Econometrica, Vol. 71, No. 5 (September, 2003), 1565–1578

INSTRUMENTAL VARIABLE ESTIMATION OF NONPARAMETRIC MODELS[1]

BY WHITNEY K. NEWEY AND JAMES L. POWELL

In econometrics there are many occasions where knowledge of the structural relationship among dependent variables is required to answer questions of interest. This paper gives identification and estimation results for nonparametric conditional moment restrictions. We characterize identification of structural functions as completeness of certain conditional distributions, and give sufficient identification conditions for exponential families and discrete variables. We also give a consistent, nonparametric estimator of the structural function. The estimator is nonparametric two-stage least squares based on series approximation, which overcomes an ill-posed inverse problem by placing bounds on integrals of higher-order derivatives.

KEYWORDS: Structural models, nonparametric estimation, instrumental variables.

1. INTRODUCTION

IN ECONOMETRICS THERE ARE MANY occasions where knowledge of the structural relationship among dependent variables is required to answer questions of interest. For parametric models with additive disturbances, moment restrictions on the disturbances are often imposed to identify and consistently estimate the parameters of interest; for a linear simultaneous equation, zero covariance between the instruments and disturbance, along with identification, suffices for consistent estimation. In a nonparametric setting, a stronger restriction that the disturbance has conditional mean zero given instruments is important; a finite number of zero covariance restrictions will generally not suffice to identify an infinite dimensional function.

We characterize identification of structural functions as completeness of certain conditional distributions, and give sufficient conditions for exponential families and discrete variables. Estimation is more difficult. The relationship between structure and reduced form is a Fredholm integral equation of the first kind, leading to an ill-posed inverse problem (e.g., see Kress (1989)). Difficulties associated with such problems are existence and computation of a solution and continuity of that solution in the reduced form. We consider solutions to each of these problems. Existence of an estimator is obtained using nonparametric extensions of minimum distance methods. Also, computation can be carried out using a parametric approximation to the unknown function. Noncontinuity is harder to overcome. Our approach is to restrict the true function to be an element of a compact set of functions, by imposing bounds on higher-order derivatives, which makes the mapping from reduced form to structure continuous. Such

[1] This research was supported by the National Science Foundation, the Alfred P. Sloan Foundation, and the Center for Advanced Study in the Behavioral Sciences. An early version of the paper was presented at the 1988 North American Winter Meetings of the Econometric Society. I. Fernandez and C. Hansen provided able research assistance. Discussion with Angus Deaton inspired this paper; helpful comments were also provided by R. Blundell, X. Chen, V. Chernozhukov, L. Hansen, J. Heckman, H. Ichimura, C. Manski, C. Meghir, J. Scheinkman, G. Wahba, a co-editor, two referees, and participants in seminars at BellCore, and Chicago, Northwestern, Princeton,Yale, Wisconsin Universities.

restrictions on derivatives are common in the work on sieve estimation, including that of Gallant (1987) and Chen and Shen (1998).

The estimator we propose is a nonparametric analog to the familiar two-stage least squares (2SLS) estimator for linear models with endogenous regressors; the first stage involves nonparametric estimation of conditional means of "basis" functions which are used in a second-stage series approximation of the unknown function to be estimated. When bounds on integrals of derivatives are imposed, a quadratic objective form for the estimator results. We also generalize this estimator to obtain a nonparametric analog of Amemiya's (1974) nonlinear 2SLS estimator. A consistency result for the estimator is obtained, with proofs provided in a mathematical appendix.

Ill-posed inverse problems have been previously considered in the statistics literature, particularly for the deconvolution problem of estimating a density from its convolution with a known distribution; e.g. see the survey of O'Sullivan (1986). A nonparametric transformation model was considered by Breiman and Friedman (1985), although their model is different than the one here, being based on a least squares definition of the structural function rather than on instrumental variable identification. This work, as well as more recent work by Opsomer and Ruppert (1997) and Mammen, Linton, and Nielsen (1999) is about solving integral equations that are not ill-posed inverse problems.

Prior work on nonparametric structural models includes identification results of Roehrig (1988), although those are for a different model, where disturbances are independent of instruments. Also, following the first version of this paper (Newey and Powell (1988)), Newey, Powell, and Vella (1999), Brown and Matzkin (1998), Altonji and Matzkin (2001), and Imbens and Newey (2001) consider identification and estimation of other nonparametric structural models. Recently, Darolles, Florens, and Renault (2000) have developed a kernel estimator for a special case of our model and derived convergence rates. Also, Ai and Chen (2001) obtained estimators for parameters of semiparametric models that are $\sqrt{n}$-consistent and asymptotically efficient with conditional moment restrictions.

2. THE CONDITIONAL MEAN MODEL

We focus on identification of a model of the form

$$(2.1) \qquad y = g_0(x, z_1) + \varepsilon, \qquad E[\varepsilon|z] = 0, \qquad z = (z_1, z_2),$$

where y is an observable scalar random variable, $g_0(\cdot)$ denotes the true, unknown structural function of interest, x is a $d_x \times 1$ vector of explanatory variables, z_1 and z_2 are $d_1 \times 1$ and $d_2 \times 1$ vectors of instruments variables, and ε is a disturbance. This model reduces to the usual nonparametric regression model when $x = z_2$, but otherwise allows x to be endogenous.

The conditional expectation of equation (2.1) yields the integral equation

$$(2.2) \qquad \pi(z) \equiv E[y|z] = E[g_0(x, z_1)|z] = \int g_0(x, z_1) F(dx|z),$$

where F denotes the conditional c.d.f. of x given z. The functions π and F constitute a nonparametric generalization of the reduced form for y and x. Since π and F are functionals of the distribution function for the observable random vector (y, x, z), they are identified; identification of g_0 thus depends on the existence of a unique solution to the integral equation (2.2). This uniqueness is equivalent to completeness in z_2 of the conditional distribution of x given z, a concept we borrow from the literature on minimum variance unbiased estimation. By subtracting equation (2.2) from the same equation with $\tilde{g}(x, z_1)$ substituted for $g_0(x, z_1)$, it is easily seen that identification is equivalent to the nonexistence of any function $\delta(x, z_1) \equiv \tilde{g}(x, z_1) - g_0(x, z_1) \neq 0$ such that $E[\delta(x, z_1)|z] = 0$.

PROPOSITION 2.1: *If equation* (2.1) *is satisfied, then* g_0 *is identified if and only if for all* $\delta(x, z_1)$ *with finite expectation,* $E[\delta(x, z_1)|z] = 0$ *implies* $\delta(x, z_1) = 0$.

A sufficient condition for identification can be obtained from the well-known completeness property of exponential families.

THEOREM 2.2: *If equation* (2.1) *is satisfied, with probability one conditional on z the distribution of x is absolutely continuous with density* $f(x|z) = s(x, z_1)t(z)\exp\{\mu(z)' \times \tau(x, z_1)\}$, $s(x, z_1) > 0$, $\tau(x, z_1)$ *is one-to-one in x, and the support of* $\mu(z)$ *given* z_1 *is an open set, then* $g_0(x, z_1)$ *is identified.*

An example that helps illustrate the connection with the parametric linear model is the conditional normal case.

THEOREM 2.3: *If equation* (2.1) *is satisfied, with probability one conditional on z the distribution of x is* $N(\Psi(z_1) + \Gamma(z_1)z_2, \Omega(z_1))$, $\Omega(z_1)$ *is nonsingular, and the support of* z_2 *given* z_1 *contains an open set, then* $g_0(x, z_1)$ *is identified if and only if* $\Pr(\text{rank}(\Gamma(z_1)) = d_x) = 1$.

This result is the nonparametric analog of the necessary and sufficient conditions for identification under conditional normality in a linear model, with the matrix $\Gamma(z_1)$ being the analog of the coefficients of the excluded instruments in the reduced form for the right-hand side variables. A necessary order condition for identification in this case is that $d_2 \geq d_x$, just as in a linear model. We conjecture that such a necessary condition holds more generally.

The completeness condition also is useful when both x and z_2 are discrete with finite support $\{x_1, \dots, x_s\}$ and $\{z_{21}, \dots, z_{2t}\}$. Let $P(z_1)$ be the $s \times t$ matrix with $P(z_1)_{jk} = \Pr(x = x_j \mid z_2 = z_{2k}, z_1)$.

THEOREM 2.4: *If equation* (2.1) *is satisfied and x and* z_2 *have finite support, then* $g_0(x, z_1)$ *is identified if and only if* $\Pr(\text{rank}(P(z_1)) = s) = 1$.

The rank condition here implies the order condition that $t \geq s$. Das (1999) has considered estimation in this setting.

In many settings various generalizations of model (2.1) are useful, including semiparametric models and those where the residual is nonlinear in unknown functions. Examples include the measurement error model of Hausman et al. (1991a, b) and the partially linear models with endogeneity of Newey, Powell, and Vella (1999) and Ai and Chen (2001). To include such cases it is important to consider a generalization of equation (2.1),

$$(2.3) \qquad E[\rho(y, x, \theta_0)|z] = 0,$$

where $\theta = (\beta', g_1, \ldots, g_L)'$ for a vector of parameters β and functions g_ℓ ($\ell = 1, \ldots, L$) and $\rho(y, x, \theta)$ is a vector of residuals. Identification of θ_0 in this general model is difficult to analyze, just as for parametric nonlinear models. Generally, the only primitive conditions will be local ones, as in Florens (2000). However, given identification it is straightforward to obtain consistency, and we follow that approach.

3. NONPARAMETRIC TWO-STAGE LEAST SQUARES

We consider estimation of g_0 from equation (2.1) using i.i.d. data $((y_i, x_i, z_i), i = 1, \ldots, n)$. To motivate the estimator it is helpful to work with an estimation analog of equation (2.1). For reduced-form estimators $\hat{\pi}$ and $\widehat{F}$ obtained from preliminary nonparametric estimation, consider

$$\hat{\pi}(z) = \int g(x, z_1)\widehat{F}(dx|z). \tag{3.1}$$

A basic approach to estimation consists of "solving" this equation for $\hat{g}$. As outlined in the introduction, there are several difficulties with this approach, including existence, computation, and noncontinuity of $\hat{g}$ in the reduced form estimators. We deal with existence and computation by minimum distance with a linear in parameters approximation described below.

Noncontinuity of $\hat{g}$ in the reduced form estimators is the biggest obstacle to overcome. The lack of continuity of $\hat{g}$ in $\hat{\pi}$ and $\widehat{F}$ means that small inaccuracies in the reduced form estimates can translate into large inaccuracies in $\hat{g}$. Thus, unlike most other estimation problems, consistency of $\hat{g}$ does not automatically follow from consistency of $\hat{\pi}$ and $\widehat{F}$. This "ill-posed inverse" problem is more apparent using a linear-in-parameters (i.e., series) approximation for g_0, which we will adopt for our estimation approach. Let $w = (x, z_1)$ be the $d = d_x + d_1$ dimensional vector of all right-hand-side variables. Suppose the structural function $g_0(w)$ can be approximated as

$$g_0(w) \cong \sum_{j=1}^{J} \gamma_j p_j(w),$$

where $\{p_1(w), p_2(w), \ldots\}$ is a sequence of "basis" functions, and γ is a corresponding vector of coefficients. Substitution into equation (2.2) yields

$$E[y|z] \cong \sum_{j=1}^{J} \gamma_j \int p_j(w)F(dx|z) = \sum_{j=1}^{J} \gamma_j E[p_j(w)|z].$$

This equation suggests a two-stage estimation procedure, with the conditional expectations $E[p_j(w)|z]$ being estimated (by nonparametric regression) in the first stage, followed by a second-stage regression of y on the estimator of $E[p_j(w)|z]$ to estimate the γ coefficients. However, the "true" second-stage regressors $E[p_j(w)|z]$ may not have much variance even when the basis functions do. Indeed the essence of the noncontinuity problem lies in the existence (under certain regularity conditions) of a basis $p_j(w)$ that is orthogonal for $E[\cdot]$, $E[p_j(w)|z]$ being orthogonal also, with $E[p_j(w)^2] = 1$ but $\lim_{j\to\infty} E[E[p_j(w)|z]^2] = 0$; e.g., see Kress (1989, p. 235). This property makes the second stage sensitive to the number of approximating functions J and the precision of the first-stage estimators.

In the literature on integral equations various methods of dealing with noncontinuity have been proposed, often referred to as "regularization." Generally they consist of careful choice of a minimum distance problem to be solved. Our approach is to focus on the case where g_0 is known to belong to a compact set, and to restrict the estimator $\hat{g}$ to belong to this set. This approach eliminates the ill-posed inverse problem essentially because integration is a continuous mapping, so by compactness the inverse is continuous (e.g., see Theorem 5.6 of Munkres (1975)). It can also be viewed as a "regularization" method (Kress (1989)), although for our purposes it is more than that, as we restrict the true structural function to be in the compact set.

We will consider functions of the form $a(w)'\beta + g_1(w)$, where $a(w)$ and β are $r \times 1$ vectors of known functions and unknown parameters respectively, where bounds are placed on β, and where $g_1(w)$ and its derivatives are required to be small in the tails. Thus the unknown function is allowed to be nonparametric over the "middle" of the distribution but is restricted to be parametric in the tails. This specification allows for unbounded w_i, which is difficult to do for a compact set of functions. Alternatively, if w_i is bounded, one can drop the parametric part and obtain a full nonparametric specification. We have adopted this specification because it seems important to allow for unbounded endogenous variables in many applications.

To be precise about the compactess restrictions some additional notation is needed. Let λ denote a $d \times 1$ multi-index (vector of nonnegative integers), $|\lambda| = \sum_{\ell=1}^{d} \lambda_\ell$, $w^\lambda \equiv \prod_{i=1}^{d}(w_i)^{\lambda_i}$, and $D^\lambda g_1(w) = \partial^{|\lambda|} g_1(w)/\partial w_1^{\lambda_1} \cdots \partial w_d^{\lambda_d}$. Also, let $m_0 > d/2$, $\delta_0 > d/2$, and m denote positive integers. Assuming that the mean μ_w and variance Σ_w of w_i exists, and Σ_w is nonsingular, let $\widetilde{w} = \Sigma_w^{-1/2}(w - \mu_w)$. Define

$$\|g_1\|_1 = \left\{ \sum_{|\lambda| \le m+m_0} \int \left[D^\lambda g_1(\widetilde{w})\right]^2 \cdot (1 + \widetilde{w}'\widetilde{w})^{\delta_0} dw \right\}^{1/2}.$$

Let B_1 and B_β be known positive constants and $\mathcal{G}_1 = \{g_1(w) : (\|g_1\|_1)^2 \le B_1\}$. The set of functions that we consider is

$$\mathcal{G} = \left\{a(w)'\beta + g_1(w) : \beta'\beta \le B_\beta, g_1 \in \mathcal{G}_1\right\}.$$

We will assume that $g_0(w) = a(w)'\beta_0 + g_{10}(w) \in \mathcal{G}$ and also impose the restriction that the estimator is an element of $\mathcal{G}$. For any δ with $d/2 < \delta < \delta_0$ let $\|g_1\|_{1\delta} = \max_{|\lambda| \le m} \sup_w |D^\lambda g_1(\widetilde{w})|(1 + \widetilde{w}'\widetilde{w})^\delta$. Define the norm

$$\|g\| = \sqrt{\beta'\beta} + \|g_1\|_{1\delta}.$$

Then the closure $\overline{\mathcal{G}}$ of $\mathcal{G}$ with respect to the norm $\|g\|$ will be compact for the norm $\|g\|$, as follows from Gallant and Nychka (1987). Consistency will be shown for this norm, meaning that $\|\hat{g} - g_0\| \xrightarrow{p} 0$.

Using a finite dimensional approximation to g_1 aids in computation. Let $\hat{\mu}_1$ and $\widehat{\Sigma}_1$ be the sample mean and variance of w respectively and $\widehat{w} = \widehat{\Sigma}_1^{-1/2}(w - \hat{\mu}_1)$. We consider a Hermite polynomial approximation to g_1 of the form

$$g_1(w) \cong \sum_{j=1}^{J} \gamma_j p_j(\widehat{w}), \qquad p_j(w) \equiv \exp\{-w'w\} \cdot w^{\lambda(j)}, \tag{3.2}$$

where $|\lambda(j)|$ is increasing in j. The compactness restriction will be imposed by bounding the coefficients γ. Let $\gamma = (\gamma_1, \ldots, \gamma_J)'$ and

$$\Lambda_J = \sum_{|\lambda| \leq m+m_0} \int \left[D^\lambda p^J(\widehat{w}) D^\lambda p^J(\widehat{w})'\right] \cdot (1 + \widehat{w}'\widehat{w})^{\delta_0}\, dw,$$

where $p^J(\widehat{w}) = (p_1(\widehat{w}), \ldots, p_J(\widehat{w}))'$. The restriction that $\sum_{j=1}^J \gamma_j p_j(\cdot) \in \mathcal{G}_1$ is then the same as the quadratic inequality restriction $\gamma'\Lambda_J\gamma \leq B_1$.

For estimation substitute equation (3.2) into equation (3.1) to obtain

$$\hat{\pi}(z) \cong \widehat{E}[a|z]'\beta + \sum_{j=1}^{J} \gamma_j \widehat{E}[p_j|z], \tag{3.3}$$

where $\widehat{E}[(\cdot)|z] = \int(\cdot)\widehat{F}(dx|z)$. An objective function measuring the distance between the left and right-hand sides can be obtained as the sum of squared differences evaluated at the observations for z. Let

$$\widetilde{Q}(\beta, \gamma) = \sum_{i=1}^{n} \left\{ y_i - \widehat{E}[a|z_i]'\beta - \sum_{j=1}^{J} \gamma_j \widehat{E}[p_j|z_i] \right\}^2 \Big/ n. \tag{3.4}$$

This objective function is a nonparametric analog of the two-stage least squares objective function, with right-hand side variables being conditional expectation estimators, rather than predicted values from a parametric regression, and the parameters being those of a functional approximation rather than the true parameters of the model.[2] A nonparametric two-stage least squares estimator can be obtained by minimizing this objective function subject to restrictions on the parameters. That is,

$$\begin{aligned} &\hat{g}(w) = a(w)'\hat{\beta} + \hat{g}_1(w), \qquad \hat{g}_1(w) = \sum_{j=1}^{J} \hat{\gamma}_j p_j(\widehat{w}), \\ &(\hat{\beta}, \hat{\gamma}) = \operatorname{argmin} \widetilde{Q}(\beta, \gamma) \quad \text{subject to} \quad \beta'\beta \leq B_\beta, \quad \gamma'\Lambda_J\gamma \leq B_1. \end{aligned} \tag{3.5}$$

Computation of this estimator is straightforward, because it is the solution to a quadratic programming problem.

[2] The use of y_i here rather than $\hat{\pi}(z_i)$ does not affect the estimator when the first stage is the series estimator described below.

The coefficients $\hat{\beta}$ and $\hat{\gamma}$ have a ridge regression form. Let $Y = (y_1, \ldots, y_n)'$, $\widehat{R}_i = (\widehat{E}[a|z_i]', \widehat{E}[p_1|z_i], \ldots, \widehat{E}[p_J|z_i])'$, $\widehat{R} = [\widehat{R}_1, \ldots, \widehat{R}_n]'$, $\hat{\zeta}_\beta$ and $\hat{\zeta}_g$ be the Lagrange multipliers associated with the constraints. Also let $S_J = \mathrm{diag}[\hat{\zeta}_\beta I, \hat{\zeta}_g \Lambda_J]$. The first-order condition is given by

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = (\widehat{R}'\widehat{R} + S_J)^{-1}\widehat{R}'Y.$$

To complete the description of this estimator we need to specify the first-stage nonparametric regression estimators $\widehat{E}[a|z_i]$ and $\widehat{E}[p_j|z_i]$. Here we consider series estimators.[3] Let $q^K(z) = (q_{1K}(z), \ldots, q_{KK}(z))'$ denote approximating functions, such as power series or splines, $q_i^K = q^K(z_i)$, and $\widehat{M} = \sum_{i=1}^n q_i^K q_i^{K\prime}/n$. The first-stage estimators are

$$\begin{aligned} &\widehat{E}[a'|z_i] = q_i^{K\prime}\widehat{M}^- \sum_{\ell=1}^n q_\ell^K a(w_\ell)'/n, \\ (3.6)\qquad &\widehat{E}[p_j|z_i] = q_i^{K\prime}\widehat{M}^- \sum_{\ell=1}^n q_\ell^K p_j\big(\widehat{\Sigma}_1^{-1/2}(w_\ell - \hat{\mu}_1)\big)/n, \end{aligned}$$

where $\widehat{M}^-$ denotes a generalized inverse, with $\widehat{M}\widehat{M}^-\widehat{M} = \widehat{M}$. The NP2SLS estimator can be obtained by substituting these conditional expectation estimators in equation (3.4) and carrying out the minimization in equation (3.5).

The NP2SLS estimator depends on the choice of J, B_β, B_1, and K. Together J and B_1 determine how flexible the approximation is. For consistency it is important that B_1 be set large enough so that the true structural function satisfies a constraint analogous to that for the estimator and that J grows with the sample size. However, if J and B_1 are too large, then the estimator may be too variable, particularly in view of the ill-posed inverse problem. Choosing K too large could also make the estimator too variable. It would be useful to have data-based methods for these choices, but derivation of these is beyond the scope of this paper. Of course, as always it is possible to conduct a sensitivity analysis by varying these choices.

This estimator can be modified to apply to a partially linear model by restricting the hermite polynomial approximation to be a function of those components w_b of w that enter nonparametrically. In that case the "trend" part $a(w)'\beta$ includes the parametric components as well as the leading terms for the nonparametric one. The estimator can be formed exactly as described above with w_b replacing w in the calculation of $\widehat{w}$, $p_j(\widehat{w})$, etc.

The estimator can be extended to estimate the general conditional moment restriction model of equation (2.3), using a minimum distance approach like that of Malinvaud (1980). Let $\hat{\rho}(z_i, \theta) = [\sum_{j=1}^n \rho(y_j, x_j, \theta)q_j^K/n]\widehat{M}^- q^K(z_i)$ be a series estimator of $E[\rho(y, x, \theta)|z = z_i]$. Suppose that each function g_ℓ in the vector $\theta = (\beta', g_1, \ldots, g_L)'$ depends on a vector w^ℓ of variables. Let $\hat{\mu}_\ell$ and $\widehat{\Sigma}_\ell$ be preliminary mean and variance estimators, $\widehat{w}^\ell = \widehat{\Sigma}_\ell^{-1/2}(w^\ell - \hat{\mu}_\ell)$, $\gamma^\ell = (\overset{\ell}{\gamma_1}, \ldots, \overset{\ell}{\gamma_{J_\ell}})'$, $g_\ell(\gamma^\ell) = \sum_{j=1}^{J_\ell} \gamma_j^\ell p_j(\widehat{w}^\ell)$, $g(\gamma) = (g_1(\gamma^1), \ldots, g_L(\gamma^L))'$, and $\Lambda_{J_\ell}^\ell$ be as defined above with w^ℓ replacing w. Let

[3] Results for nearest neighbor estimators are given in Newey and Powell (1988).

B_β and B_ℓ ($\ell = 1, \dots, L$) be prespecified bounds analogous to those given above and $\widehat{A}$ be a positive definite matrix, and consider the objective function

$$\widehat{Q}(\theta) = \sum_{i=1}^{n} \hat{\rho}(z_i, \theta)' \widehat{A} \hat{\rho}(z_i, \theta).$$

The estimator is given by

$$\begin{aligned} &(\hat{\beta}, g(\hat{\gamma})) = \operatorname*{argmin}_{\beta,\gamma} \widehat{Q}(\beta, g(\gamma)) \quad \text{subject to} \\ (3.7)\qquad &\beta'\beta \le B_\beta, \quad \gamma^{\ell\prime} \Lambda_{J_\ell}^{\ell} \gamma^{\ell} \le B_\ell \qquad\qquad (\ell = 1, \dots, L). \end{aligned}$$

This estimator is a nonparametric minimum distance estimator, where the distance measure is a sample average over the conditioning variables of a quadratic form in conditional expectation estimators.[4]

4. CONSISTENCY

We will first give a consistency theorem for the setting of equation (2.3) for a general norm $\|\cdot\|$ and compact set Θ. Let $\hat{\rho}(z_i, \theta) = [\sum_{j=1}^{n} \rho(y_j, x_j, \theta) q_j^K / n] \cdot \widehat{M}^{-} q^K(z_i)$ and $\widehat{A}$ be a positive definite matrix. Also, let Θ_J be a subset of Θ that can be thought of as a finite dimensional approximation to Θ in our setting. The estimator we consider takes the form

$$(4.1)\qquad \hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta_J} \sum_{i=1}^{n} \hat{\rho}(z_i, \theta)' \widehat{A} \hat{\rho}(z_i, \theta).$$

This set up is general enough to include those we have already considered and has some independent interest, e.g., as in Ai and Chen (2001).

The first condition imposes identification.

ASSUMPTION 1: *$\theta_0 \in \Theta$ is the only $\theta \in \Theta$ satisfying $E[\rho(y, x, \theta)|z] = 0$.*

The next condition requires that the first-stage series approximation can approximate any function with finite mean-square, which is known to hold for power series and splines, and that the distance matrix $\widehat{A}$ has a constant limit A:

ASSUMPTION 2: *For any $b(z)$ with $E[b(z)^2] < \infty$ there is π_K with $E[\{b(z) - q^K(z)'\pi_K\}^2] \to 0$, $K \to \infty$, and $K/n \to 0$. Also, $\widehat{A} \xrightarrow{p} A$, and A is positive definite and constant.*

The consistency result will take the form $\|\hat{\theta} - \theta_0\| \xrightarrow{p} 0$ for a norm $\|\cdot\|$. The next condition requires a Hölder continuity property for the residual $\rho(y, x, \theta)$ in this norm.

[4] This estimator includes NP2SLS as a special case when y_i is replaced with $\hat{\pi}(z_i)$ in the objective function.

ASSUMPTION 3: *$E[\|\rho(y,x,\theta_0)\|^2|z]$ is bounded and there exists $M(y,x)$, $\nu > 0$ such that for all $\tilde{\theta}$, $\theta \in \Theta$, $\|\rho(y,x,\tilde{\theta}) - \rho(y,x,\theta)\| \leq M(y,x)\|\tilde{\theta} - \theta\|^{\nu}$ and $E[M(y,x)^2|z]$ is bounded.*

Two additional conditions are needed for consistency in the setting we consider. One is that the parameter set be compact for the norm $\|\theta\|$.

ASSUMPTION 4: *$\theta_0 \in \Theta$, and Θ is compact for the norm $\|\theta\|$.*

The other one is that the approximating subspaces are dense for $\|\theta\|$.

ASSUMPTION 5: *For any $\theta \in \Theta$ there exists $\theta_J \in \Theta_J$ such that* $\lim_{J\to\infty} \|\theta_J - \theta\| = 0$.

One could also obtain consistency for the estimator that minimized the objective function over all of Θ rather than Θ_J. The purpose of considering Θ_J is to simplify computation, especially by choosing Θ_J to be finite dimensional.

A general consistency result follows from these conditions.

THEOREM 4.1: *If Assumptions* 1–5 *are satisfied and $J \to \infty$, then $\|\hat{\theta} - \theta_0\| \xrightarrow{p} 0$.*

We can use the results of Gallant and Nychka (1987) to verify the conditions of the general Theorem 4.1 for the estimator of equation (3.7). The consistency norm is related to the constraints we have imposed on β and γ. Specifically, for $\theta = (\beta', g_1, \dots, g_L)'$, the domain $\mathcal{W}_\ell$ of g_ℓ, and positive constants $d_\ell/2 < \delta_\ell < \delta_{\ell 0}$, m_ℓ, and $m_{0\ell} > d_\ell/2$, let

$$\begin{aligned} &\|\theta\| = (\beta'\beta)^{1/2} + \sum_{\ell=1}^{L} \|g_\ell\|_\ell, \\ &\|g_\ell\|_\ell = \max_{|\lambda|\leq m_\ell} \sup_{w_\ell \in \mathcal{W}_\ell} |D^\lambda g_\ell(w_\ell)|(1 + w_\ell' w_\ell)^{\delta_\ell}. \end{aligned} \tag{4.2}$$

ASSUMPTION 6: *$\Theta_J = (\beta, g(\gamma))$ such that $\beta \in B = \{\beta : \beta'\beta \leq B_\beta\}$, and $g_\ell(\gamma^\ell) \in \mathcal{G}_\ell = \{g_\ell(w_\ell) : \sum_{|\lambda|\leq m_\ell + m_{\ell 0}} \int [D^\lambda g_\ell(w_\ell)]^2 \cdot (1 + w_\ell' w_\ell)^{\delta_{\ell 0}}\, dw_\ell \leq B_\ell\}$, $\mathcal{W}_\ell$ is open and convex.*

For simplicity we assume $\hat{\mu}_\ell = 0$ and $\widehat{\Sigma}_\ell = I$. Let $\overline{\mathcal{G}}_\ell$ denote the closure of $\mathcal{G}_\ell$ in the norm $\|g_\ell\|_\ell$.

THEOREM 4.2: *If Assumptions* 1–3 *are satisfied for $\Theta = B \times \overline{\mathcal{G}}_1 \times \cdots \times \overline{\mathcal{G}}_L$, Assumption 6 is satisfied, and $J_\ell \to \infty$ $(\ell = 1, \dots, L)$, then $\hat{\theta} = (\hat{\beta}, g(\hat{\gamma}))$ from equation* (3.7) *satisfies $\|\hat{\theta} - \theta_0\| \xrightarrow{p} 0$.*

It is straightforward to specialize this result to the NP2SLS estimator, where $\rho(y,x,\theta) = y - a(w)'\beta - g_1(w)$.

THEOREM 4.3: *If the conditional distribution of x given z is complete in z_2, Assumption 2 is satisfied, $g_0(w) \in \mathcal{G}$, $E[\|a(w)\|^2] < \infty$, for any $\beta \neq 0$ we have $a(w)'\beta \notin \mathcal{G}_1$, the interior of the support of w_i is convex, and $J \to \infty$, then $\hat{\beta} \xrightarrow{p} \beta_0$ and for any $d/2 < \delta < \delta_0$,*

$$\max_{|\lambda| \le m} \sup_{w \in \mathcal{W}} \left| D^\lambda \sum_{j=1}^{J} \hat{\gamma}_j p_j(w) - D^\lambda g_{10}(w) \right| (1 + w'w)^\delta \xrightarrow{p} 0.$$

This result shows consistency of the NP2SLS estimator, in the sense that the trend component satisfies $\hat{\beta} \xrightarrow{p} \beta_0$ and that $\sum_{j=1}^{J} \hat{\gamma}_j p_j(w)$ and its derivatives up to order m converge in probability to $g_{10}(w)$ and its derivatives.

5. SIMULATION RESULTS

To investigate the practical applicability of this consistency result in finite samples, we conducted a small-scale simulation study of the sampling distribution of the estimator. Only results for a single design are reported here; results for other designs were qualitatively similar.

Our design used a simple specification for the structural function $g(x)$ and (scalar) regressor x, with

$$y = g(x) + u = \ln(|x - 1| + 1)\,\mathrm{sgn}(x - 1) + u,$$
$$x = z + v,$$

where the errors u and v and instrument z are generated as

$$\begin{pmatrix} u \\ v \\ z \end{pmatrix} \sim \text{i.i.d. } \mathcal{N}\left(0, \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right).$$

The Hermite series approximation of (3.2) used $J = 5$ terms while $a(w) = x$. The instruments $q^K(z)$ were chosen to be a cubic spline with 5 knots. For this design, 500 replications of the estimator $\hat{g}(x)$ were generated for two sample sizes, $N = 100$ and $N = 400$, and for two values of the constraint parameter for the functional compactness restriction, $B_1 = 5$ and $B_1 = 50$.

Results for this simple 2 × 2 design are summarized in Table I, which gives the root mean-squared-error (RMSE), averaged across the realized values of x and the 500 replications. For each value of the constraint coefficient, these RMSE's decline as the sample size is quadrupled, but at somewhat less than a $\sqrt{N}$-rate, as would be expected for this nonparametric estimation problem. The results in this table show considerable

TABLE I
RMSE

	$B_1 = 5$	$B_1 = 50$
$N = 100$	0.277	0.446
$N = 400$	0.208	0.356

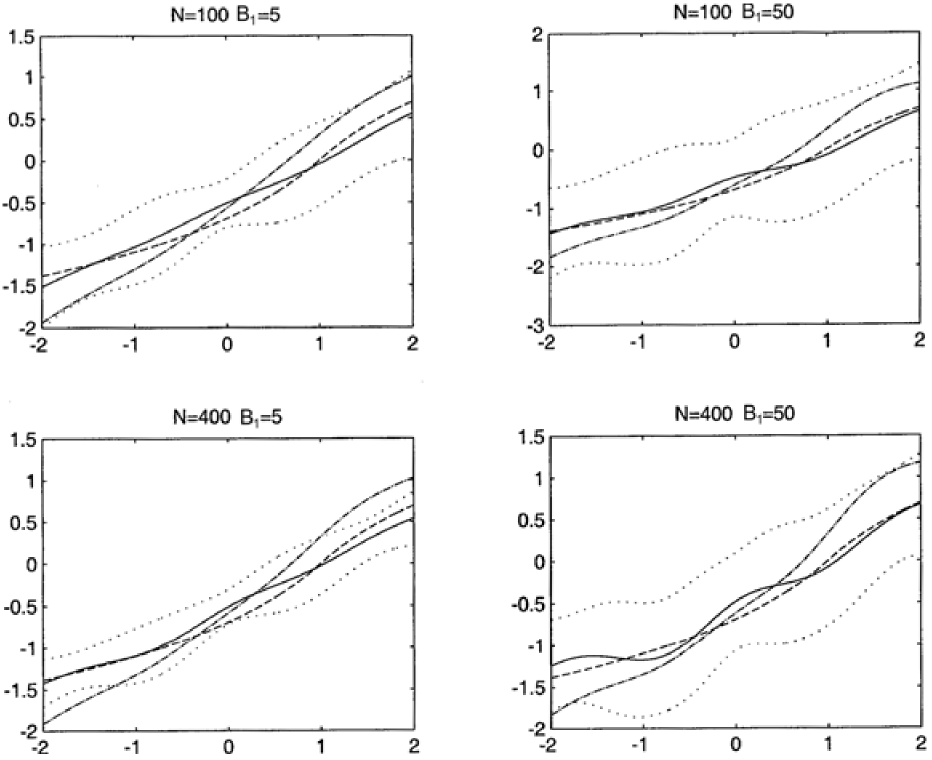

FIGURE 1.—Function value estimates for Monte Carlo.

sensitivity of the RMSE's to the variation in the constraint parameter B_1 for these designs. This sensitivity is also evident in Figure 1, which graphs the average value of the function estimates $\hat{g}$ (solid line) against the true value of $g(x)$ (dashed line); the dotted lines in these graphs plot the upper and lower two standard deviation limits for the simulation distributions of $\hat{g}$, which are considerably wider for the larger value of B_1.

The figure also plots average values of an estimator $\widehat{E}[y|x]$ (dotted-and-dashed line) of the conditional mean $E[y|x]$ of y given x, using the same form of the series approximation as for $\hat{g}(x)$ (but without the first-stage nonparametric fitting of the series terms). Without the correction for endogeneity, the average value of $\widehat{E}[y|x]$ is tilted upward relative to the true structural function $g(x)$, and occasionally strays outside the dotted standard deviation lines for $\hat{g}(x)$. In contrast, the average value of $\hat{g}(x)$ tracks the true function $g(x)$ much more closely, except for an anomalous "bump" near $x = 0$, which may be due, in part, to the Hermite form of the series approximation with $J = 5$. Though hardly definitive, these results suggest that our theoretical consistency result may be relevant in practice, though judicious choice of the "smoothing" parameter B_1 is no less important for this estimator than for other nonparametric estimation problems.

Department of Economics, MIT, E52-252D, Cambridge, MA 02139; wnewey@mit.edu
and
Department of Economics, UC Berkeley, Berkeley, CA 94720; powell@econ.berkeley.edu.

Manuscript received November, 2000; final revision received July, 2002.

APPENDIX: PROOFS

We first state and prove some lemmas that are useful for our consistency results. The first result is a convergence in probability version of Gallant's (1987) consistency result and the second is a slightly improved version of Corollary 2.2 of Newey (1991). In these results $\widehat{Q}(\theta)$ is a general objective and not just the one considered in the body of the paper. In the first one $\widehat{\Theta}$ is an approximating set that can be thought of as a finite dimensional subset of Θ.

LEMMA A1: *Suppose* (i) *$Q(\theta)$ has a unique minimum on Θ at θ_0*; (ii) *$\widehat{Q}(\theta)$ and $Q(\theta)$ are continuous*, *Θ is compact*, *and* $\max_{\theta\in\Theta} |\widehat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0$; (iii) *$\widehat{\Theta}$ are compact subsets of Θ such that for any $\theta \in \Theta$ there exists $\tilde{\theta} \in \widehat{\Theta}$ such that $\tilde{\theta} \xrightarrow{p} \theta$. Then* $\hat{\theta} = \operatorname{argmin}_{\theta\in\widehat{\Theta}} \widehat{Q}(\theta) \xrightarrow{p} \theta_0$.

PROOF: Consider any neighborhood $\mathcal{N}$ of θ_0. By compactness, continuity of $Q(\theta)$, and $Q(\theta)$ having a unique minimum at θ_0,

$$\Delta = \Big[\min_{\theta\in\Theta\cap\mathcal{N}^c} Q(\theta)\Big] - Q(\theta_0) > 0.$$

By (iii) there is $\tilde{\theta} \in \widehat{\Theta}$ such that $\tilde{\theta} \xrightarrow{p} \theta_0$. By the definition of $\hat{\theta}$, $\widehat{Q}(\hat{\theta}) \leq \widehat{Q}(\tilde{\theta})$, so that by the uniform convergence hypothesis in (ii), $Q(\hat{\theta}) < Q(\tilde{\theta}) + \Delta/2$ w.p.a. 1. Furthermore, by the definition of $\tilde{\theta}$ and continuity of $Q(\theta)$, $Q(\tilde{\theta}) < Q(\theta_0) + \Delta/2$ w.p.a. 1. Then by summing the two inequalities and subtracting $Q(\tilde{\theta})$ from both sides, $Q(\hat{\theta}) < Q(\theta_0) + \Delta$ w.p.a. 1. By the definition of Δ, this event can only happen when $\hat{\theta} \in \mathcal{N}$, which thus occurs w.p.a. 1. The conclusion follows by the $\mathcal{N}$ being any neighborhood of θ_0. *Q.E.D.*

LEMMA A2: *If* (i) *Θ is a compact subset of a space with norm $\|\theta\|$*; (ii) *$\widehat{Q}(\theta) \xrightarrow{p} Q(\theta)$ for all $\theta \in \Theta$*; (iii) *there is $v > 0$ and $B_n = O_p(1)$ such that for all θ, $\tilde{\theta} \in \Theta$, $|\widehat{Q}(\theta) - \widehat{Q}(\tilde{\theta})| \leq B_n\|\theta - \tilde{\theta}\|^\nu$, then $Q(\theta)$ is continuous and* $\sup_{\theta\in\Theta} |\widehat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0$.

PROOF: Consider any fixed $\tilde{\theta}$ and $\varepsilon > 0$. There exists M such that $\Pr(B_n/M \leq 1) > 0$ for all n. Consider $\Delta = (\varepsilon/2M)^{1/\nu}$. Note that for all θ with $\|\theta - \tilde{\theta}\| \leq \Delta$ we have $|\widehat{Q}(\theta) - \widehat{Q}(\tilde{\theta})| \leq B_n\Delta^\nu = B_n\varepsilon/2M \leq \varepsilon/2$ with positive probability. Then by the triangle inequality and (ii),

$$|Q(\theta) - Q(\tilde{\theta})| \leq |\widehat{Q}(\theta) - Q(\theta)| + |\widehat{Q}(\tilde{\theta}) - Q(\tilde{\theta})| + |\widehat{Q}(\theta) - \widehat{Q}(\tilde{\theta})| \leq \varepsilon$$

with positive probability. It then follows by $|Q(\theta) - Q(\tilde{\theta})|$ constant that $|Q(\theta) - Q(\tilde{\theta})| \leq \varepsilon$. Hence $Q(\theta)$ is continuous at $\tilde{\theta}$, and since $\tilde{\theta}$ is arbitrary, $Q(\theta)$ is continuous on Θ. The other conclusion then follows by Corollary 2.2 of Newey (1991). *Q.E.D.*

PROOF OF THEOREM 2.2: By hypothesis, with probability one the conditional density of x given z belongs to an exponential family with "parameter" vector $\mu(z_1, z_2)$ that varies over an open set. Theorem 1, p. 132 of Lehman (1959) gives the conclusion. *Q.E.D.*

PROOF OF THEOREM 2.3: By normality the density of x given z is as in Theorem 2.2 with

$$\begin{aligned} t(z) &= (2\pi)^{-d_x/2}\det(\Omega(z_1))^{-1/2} \\ &\qquad \times \exp\big\{-(\Psi(z_1)+\Gamma(z_1)z_2)'\Omega(z_1)^{-1}(\Psi(z_1)+\Gamma(z_1)z_2)/2\big\}, \\ s(x, z_1) &= \exp\{-x'\Omega(z_1)^{-1}x/2\}, \qquad \tau(x, z_1) = \Omega(z_1)^{-1}x, \\ \mu(z) &= \Psi(z_1)+\Gamma(z_1)z_2. \end{aligned}$$

Note that when $\operatorname{rank}(\Gamma(z_1)) = d_x$, then $\mu(z)$ maps open z_2 sets into open sets for each given z_1, so that the "if" conclusion follows by Theorem 2.2. Also if $\operatorname{rank}(\Gamma(z_1)) < d_x$ there is $\alpha(z_1)$ such that $\alpha(z_1)'\Gamma(z_1) = 0$ and $\Pr(\alpha(z_1) = 0) < 1$. Consider $a = \alpha(z_1)'(x - \Psi(z_1))$. Note that $\operatorname{var}(a|z) = \alpha(z_1)'\Omega(z_1)\alpha(z_1) > 0$ with positive probability, so that $a \neq 0$. Then $E[a|z] = \alpha(z_1)'\Gamma(z_1)z_2 = 0$, so that g_0 is not identified by Proposition 2.1. *Q.E.D.*

PROOF OF THEOREM 4.1: The proof will proceed by verifying the hypotheses of Lemma A1. For (i), note that by Assumptions 3 and 4, there is $\widetilde{M}(y, x)$ with

$$\begin{aligned} \|\rho(y, x, \theta)\| &\le \|\rho(y, x, \theta_0)\| + \|\rho(y, x, \theta) - \rho(y, x, \theta_0)\| \\ &\le \|\rho(y, x, \theta_0)\| + M(y, x)\|\theta - \theta_0\|^{\nu} \\ &\le \|\rho(y, x, \theta_0)\| + M(y, x)C = \widetilde{M}(y, x) \end{aligned}$$

and $E[\widetilde{M}(y, x)^2|z]$ is bounded. Let $\bar\rho(z, \theta) = E[\rho(y, x, \theta)|z]$. Then by the Cauchy–Schwartz inequality,

$$E\big[\|\bar\rho(z, \theta)\|^2\big] \le E\big[E[\|\rho(y, x, \theta)\|^2|z]\big] \le E\big[\widetilde{M}(y, x)^2\big] < \infty.$$

Let $Q(\theta) = E[\bar\rho(z, \theta)'A\bar\rho(z, \theta)]$. By Assumption 1, $Q(\theta_0) = 0$ and by A positive definite, $Q(\theta) > Q(\theta_0)$ for $\theta \neq \theta_0$, showing (i) of Lemma A1. For the series estimator of $\bar\rho(z, \theta)$, hypothesis (ii) follows by Corollary 4.2 of Newey (1991). Finally, (iii) of Lemma A1 follows by choosing $\theta_J \in \Theta_J$ such that $\|\theta_J - \theta\| \to 0$. *Q.E.D.*

PROOF OF THEOREM 4.2: We note that the constraints $\gamma^{\ell\prime}\Lambda^{\ell}_{J_\ell}\gamma^{\ell} \le B_\ell$ are equivalent to $g_\ell(\gamma^\ell) \in \mathcal{G}_\ell$. Compactness of $\overline{\mathcal{G}}_\ell$ in the norm $\|g_\ell\|_\ell$ follows by Theorem 1 of Gallant and Nychka (1987), so Assumption 4 holds by the Tychonoff Theorem. Assumption 5 follows from Theorem 2 of Gallant and Nychka (1987). *Q.E.D.*

PROOF OF THEOREM 4.3: To prove Assumption 1, note that by completeness any β and g_1 satisfying $\pi(z) = E[a(w)'\beta + g_1(w) \mid z]$ satisfies $a(w)'\beta + g_1(w) = a(w)'\beta_0 + g_{10}(w)$, implying $a(w)'(\beta - \beta_0) = g_{10}(w) - g_1(w)$. Note that by hypothesis $g_{10} \in \mathcal{G}_1$ and $g_1 \in \mathcal{G}_1$, so by the triangle inequality $(g_{10} - g_1)/2 \in \mathcal{G}_1$. It follows that $a'(\beta - \beta_0)/2 = (g_{10} - g_1)/2 \in \mathcal{G}_1$, and hence $\beta - \beta_0 = 0$. For Assumption 3, note that

$$\begin{aligned} \big|\rho(y, x, \tilde\theta) - \rho(y, x, \theta)\big| &\le \|a(w)\|\|\tilde\beta - \beta\| + |\tilde g(w) - g(w)| \\ &\le (1 + \|a(w)\|) \\ &\qquad \times \Big[\|\tilde\beta - \beta\| + \max_{|\lambda|\le m}\sup_{w\in\mathcal{W}} |D^\lambda[\tilde g(w) - g(w)]|(1 + w'w)^\delta\Big]. \end{aligned}$$

It follows that all of the Assumptions of Theorem 4.2 are satisfied, so the conclusion follows by Theorem 4.2. *Q.E.D.*

REFERENCES

AI, C., AND X. CHEN (2001): "Efficient Sieve Minimum Distance Estimation of Semiparametric Conditional Moment Models," Working Paper.

ALTONJI, J. G., AND R. L. MATZKIN (2001): "Panel Data Estimators for Nonseparable Models with Endogenous Regressors," NBER Working Paper No. TO267, March.

AMEMIYA, T. (1974): "The Nonlinear Two-Stage Least Squares Estimator," *Journal of Econometrics*, 2, 105–110.

BREIMAN, L., AND J. H. FRIEDMAN (1985): "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–598.

BROWN, D. J., AND R. MATZKIN (1998): "Estimation of Nonparametric Functions in Simultaneous Equations Models, with an Application to Consumer Demand," Cowles Foundation Working Paper, March.

CHEN, X., AND X. SHEN (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314.

DAROLLES, S., J.-P. FLORENS, AND E. RENAULT (2000): "Nonparametric Instrumental Regression," Manuscript, GREMAQ, University of Toulouse, April.

DAS, M. (1999): "Instrumental Variable Estimation of Models with Discrete Endogenous Regressors," Manuscript presented at 2000 World Congress of the Econometric Society.

FLORENS, J.-P. (2000): "Inverse Problems and Structural Econometrics: The Example of Instrumental Variables," Invited Presentation, World Congress of the Econometric Society.

GALLANT, A. R. (1987): "Identification and Consistency in Nonparametric Regression," in *Advances in Econometrics: Fifth World Congress*, ed. by T. F. Bewley. Cambridge: Cambridge University Press, 145–169.

GALLANT, A. R., AND D. W. NYCHKA (1987): "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.

HAUSMAN, J. A., H. ICHIMURA, W. K. NEWEY, AND J. L. POWELL (1991a): "Identification and Estimation of Polynomial Errors-in-Variables Models," *Journal of Econometrics*, 50, 273–295.

——— (1991b): "Nonlinear Errors in Variables," *Journal of Econometrics*, 65, 205–233.

IMBENS, G. W., AND W. K. NEWEY (2001): "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity," Preprint, March.

KRESS, R. (1989): *Linear Integral Equations*. New York: Springer-Verlag.

LEHMAN, E. L. (1959): *Testing Statistical Hypotheses*. New York: Wiley.

MALINVAUD, E. (1980): *Statistical Methods of Econometrics*. New York: North-Holland.

MAMMEN, E., O. LINTON, AND J. NIELSEN (1999): "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions," *Annals of Statistics*, 27, 1443–1490.

MUNKRES, J. R. (1975): *Topology: A First Course*. Englewood Cliffs, NJ: Prentice-Hall.

NEWEY, W. K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59, 1161–1167.

NEWEY, W. K., AND J. L. POWELL (1988): "Instrumental Variables Estimation for Nonparametric Models," Manuscript, Department of Economics, Princeton University.

NEWEY, W. K., J. L. POWELL, AND F. VELLA (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

OPSOMER, J. D., AND D. RUPPERT (1997): "Fitting a Bivariate Additive Model by Local Polynomial Regression," *Annals of Statistics*, 25, 186–211.

O'SULLIVAN, F. (1986): "Ill Posed Inverse Problems" (with discussion), *Statistical Science*, 4, 503–527.

ROEHRIG, C. S. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56, 433–447.