Instrumental Variable Estimation of Nonparametric Models*

Whitney K. Newey
Department of Economics
M.I.T.

James L. Powell
Department of Economics
Berkeley

1988, Revised January 2002

JEL Classification: C13, C30
Keywords:

[*] 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. Discussion with Angus Deaton inspired this paper; helpful comments were also provided by R. Blundell, X. Chen, L. Hansen, J. Heckman, J. Horowitz, H. Ichimura, C. Manski, C. Meghir, J. Scheinkman, G. Wahba, two referees, and participants in seminars at BellCore, and Chicago, Northwestern, Princeton,Yale, Wisconsin Universities.

1 Introduction

Nonparametric regression provides a way of uncovering the reduced-form relationship between a dependent variable and explanatory variables, without imposing functional form restrictions. In econometric applications, 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 error terms, moment restrictions on unobservable errors are often imposed to identify and consistently estimate the parameters of interest; in linear models, zero covariance between the "instruments" (conditioning variables) and disturbances, 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 may be 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. Existence is easy to obtain using nonparametric extensions of minimum distance methods that are standard approaches in parametric models. 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 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 the integral bounds on derivates 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 instrumental variable identification. This work, as well as more recent work by Opsomer and Rupert (1997) and Mammen, Linton, and Nielsen (1999) is about solving nonlinear integral equations that have a slightly different structure than those given here.

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 (1988), Altonji and Matzkin (2001) and Imbens and Newey (2001) consider identification and estimation of other nonparametric structural models. Recently, Darroles, 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 and estimation of a model of the form

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

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

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

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 $E[\delta(x, z_1)|x] = 0$.

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 $f(x|z) = s(x, z_1)t(z)\exp\{\mu(z)'\tau(x, z_1)\}$,*

$s(x, z_1) > 0$, and $\tau(x, z_1)$ 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 $N(\Gamma(z_1)z_2, \Omega(z_1))$, $\Omega(z_1)$ 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* $\Pr(rank(\Gamma(z_1)) = d_x) = 1$.

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 $\{x_1, ..., x_s\}$ and $\{z_{21}, ..., z_{2t}\}$. Let $P(z_1)$ be the $t \times x$ 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* $\Pr(rank(\Gamma(z_1)) = d_x) = 1$.

The rank condition here implies the order condition that $t > s$. 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.* (1991) and the partially linear models with endogeniety 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),

$$E[\rho(y, x, \theta_0)|z] = 0, \tag{2.4}$$

where $\theta = (\beta', g_1, ..., g_L)$ for a vector of parameters β and functions g_ℓ and $\rho(y, x, \theta)$ is a vector of residuals. Identification of g_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 $((y_i, x_i, z_i), i = 1, ..., n)$. To motivate the estimator it is helpful to work with an estimation analog of equation (2.1). For reduced-form estimators $\hat{\pi}$ and $\hat{F}$ obtained from preliminary nonparametric estimation, consider

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

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 $\hat{F}$ 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 $\hat{\pi}$ and $\hat{F}$. 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 $\{p_1(w), p_2(w), ...\}$ 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 $\hat{E}[p_j(w)|z]$ 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 non-continuity have been proposed, often referred to as "regularization." Generally they consist of careful choice of the 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 of sufficiently smooth functions, and to restrict the estimator $\hat{g}$ to belong to this set. This avoids the continuity problem essentially because integration is a continuous mapping, so by compactness the inverse is continuous (REF). 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 use an approximation for g of a special form. Let $a(w)$ be a vector of functions, and β a vector of parameters, λ denote a $d \times 1$ multi-index, vector of nonnegative integers, $|\lambda| = \sum_{\ell=1}^{d} \lambda_\ell$, and $w^\lambda \equiv \Pi_{i=1}^{d} (w_i)^{\lambda_i}$. Also, let $\bar{w}$ and $\hat{\Sigma}$ estimate the mean and variance of w respectively, and $\tilde{w} = \hat{\Sigma}^{-1/2}(w - \bar{w})$. We approximate $g(w)$ by

$$g(w) \cong a(w)'\beta + \sum_{j=1}^{J} \gamma_j p_j(\tilde{w}), \quad p_j(w) \equiv \exp\{-w'w\} \cdot w^{\lambda(j)}, \tag{3.2}$$

where $|\lambda(j)|$ is increasing in j. This approximation consists of two parts, a "detrending" term $a(w)'\beta$ and a Hermite polynomial approximation. The Hermite polynomial is useful for imposing compactness on the function while allowing for unbounded support of w, as in Gallant and Nychka (1978). The term $a(w)'\beta$ allows for the function g to be unbounded in the tails of the distribution, although its tail behavior is entirely determined by this parametric component.

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

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

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

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

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.[1] A nonparametric two-stage least squares estimator can be obtained by minimizing this objective function subject to restrictions on the parameters.

The restrictions on β and γ are used to impose compactness. To describe these restrictions, let m denote an integer, equal to the number of derivatives of g_0 of interest in estimation. In applications where just the function itself and not its derivatives, are of interest, $m = 0$ is appropriate. Then let $m_0 > d/2$, $\delta_0 > d/2$ denote positive integers, B_g be a known positive constant, and for a function $h(w)$ let $D^\lambda h(w) = (\partial/\partial w_1)^{\lambda_1} ... (\partial/\partial w_d)^{\lambda_d} h(w)$. Let $p^J(w) = (p_1(w), ..., p_J(w))'$ and the $J \times J$ matrix Λ_J be given by

$$\Lambda_J = \sum_{|\lambda| \leq m+m_0} \int D^\lambda [p^J(\tilde{w}) p^J(\tilde{w})'] \cdot (1 + \tilde{w}'\tilde{w})^{\delta_0} dw. \tag{3.5}$$

[1] 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 restrictions we impose are $\gamma'\Lambda_J\gamma \leq B_g$ and $\beta'\beta \leq B_\beta$ for prespecified constants B_g and B_β. The nonparametric two-stage least squares (NP2SLS) estimator is then given by equation (3.2) with (β,γ) replaced by $(\hat{\beta},\hat{\gamma})$ for

$$(\hat{\beta},\hat{\gamma}) = \operatorname{argmin} \hat{Q}(\beta,\gamma) \quad \text{s.t.} \quad \beta'\beta \leq B_{\beta'} \quad \gamma'\Lambda_J\gamma \leq B_g.$$

Computation of this estimator is straightforward, because it is the solution to a quadratic programming problem.[2]

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

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

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

$$\begin{aligned} \hat{E}[a|z_i] \quad &= \quad q_i^{K\prime}\hat{Q}^- \sum_{\ell=1}^n q_\ell^K a(w_\ell)'/n, \quad \hat{E}[p_j|z_i] \\ &= \quad q_i^{K\prime}\hat{Q}^- \sum_{\ell=1}^n q_\ell^K p_j(\hat{\Sigma}^{-1/2}(w_\ell - \bar{w}))'/n \end{aligned} \tag{3.6}$$

where $\hat{Q}^-$ denotes a generalized inverse, with $\hat{Q}\hat{Q}^-\hat{Q} = \hat{Q}$. The NP2SLS estimator can be obtained from equation (3.5) using these as predicted values from the first-stage nonparametric regression.

[2] An alternative approach is available when w is bounded, or can be transformed to be bounded. In that case compactness can also be imposed by specifying bounds on the supremum of certain higher-order derivatives, as was done in Newey and Powell (1988).

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

The NP2SLS estimator depends on the choice of J, B_β, B_g, and K. Together J and B_g determine how flexible the approximation is. For consistency it is important that B_g be set large enough so that the true structural function satisfies a constraint analogous to that for the estimator and that J grow with the sample size. However, if J and B_g are too large then the estimator may be too variable, particular in view of the ill-posed inverse problem. Also, 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 theleading terms for the nonparametric one. The estimator can be formed exactly as described above with w_b replacing w in the calculation of $\tilde{w}$, etc.

The estimator can be extended to estimate the general conditional moment restriction model of equation (2.4), using a minimum distance approach like that of Malinvaud (1980). Let $\hat{\rho}(z_i,\theta) = \left[\sum_{j=1}^{n}\rho(y_j,x_j,\theta)q_j^K/n\right]\hat{Q}^{-}q^K(z_i)$ be a series estimator of $E[\rho(y,x,\theta)|z=z_i]$. Suppose that each function g_ℓ in the fector $\theta=(\beta,g_1,...,g_L)$ depends on a vector w^ℓ of variables. Let $\bar{w}_\ell$ and $\bar{\Sigma}$ be preliminary mean and variance estimators, $\tilde{w}^\ell=\hat{\Sigma}_\ell^{-1/2}(w^\ell-\bar{w}_\ell)$, $g_\ell(\gamma^\ell)=\sum_{j=1}^{J\ell}\gamma_j^\ell p_j(\tilde{w}^\ell)$, and $\Lambda_{J_\ell}^\ell$ be as defined in equation (3.5) with w^ℓ replacing w. Let B_β and $B_{\ell'}$ $(\ell=1,...,L)$ be prespecified bounds analogous to those given above and $\hat{A}$ be a positive definite matrix and consider the objective function

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

The estimator is given by

$$\begin{aligned}\left(\hat{\beta},g(\hat{\gamma})\right) &= \operatorname{argmin}\hat{Q}(\beta,g(\gamma)) \qquad (3.7)\\ \text{s.t.}\quad \beta'\beta &\leq B_\beta,\quad \gamma^{\ell\prime}\Lambda_J^\ell\gamma^\ell\leq B_\ell,\quad (\ell=1,...,L).\end{aligned}$$

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

4 Nonparametric Two Stage Least Squares

We will first give a consistency theorem for an estimator of the general form

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta_J} \sum_{i=1}^{n} \hat{\rho}(z_i, \theta)' \hat{A} \hat{\rho}(z_i, \theta), \tag{4.1}$$

That is applicable to any norm $\|\cdot\|$ on θ and approximating subset Θ_J and then specialize to the specific cases in Section 3. This result may be os some independent interest, e.g. as in Ai and Chen (2000). 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 concerns the first stage and the distance matrix $\hat{A}$:

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

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

Assumption 3: $E[\|\rho(y, x, \theta_0)\|^2 |z]$ is bounded and there exists $M(y, x)$, $\delta > 0$ such that for all $\tilde{\theta}$, $\theta \in \Theta$, $\left\|\rho(y, x, \tilde{\theta}) - \rho(y, x, \theta)\right\| \le M(y, x) \left\|\tilde{\theta} - \theta\right\|^{\delta}$ 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 in the norm $\|\theta\|$.

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

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

The other one is that the approximating subspaces are dense.

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

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 obtain a consistency result for the set up of Section 3. The consistency norm is related to the constraints we have imposed on β and γ. Specifically, for $\theta = (\beta, g_1, ..., g_L)$ and domain $\mathcal{W}_\ell$ of g_ℓ,

$$\begin{aligned} \|\theta\| &= (\beta'\beta)^{1/2} + \sum_{\ell=1}^{L} \|g_\ell\|_\ell \,,\ \|g_\ell\| \qquad (4.2) \\ &= \max\nolimits_{|\lambda|\le m_\ell} \sup\nolimits_{w_\ell \in W_\ell} \mid D^\lambda g_\ell(w_\ell) \mid (1 + w_\ell' w_\ell)^{\delta_\ell}. \end{aligned}$$

Assumption 6: $\Theta = B \times G_1 \times ... \times G_L$, $B = \{\beta : \beta'\beta \le B_\beta\}$, W_ℓ is open and convex, $G_\ell = \{g_\ell(w_\ell)$, continuously differentiable to order $m_\ell + m_{\ell 0}, m_{\ell 0} > d_\ell/2$, $\sum_{|\lambda|\le m+m_0} \int D^\lambda g_{0\ell}(w_\ell)^2 \cdot (1 + w_\ell' w_\ell)^{\delta_0} dw_\ell \le B_\ell,\ (\ell = 1, ..., L)\}$.

The next result shows consistency in the norm given in equation (4.2) for the estimator given in Section 3. For simplicity we assume $\bar{w}_\ell = 0$ and $\hat{\Sigma}_\ell = I$.

Theorem 4.2: *If Assumptions 1-3 and 6 are satisfied and $J_\ell \to \infty, (\ell = 1, ..., J)$ then $(\hat{\beta}, g(\hat{\gamma}))$ from equation (3.7) satisfies $\|\hat{\theta} - \theta_0\| \xrightarrow{p} 0$.*

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

Theorem 4.2: *If the conditional distribution of x given z is complete in z_2, Assumption 2 is satisfied, $g_0(w) = a(w)'\beta_0 + h(w)$, $E[\|a(w)\|] < \infty$, $\beta_0'\beta_0 \le B_\beta$, $\sum_{|\lambda|\le m+m_0} \int [D^\lambda h(w)]^2 (1 + w'w)^{\delta_0} dw \le B_g$, for any $\beta \ne 0$, $\sum_{|\lambda|\le m+m_0} \int [D^\lambda a(w)'\beta]^2 (1 + w'w)^{\delta_0} dw = \infty$, the interior of the support of w_i is convex, and $J \to \infty$ then $\hat{\beta} \xrightarrow{p} \beta_0$ and $\max_{|\lambda|\le m} \sup_{w\in W} |D^\lambda \sum_{j=1}^{J} \hat{\gamma}_j p_j(w) - D^\lambda h(w)| (1 + w'w)^{\delta_0} \xrightarrow{p} 0$.*

This result shows consistency of the NP2SLS estimator, in the sense that the trend component satisfies $\hat{\beta} \xrightarrow{p} \beta_0$ and $\sum_{j=1}^{J} \hat{\gamma}_j p_j(w) \xrightarrow{p} h(w)$, in the norm given in the conclusion.

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

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

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

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

The Hermite series approximation of (3.2) used $J = 5$ terms; while the the parametric component $a(w)'\beta$ was taken to be linear in x. For this design, 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, $B_g = 5$ and $B_g = 50$.

Results for this simple 2×2 design are summarized in Table 1, which gives the root mean-squared-error (RMSE), averaged across the realized values of x and the simulation replications. For each value of the constraint coefficient, these RMSEs 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

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

Table 1: Root Mean Squared Errors for Monte Carlo Design

sensitivity of the RMSEs to the variation in the constraint parameter B_g 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 deciles for the simulation distributions of $\hat{g}$, which are considerably wider for the larger value of B_g.

The figure also plots average values of an estimator $\hat{E}[y|x]$ (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 $\hat{E}[y|x]$ is tilted upward upward relative to the true structural function $g(x)$, and occasionally strays outside the dotted quantile 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 anomolous "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_g 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

Appendix: Proofs

Throughout the appendix C will denote a generic constant that can be different in different uses and abbriviate " with probability approaching one" as w.p.a.1. Also, for notational simplicity let $\sum_i = \sum_{i=1}^n$, $\sum_{ij} = \sum_{i,j=1}^n$, T, CS, and M denote references to the triangle, Cauchy-Schwartz, and Markov inequalities respectively. We first state and prove some Lemmas that are useful for our consistency results. The first is a convergence inprobability version of Gallant (1987) consistency result and the second is a slightly improved version of Corollary 2.2 of Newey (1991).

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

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

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

By iii) there is $\tilde{\theta} \in \hat{\Theta}$ such that $\tilde{\theta} \xrightarrow{p} \theta_0$. By the definition of $\hat{\theta}$, $\hat{Q}(\hat{\theta}) \leq \hat{Q}(\tilde{\theta})$, 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 the triangle inequality, $Q(\tilde{\theta}) < Q(\theta_0) + \Delta$ w.p.a.1. By the defnition of Δ, this even 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) $\hat{Q}(\theta) \xrightarrow{p} Q(\theta)$ for all $\theta \in \Theta$; iii) there is $\delta > 0$ and $B_n = O_p(1)$ such that for all θ, $\tilde{\theta} \in \Theta$, $\left|\hat{Q}(\theta) - \hat{Q}(\tilde{\theta})\right| \leq B_n \left\|\theta - \tilde{\theta}\right\|^\delta$, then $Q(\theta)$ is continuous and* $\sup_{\theta\in\Theta}\left|\hat{Q}(\theta) - Q(\theta)\right| \xrightarrow{p} 0$.

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 $\Delta = (2M/\varepsilon)^{1/\delta}$. Note that for all θ with $\left\|\theta - \tilde{\theta}\right\| \leq \Delta$ we have $\left|\hat{Q}(\theta) - \hat{Q}(\tilde{\theta})\right| \leq B_n\Delta^\delta = B_n\varepsilon/2M \leq \varepsilon/2$ with positive probability. Then by T and ii),

$$\left|Q(\theta) - Q(\tilde{\theta})\right| \leq \left|\hat{Q}(\theta) - Q(\theta)\right| + \left|\hat{Q}(\tilde{\theta}) - Q(\tilde{\theta})\right| + \left|\hat{Q}(\theta) - Q(\tilde{\theta})\right| \leq \varepsilon.$$

Proof. with positive probability. It then follows by $\left|Q(\theta) - Q(\tilde{\theta})\right|$ a constant that $\left|Q(\theta) - Q(\tilde{\theta})\right| \leq \varepsilon$. 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 y given z is as in Theorem 2.2 with

$$\begin{aligned} q(z) &= (2\pi)^{-d/2}\det(\Omega(z_1))^{-1/2}\exp\{(\Psi(z_1)+\Gamma(z_1)z_2)'\Omega(z_1)^{-1}(\Psi(z_1)+\Gamma(z_1)z_2)/2\} \\ r(x, z_1) &= \exp\{x'\Omega(z_1)^{-1}x/2\}, \tau(x, z_1) = -x/2, \mu(z) = \Psi(z_1) + \Gamma(z_1)z_2. \end{aligned}$$

Proof. Note that when $\text{rank}(\Gamma(z_1)) = d_x$ then $\mu(z)$ maps open z_2 sets into open sets for given z_1 so "if" conclusion follows by Theorem 2.2. Also, if $\text{rank}(\Gamma(z_1)) < d_x$ there is $\alpha(z_1)$ such that α

$(z_1)'\Gamma(z_1) = 0$ and $\Pr(\alpha(z_1) = 0) < 1$. Consider $a = \alpha(z_1)'x$. Note that $\text{var}(a|z) = \alpha(z_1)'\Omega(z_1)\alpha(z_1) > 0$ with positive probability, so that $a \neq 0$, but $E[a|z] = \alpha(z_1)'\Gamma(z_1)z_2 = 0$. Thus, 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 2 and 4, there is $\tilde{M}(y, x)$ with

$$\begin{aligned} \|\rho(y,x,\theta\| &\leq \|\rho(y,x,\theta_0)\| + \|\rho(y,x,\theta) - \rho(y,x,\theta_0)\| \leq \|\rho(y,x,\theta_0\| + M(y,x)\left\|\theta - \theta_0\right\|^{\delta} \\ &\leq \|\rho(y,x,\theta_0)\| + M(y,x)C = \tilde{M}(y,x) \end{aligned}$$

Proof. and $E[\tilde{M}(y,x)^2\,|z|$ is bounded. Let $\bar{\rho}(z,\theta) = E[\rho(y,x,\theta) \mid z]$. Then by CS, $E[\|\bar{\rho}(z,\theta)\|^2] \leq E[E[\|\rho(y,x,\theta)\|^2 \mid z]] \leq E[\tilde{M}(y,x)^2] < \infty$. 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). To check iii) of Lemma A1, consider $\theta \in \Theta$, and choose θ_J such that $\|\theta_J - \theta\| \to 0$ as $J \to \infty$. *Q.E.D.* ■

Proof of Theorem 4.2: We note that the constraints $\gamma^\ell \Lambda^\ell_{J_\ell} \gamma^\ell \le B_\ell$ are equivalent to $g_\ell(\gamma^\ell) \in \mathcal{G}_\ell$. Compactness of $\mathcal{G}_\ell$ in the norm $\|\cdot\|_\ell$ 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 satisfying $\pi(z) = E[a(w)'\beta + g(w) \mid z]$ satisfies $a(w)'\beta + g(w) = a(w)'\beta_0 + h(w)$, implying $a(w)'(\beta - \beta_0) = h(w) - g(w)$. Note that by the constraints, $\sum_{|\lambda| \le m+m_0} \int [D^\lambda a(w)'(\beta - \beta_0)]^2 (1 + w'w)^{\delta_0} dw < \infty$, implying $\beta = \beta_0$. For assumption 3, note that

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

Proof. 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

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.

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.

Ferguson, T.S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press.

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): "Indentification and Consistency in Nonparametric Regression," in T.F. Bewley, ed., *Advances in Econometrics: Fifth World Congress*, Cambridge: Cambridge University Press, 145-169.

Gallant, A.R. and D.W. Nychka (1987): "Semi-Nonparamtric Maximum Likelihood Estimation," *Econometrica 55*: 363-390.

Hausman, J.A., H. Ichimura, W.K. Newey, J.L. Powell (1991): "Identification and Estimation of Polynomial Errors-in-Variables Models," *Journal of Econometrics, 50*: 273-295.

Hausman, J.A., H. Ichimura, W.K. Newey, J.L. Powell (1991): "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.

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.

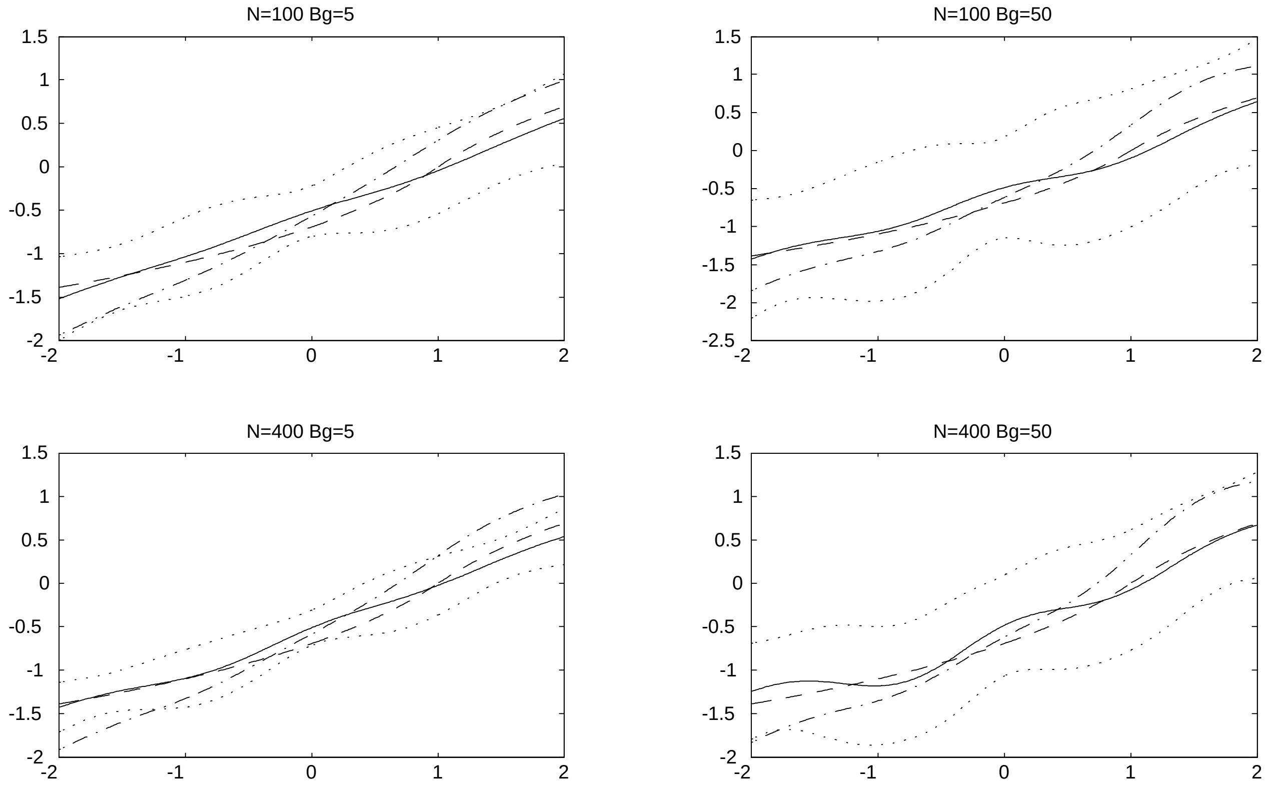

Figure 1: Function Value Estimates for Monte Carlo Design