z) = \int_0^1 h(\zeta) \ell(\zeta, z) d\zeta .$$

Define $q(x,z) = g(x,z) - G(z)$. Let f_{XZW} be fixed. For each $n = 1,2,...$ and finite $C > 0$, define $\mathcal{F}_{nc}$ as a set of distributions of (Y,X,Z,W) such that: (i) f_{XZW} satisfies assumption 1; (ii) $E[Y - g(X,Z) \mid Z,W] = 0$ for some function g that satisfies assumption 2 with $U = Y - g(X,Z)$; (iii) $E(Y \mid Z = z) = G(z)$ for some function G that satisfies assumption 3 with $V = Y - G(Z)$; (iv) $\|LT_z q\| \geq n^{-1/2} C$, where $\|\cdot\|$ denotes the $L_2[0,1]^2$ norm; and (v) $h_1^s (\log n) \|q\| / \|LT_z q\| = o(1)$ as $n \to \infty$. $\mathcal{F}_{nc}$ is a set of distributions of (Y,X,Z,W) for which the distance of g from G shrinks to zero at the rate $n^{-1/2}$ in the sense that $\mathcal{F}_{nc}$ includes distributions for which $\|q\| = O(n^{-1/2})$. Condition (v) rules out distributions for which q depends on (x,z) only through sequences of eigenvectors of T_z whose eigenvalues converge to 0 too rapidly. The practical significance of condition (v) is that the $\tilde{\tau}_n$ test has low power when g differs from G only through eigenvectors of T_z with very small eigenvalues. Such differences tend to oscillate rapidly (that is, to be very wiggly) and are unlikely to be important in most applications. The uniform consistency result is as follows.

3. Given any $\varepsilon > 0$, any α such that $0 < \alpha < 1$, and any sufficiently large (but finite) C, $\liminf_{n \to \infty} \inf_{\mathcal{F}_{nc}} \boldsymbol{P}(\tilde{\tau}_n > \xi_\alpha) \geq 1 - \varepsilon$.

The remainder of this section explains how to obtain an approximate asymptotic critical value for the $\tilde{\tau}_n$ test. The method is based on replacing the asymptotic distribution of $\tilde{\tau}_n$ with an approximate distribution. The difference between the true and approximate distributions can be made arbitrarily small under both the null hypothesis and alternatives. Moreover, the quantiles of the approximate distribution can be estimated consistently as $n \to \infty$. The approximate $1 - \alpha$ critical value of the $\tilde{\tau}_n$ test is a consistent estimator of the $1 - \alpha$ quantile of the approximate distribution.

We now describe the approximation to the asymptotic distribution of $\tilde{\tau}_n$. Given any $\varepsilon > 0$, there is an integer $K_\varepsilon < \infty$ such that

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

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 $0 < \boldsymbol{P}(\tilde{\tau}_\infty > z_{\varepsilon\alpha}) - \alpha < \varepsilon$. Thus, using $z_{\varepsilon\alpha}$ to approximate the asymptotic $1-\alpha$ critical value of $\tilde{\tau}_n$ 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 $\tilde{\tau}_n$ test when the null hypothesis is false. The approximate $1-\alpha$ critical value for the $\tilde{\tau}_n$ 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 approximate critical value of $\tilde{\tau}_n$ is the $1-\alpha$ quantile of the distribution of

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

This quantile 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 $\tilde{\tau}_n$ test is not highly sensitive to K_ε, so it is not necessary to attempt precision in making the choice.

We now explain how to obtain the estimated eigenvalues $\{\hat{\omega}_j\}$. Let $\hat{f}_{XZW}$ be a kernel estimator of f_{XZW}. Define

$$\hat{t}_z(x_1, x_2) = \int_0^1 \hat{f}_{XZW}(x_1, z, w)\, \hat{f}_{XZW}(x_2, z, w) dw .$$

Estimate the V_i's by generating data from an estimated version of the model

(A.10) $\tilde{Y} = G(Z) + \tilde{V}$,

where $\tilde{Y} = Y - E[Y - G(Z) \mid Z, W]$ and $\tilde{V} = \tilde{Y} - G(Z)$. Model (A.10) is identical to model (1.3)-(1.4) under H_0. Moreover, the moment condition $E(\tilde{V} \mid Z, W) = 0$ holds regardless of whether H_0 is true. Observe that $\tilde{V} = Y - E(Y \mid Z, W)$. Let $\hat{E}^{(-i)}(Y \mid Z, W)$ denote the leave-observation-*i*-out nonparametric regression of Y on (Z, W). Estimate V_i by

$$\hat{V}_i = Y_i - \hat{E}^{(-i)}(Z_i, W_i) .$$

Now define

$$\hat{r}(x,Z_i,W_i)=\hat{f}_{XZW}(x,Z_i,W_i)-\frac{1}{\hat{f}_Z(Z_i)}\int_0^1\hat{t}_{Z_i}(\xi,x)d\xi$$

$R(x_1,z_1;x_2,z_2)$ is estimated consistently by

$$\hat{R}(x_1,z_1,x_2,z_2)=n^{-1}\sum_{i=1}^{n}\hat{r}(x_1,Z_i)\hat{r}(x_2,Z_i)\ell(Z_i,z_1)\ell(Z_i,z_2)\hat{V}_i^2\,.$$

Define the operator $\hat{\Omega}$ on $L_2[0,1]$ by

$$(\hat{\Omega}\psi)(x,z)=\int_0^1\hat{R}(x,z;\xi,\zeta)\psi(\xi,\zeta)d\xi d\zeta\,.$$

Denote the eigenvalues of $\hat{\Omega}$ by $\{\hat{\omega}_j : j=1,2,...\}$ and order them so that $\hat{\omega}_{M1}\geq\hat{\omega}_{M2}\geq...\geq 0$. Then the $\hat{\omega}_j$'s are consistent estimators of the ω_j's.

To obtain an accurate numerical approximation to the $\hat{\omega}_j$'s, let $\hat{F}(x,z)$ denote the $n\times 1$ vector whose i'th component is $\hat{r}(x,Z_i,W_i)\ell(Z_i,z)$, and let Υ denote the $n\times n$ diagonal matrix whose (i,i) element is $\hat{V}_i^2$. Then

$$\hat{R}(x_1,z_1;x_2,z_2)=n^{-1}\hat{F}(x_1,z_1)'\Upsilon\hat{F}(x_2,z_2)\,.$$

The computation of the eigenvalues can now be reduced to finding the eigenvalues of a finite-dimensional matrix. To this end, let $\{\phi_j : j=1,2,...\}$ be a complete, orthonormal basis for $L_2[0,1]^2$. Let $\{\psi_j\}$ be a complete orthonormal basis for $L_2[0,1]$. Then

$$\hat{f}_{XZW}(x,Z,W)\ell(Z,z)=\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\hat{d}_{jk}\phi_j(x,z)\phi_k(Z,W)\,,$$

where

$$\hat{d}_{jk}=\int_0^1dx\int_0^1dz_1\int_0^1dz_2\int_0^1dw\hat{f}_{XZW}(x,z_2,w)\ell(z_2,z_1)\phi_j(x,z_1)\phi_k(z_2,w)\,,$$

and

$$\ell(Z,z)\int_0^1\hat{t}_Z(\xi,x)d\xi=\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\hat{a}_{jk}\phi_j(x,z)\psi_k(Z)\,,$$

where

$$\hat{a}_{jk}=\int_0^1dx\int_0^1dz_1\int_0^1dz_2\int_0^1d\xi\hat{t}_{z_1}(\xi,x)\ell(z_1,z_2)\phi_j(x,z_2)\psi_k(z_1)\,.$$

Approximate $\hat{f}_{XZW}(x,Z,W)\ell(Z,z)$ and $\ell(Z,z)\int_0^1\hat{t}_Z(\xi,x)d\xi$, respectively, by the finite sums

$$\Pi_f(x,z,W,Z) = \sum_{j=1}^{M} \sum_{k=1}^{M} \hat{d}_{jk} \phi_j(x,z) \phi_k(Z,W)$$

and

$$\Pi_t(x,z,Z)) = \sum_{j=1}^{M} \sum_{k=1}^{M} \hat{a}_{jk} \phi_j(x,z) \psi_k(Z).$$

for $M < \infty$. Since $\hat{f}_{XZW}\ell$ and $\ell \int_0^1 \hat{t}_Z d\xi$ are known functions, M can be chosen to approximate them with any desired accuracy. Let Φ be the $n \times L$ matrix whose (i, j) component is

$$\Phi_{ij} = n^{-1/2} \sum_{k=1}^{L} [\hat{d}_{jk} \phi_k(Z_i, W_i) - \hat{a}_{jk} \psi_k(Z_i) / \hat{f}_Z(Z_i)].$$

The eigenvalues of $\hat{\Omega}$ are approximated by those of the $L \times L$ matrix $\Phi' \Upsilon \Phi$.

REFERENCES

Ai, C. and X. Chen (2009). Semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions, working paper, Department of Economics, Yale University, New Haven, CT, USA.

Angrist, J.D. and V. Lavy (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement, *Quarterly Journal of Economics*, 114, 533-575.

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

Blundell, R. and J.L. Horowitz (2007). A nonparametric test for exogeneity, *Review of Economic Studies*, 74, 1035-1058.

Blundell, R. and J.L. Powell (2003). Endogeneity in nonparametric and semiparametric regression models, in M. Dewatripont, L.P. Hansen, and S. Turnovsky, eds., *Advances in Economics and Econometrics: Theory and Applications: Eighth World Congress*, Vol. 2, pp. 312-357, Cambridge, U.K.: Cambridge University Press.

Card, D. (1995). Using geographic variation in college proximity to estimate returns to schooling. In L.N. Christofides, E.K. Grant, and R. Swidinsky, editors, *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, Toronto: University of Toronto Press.

Carrasco, M., J.-P. Florens, and E. Renault (2007): Linear inverse problems in structural econometrics: estimation based on spectral decomposition and regularization, in *Handbook of Econometrics*, Vol. 6, ed. by E.E. Leamer and J.J. Heckman. Amsterdam: North-Holland, pp. 5634-5751.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models, in James J. Heckman and Edward E. Leamer, eds., *Handbook of Econometrics*, vol. 6b, Amsterdam: North-Holland, pp. 5549-5632.

Chen, X. and D. Pouzo (2008). Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Moments, working paper, Department of Economics, Yale University.

Chen, X. and M. Reiss (2007). On Rate Optimality for Ill-Posed Inverse Problems in Econometrics, working paper CWP20/07, Centre for Microdata Methods and Practice, Department of Economics, University College London.

Chernozhukov, V. and C. Hansen (2005). An IV model of quantile treatment effects, *Econometrica*, 73, 245-261.

Chernozhukov, V., G.W. Imbens, and W.K. Newey (2007). Instrumental variable identification and estimation of nonseparable models via quantile conditions, *Journal of Econometrics*, 139, 4-14.

Chesher, A. (2004). Identification in additive error models with discrete endogenous variables, working paper CWP11/04, Centre for Microdata Methods and Practice, Department of Economics, University College London.

Chesher, A. (2005). Nonparametric identification under discrete variation, *Econometrica*, 73, 1525-1550.

Conway, John B. (1990). *A Course in Functional Analysis*, 2nd edition, New York: Springer-Verlag.

Darolles, S., J.-P. Florens, and E. Renault (2006). Nonparametric Instrumental Regression, working paper, University of Toulouse.

Engl, H.W., M. Hanke, and A. Neubauer (1996): *Regularization of Inverse Problems*. Dordrecht: Kluwer Academic Publishers.

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 O. Linton (1994). Applied Nonparametric Methods, in *Handbook of Econometrics*, vol. 4, R. F. Engle and D. F. McFadden, eds., Amsterdam: Elsevier, Ch. 38.

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

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators, *Econometrica*, 50, 1029-1054.

Heckman, J.J. and E.J. Vylacil (2007). Econometric evaluation of social programs, Part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs and to forecast their effects in new environments, in *Handbook of Econometrics*, vol. 6B, J.J. Heckman and E.E. Leamer, eds., Amsterdam: Elsevier, Ch. 71.

Horowitz, J.L. (2006). Testing a parametric model against a nonparametric alternative with identification through instrumental variables, *Econometrica*, 74, 521-538.

Horowitz, J.L. (2007). Asymptotic normality of a nonparametric instrumental variables estimator, *International Economic Review*, 48, 1329-1349.

Horowitz, J.L. (2009a). Specification testing in nonparametric instrumental variables estimation, *Journal of Econometrics*, forthcoming.

Horowitz, J.L. (2009b). *Semiparametric and Nonparametric Methods in Econometrics*, New York: Springer-Verlag.

Horowitz, J.L. (2010). Empirical selection of the regularization parameter in nonparametric instrumental variables estimation, working paper, Department of Economics, Northwestern University, Evanston, IL, USA.

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

Horowitz, J.L. and S. Lee (2009). Uniform confidence bands for functions estimated nonparametrically with instrumental variables, working paper CWP18/09, Centre for Microdata Methods and Practice, Department of Economics, University College London.

Horowitz, J.L. and C.F. Manski (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data, *Journal of the American Statistical Association*, 95, 77-84.

Imbens, G. and C.F. Manski (2004). Confidence intervals for partially identified parameters, *Econometrica*, 72, 1845-1857.

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

Liusternik, L.A. and V.J. Sobolev (1961). *Elements of Functional Analysis*, New York: Ungar Publishing Company.

Manski, C.F. and J.V. Pepper (2000). Monotone instrumental variables: With an application to returns to schooling, *Econometrica*, 68, 997-1010.

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

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

Pietsch, A. (1980). Eigenvalues of integral operators. I, *Mathematische Annalen*, 247, 169-178.

Pinkse, J. (2000). Nonparametric two-step regression estimation when regressor and error are dependent, *Canadian Journal of Statistics*, 28, 289-300.

Stoye, J. (2009). More on confidence intervals for partially identified parameters, *Econometrica*, 77, 1299-1315.

Table 1: Empirical Probabilities of Various Levels of Education[a]

Years of Education	With Nearby College	Without Nearby College
11	0.136	0.228
	(0.022)	(0.028)
12	0.456	0.578
	(0.016)	(0.021)
15	0.707	0.775
	(0.012)	(0.015)
16	0.866	0.915
	(0.008)	(0.009)

a. Table entries are the empirical probabilities that years of education is less than or equal to 11, 12, 15, and 16 conditional on whether there is a 4-year accredited college in an individual's local labor market. Quantities in parentheses are standard errors.

Table 2: Manski-Pepper (2000) Upper Bounds on Returns to a University Education

Years of Experience	Point Estimate of Upper Bound	Upper 95% Confidence Limit
6-7	0.38	0.44
8-10	0.40	0.46
11-23	0.52	0.61

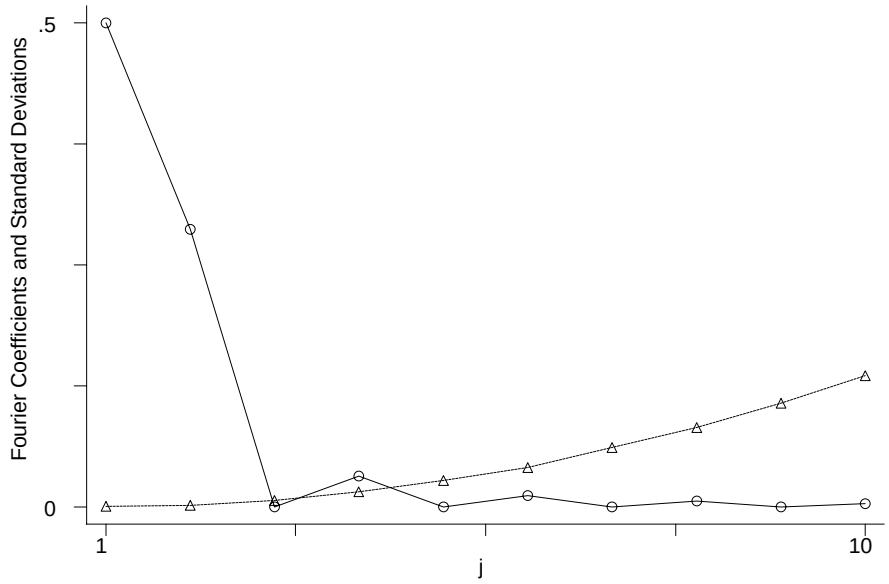

Figure 1: Illustration of Ill-Posed Inverse Problem. The solid line is the absolute values of the generalized Fourier coefficients. The dashed line is the standard deviation of maximum likelihood estimates of these coefficients.

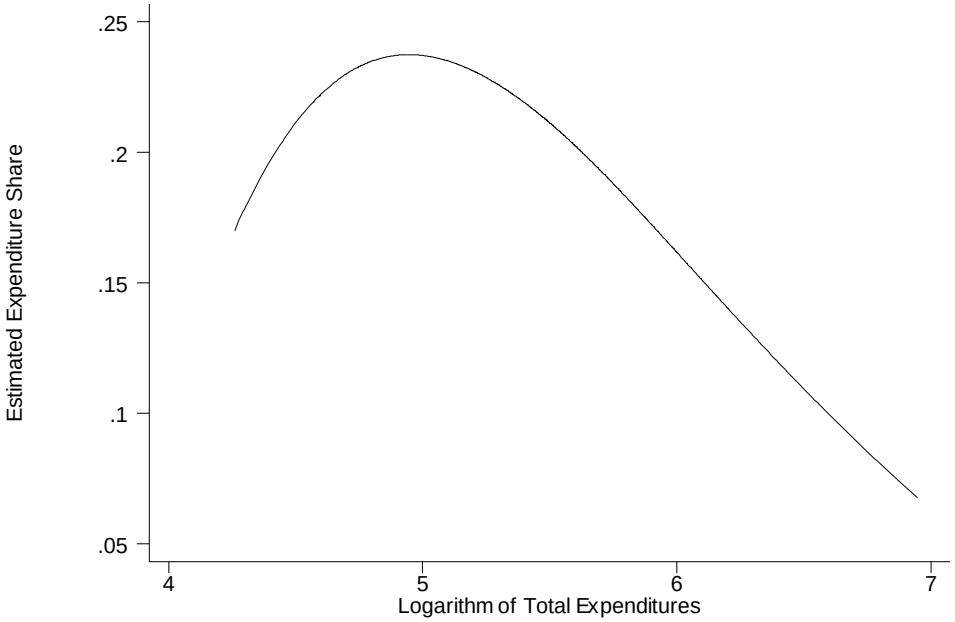

Figure 2: Estimated Engel Curve for Food

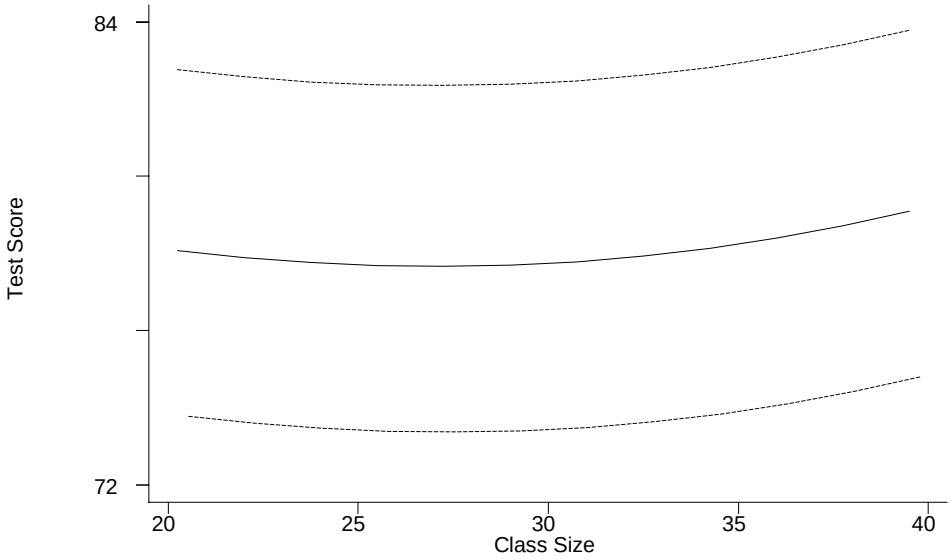

Figure 3: Estimate of test score as function of class size. Solid line is estimate. Dashed lines indicate a uniform 95% confidence band.