Available online at www.sciencedirect.com
ScienceDirect

Journal of Econometrics 139 (2007) 4–14

JOURNAL OF Econometrics

www.elsevier.com/locate/jeconom

Instrumental variable estimation of nonseparable models

Victor Chernozhukov[a], Guido W. Imbens[b], Whitney K. Newey[c,*]

[a]*Department of Economics, MIT, Cambridge, MA, USA*
[b]*Department of Economics, Harvard University, MA, USA*
[c]*Department of Economics, MIT, Cambridge, MA 02139, USA*

Available online 29 November 2006

Abstract

There are many environments where knowledge of a structural relationship is required to answer questions of interest. Also, nonseparability of a structural disturbance is a key feature of many models. Here, we consider nonparametric identification and estimation of a model that is monotonic in a nonseparable scalar disturbance, which disturbance is independent of instruments. This model leads to conditional quantile restrictions. We give local identification conditions for the structural equations from those quantile restrictions. We find that a modified completeness condition is sufficient for local identification. We also consider estimation via a nonparametric minimum distance estimator. The estimator minimizes the sum of squares of predicted values from a nonparametric regression of the quantile residual on the instruments. We show consistency of this estimator.

JEL classification: C13; C30

Keywords: Structural models; Nonparametric estimation; Instrumental variables; Quantile estimation

1. Introduction

In econometrics, there are many environments where knowledge of a structural relationship is required to answer questions of interest. Also, nonseparability of a structural disturbance is a key feature of many economic models. Here we consider nonparametric identification and estimation of a model that is monotonic in a

*Corresponding author. Tel.: +1 617 253 6420; fax: +1 617 253 1330.
E-mail address: wnewey@mit.edu (W.K. Newey).

0304-4076/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.jeconom.2006.06.002

nonseparable scalar disturbance, which disturbance is independent of instruments. This model leads to conditional quantile restrictions. We give local identification conditions for the structural equations from those quantile restrictions. We find that a completeness condition is sufficient for local identification. We give sufficient exponential family conditions for local identification and show that with a linear, Gaussian reduced form, the standard rank condition suffices.

The estimator we consider is a nonparametric minimum distance estimator with a parametric approximation to the unknown function. The estimator minimizes the sum of squares of predicted values from a nonparametric regression of the quantile residual on the instruments. The estimator is, in effect, nonlinear two-stage least squares with residuals that are not continuous in the parameters. We show consistency of this estimator when the unknown function and estimator are restricted to have known uniform and Lipschitz bounds, which are weaker conditions than the bounded derivative conditions of Newey and Powell (2003).

The nonparametric model we consider is a special case of the model considered by Chernozhukov and Hansen (2005). Their model allows for a nonscalar disturbance, whereas this paper considers only the scalar disturbance case. They also give innovative global identification conditions, whereas we give conditions for local identification. The local identification conditions developed here extend those in Newey and Powell (2003) to allow for residuals that are nonlinear in an unknown function. The nonparametric minimum distance estimator falls in the class of estimators proposed by Newey and Powell (2003).

The model can also be thought of as extending conditional quantile identification and estimation as considered by Hendricks and Koenker (2002), Matzkin (2003), and others, to allow for endogeneity. We do this by replacing the assumption of independence of disturbance and regressor with independence of the disturbance and an instrument. In relation to Blundell and Powell (2003), Chesher (2003), and Imbens and Newey (2003), the model does not require existence of a reduced form with a single disturbance, but does impose that the structural equation has a single disturbance.

Section 2 of the paper describes the model. Section 3 gives the identification results. Section 4 discusses the minimum distance estimation method and gives a consistency result for this estimator.

2. The model

The model we consider takes the following form:

$$
\begin{aligned}
&Y = g_0(W, U), Z \text{ and } U \text{ are independent,}\\
&\qquad g_0(w, u) \text{ is strictly monotonic increasing in } u,
\end{aligned}
\tag{1}
$$

where u and y are scalars, $w = (x, z_1)$, x is a $d_x \times 1$ vector of endogenous explanatory variables, z_1 and z_2 are $d_1 \times 1$ and $d_2 \times 1$ vectors of instrumental variables, $z = (z_1, z_2)$, and upper and lower case represents a random vector and its realization, respectively. Also $g_0(\cdot)$ denotes the true, unknown structural function of interest and U is continuously distributed with positive density over its support. This model allows the structural disturbance to enter in a fully nonseparable way.

The structural function $g_0(w,u)$ has a quantile interpretation. By monotonicity in U the τth quantile of $g_0(w,U)$ is $g_0(w,q_\tau)$, where q_τ is the τth quantile of the marginal distribution of U. Thus, for different values $\bar{w}$ and $\tilde{w}$ of W,

$$g_0(\tilde{w},q_\tau) - g_0(\bar{w},q_\tau)$$

can be interpreted quantile treatment effect of Lehman (1974), that is the difference of the τth quantiles of Y when W is set to $\bar{w}$ and $\tilde{w}$. Knowledge of these quantiles also lead to knowledge of the average structural function $\mu(w)$ (see Blundell and Powell, 2003) through the well-known relationship,

$$\mu(w) = \int g_0(w,u)F_U(\mathrm{d}u) = \int_0^1 g_0(w,q_\tau)\,\mathrm{d}\tau.$$

The quantile function $g_0(w,q_\tau)$ is interesting because it describes how w affects a particular point in the distribution Y. The average structural function is also interesting, because it summarizes the effect on the whole distribution.

For identification and estimation purposes we focus in this paper on conditional quantile restrictions implied by the model. Independence of U and Z imply that for each τ with $0<\tau<1$, by $g_0(W,u)$ strictly monotonic in u,

$$\begin{aligned}\tau &= E[1(U<q_\tau)] = E[1(U<q_\tau)|Z] = E[E[1(U<q_\tau)|W,Z]|Z]\\ &= E[E[1(g_0(W,U)<g_0(W,q_\tau))|W,Z]|Z] = E[1(Y<g_0(W,q_\tau))|Z].\end{aligned} \tag{2}$$

Letting $\theta^\tau(W) = g(W,q_\tau)$, this equation can be written as a conditional moment restriction

$$\begin{aligned}&E[\rho_\tau(Y,W,\theta_0^\tau)|Z] = 0,\\ &\qquad \rho_\tau(Y,W,\theta) = 1(Y<\theta(W)) - \tau.\end{aligned} \tag{3}$$

Thus we see that the quantile structural effect $g_0(w,q_\tau)$ will satisfy this conditional moment restriction.

The quantile conditions are not the only restrictions implied by the model. The monotonicity of $g_0(w,u)$ implies monotonicity of $\theta_0^\tau(w)$ in τ, i.e.

$$\theta_0^\tau(w) < \theta_0^{\tilde{\tau}}(w) \quad \text{for all } \tilde{\tau} > \tau, \text{ all } w.$$

In general, this monotonicity restriction is not implied by the conditional quantile restrictions except in the exogenous W case where $Z = W$, in which case $\theta_0^\tau(W)$ will be the τth conditional quantile of Y given W, which is known to be monotonic in τ. In general, with endogeneity, monotonicity of $\theta_0^\tau(w)$ in τ will be an extra condition that could be useful for identification and estimation. It would also be interesting to know if there were other restrictions implied by the model that could be used in estimation. We leave this topic to future work and here focus on the conditional quantile restrictions.

3. Identification

For identification purposes we will focus on the nonparametric conditional moment restriction in Eq. (3). This equation is of the general form considered by Newey and Powell (2003) but their identification conditions do not apply, because $\rho(y,w,\theta)$ is nonlinear in θ. This nonlinearity results from θ being inside the indicator function in $\rho(y,w,\theta)$. As is well known, conditions for identification in nonlinear models are somewhat harder to specify than in linear models. It is possible to specify conditions for local identification analogous to those for nonlinear parametric models in Rothenberg (1971), and that is the approach we adopt here. We find that these lead to certain completeness conditions for identification that extend those of Newey and Powell (2003). Chernozhukov and Hansen (2005) give some conditions for global identification, that only require bounded completeness, but may be harder to check because they are nonlinear in θ.

We begin by giving a sufficient condition for local identification from a nonparametric, nonlinear conditional moment restriction. We will then obtain identification conditions for the conditional quantile restriction by applying this result. For the moment, let $\rho(y, w, \theta)$ denote a function that need not have the form given in Eq. (3) above where θ denotes a function of a vector V that may include Y, W, and Z. Let $\|\theta\|_2 = \{E[\theta(V)^2]\}^{1/2}$ denote the mean-square norm.

Theorem 3.1. *If $E[\rho(Y, W, \theta_0)|Z] = 0$ and there exists a function $D(V)$ such that $E[D(V)^2] < \infty$ and* (i) *for every bounded $a(z)$ the functional $E[a(Z)\rho(Y, W, \theta)]$ satisfies*

$$E[a(Z)\rho(Y, W, \theta)] = E[a(Z)D(V)\{\theta(V) - \theta_0(V)\}] + \mathrm{o}(\|\theta - \theta_0\|_2);$$

and (ii) *$E[D(V)\Delta(V)|Z] = 0$ implies $\Delta(V) = 0$; then there is $\varepsilon > 0$ such that the only θ with $\|\theta - \theta_0\|_2 \leqslant \varepsilon$ and $E[\rho(Y, W, \theta)|Z] = 0$ is $\theta = \theta_0$.*

Condition (i) of this result is Frechet differentiablity of $E[a(Z)\rho(Y, W, \theta)]$ in the norm $\|\theta\|_2$, with derivative $a(Z)D(V)$. When $D(V)$ is positive, condition (ii) is completeness for the conditional expectation $E[D(V)(\cdot)|Z]/E[D(V)|Z]$. This is a weighted version of the identification condition in Newey and Powell (2003). Here $D(V)$ is present because the derivative of the residual with respect to θ depends on the data.

This result can be used to derive sufficient conditions for local identification of θ_0 from the moment condition of Eq. (3), as given in the following result:

Theorem 3.2. *If Y is continuously distributed conditional on X and Z with density $f(y|x, z)$ and there is $C > 0$ such that $|f(y|x, z) - f(\tilde{y}|x, z)| \leqslant C|y - \tilde{y}|$ and for $D(V) = f(\theta_0(W)|W, Z)$, $E[D(V)\Delta(V)|Z] = 0$ implies $\Delta(V) = 0$ then $\theta_0(W)$ is locally identified.*

Thus, we see that a sufficient condition for local identification is completeness with respect to the conditional density

$$D(V)/E[D(V)|W] = f(\theta_0(W)|W, Z)f_{W|Z}(W|Z)\Big/\int f(\theta_0(w)|W = w, Z)f_{W|Z}(w, Z)\,\mathrm{d}w.$$

An example helps to fix ideas and relate this result to Chernozhukov and Hansen (2005). Suppose that $X_i \in \{0, 1\}$ and $Z \in \{0, 1\}$ are binary and that Z_1 does not exist. Here Z is the sole instrument and excluded from the structural equation. In this case, the object of interest consists of the pair of numbers $(\theta_0(0), \theta_0(1))$. Here the equation $E[D(V)\Delta(V)|Z] = 0$ is a system of two equations in two unknowns $\Delta(0)$ and $\Delta(1)$, given by

$$\Pi(\Delta(0), \Delta(1))' = 0, \Pi = [\pi_{jk}]_{j,k=0}^{1}, \pi_{jk} = f_{Y|X,Z}(\theta_0(j)|j, k)\Pr(X_i = j|Z_i = k).$$

Then the condition for local identification is that the matrix Π is nonsingular. This is a local version of the global identification condition of Chernozhukov and Hansen (2005) for this example. They show that if Π remains nonsingular when $\theta_0(0)$ and $\theta_0(1)$ are replaced by all values in a certain set then global identification holds. Here this nonsingularity is only required to hold at $\theta_0(0)$ and $\theta_0(1)$.

We can use well-known exponential family results on completeness to obtain more primitive conditions for local identification. This is done in the following result.

Theorem 3.3. *If Eq.* (1) *is satisfied*, $g_0(W,u)$ *is differentiable at* $u=q_\tau$ *with* $\partial g_0(X,q_\tau)/\partial u$ *bounded and bounded away from zero, and with probability one the distribution of* (X,U) *conditional on* Z *is absolutely continuous with density satisfying* $f_{(X,U)|Z}(x,q_\tau|z)=s_\tau(x,z_1)t_\tau(z)\exp\{\mu_\tau(z)'\zeta_\tau(x,z_1)\}$, $s_\tau(x,z_1)>0$, $\zeta(x,z_1)$ *is one-to-one in* x, *and the support of* $\mu(Z)$ *given* Z_1 *is an open set with probability one, then* $g_0(W,q_\tau)$ *is locally identified.*

This result imposes an exponential family hypothesis on the conditional density of (X,U) given Z when U is set equal to it τth quantile. This is a different form of the exponential family condition than is given in Newey and Powell (2003), here requiring that a joint conditional density of X and U have an exponential form rather than the conditional density of X. It is possible to give further sufficient conditions for this hypothesis that are straightforward to interpret. The following is a result for joint conditional normality of X and U.

Theorem 3.4. *If Eq.* (2.1) *is satisfied, with probability one, conditional on* $Z=z$ *the distribution of* (X,U) *is* $\mathrm{N}(\binom{\Psi(z_1)+\Gamma(z_1)z_2}{0},\Omega(z_1))$, $\Omega(z_1)$ *is nonsingular, the last element of the diagonal of* $\Omega(z_1)$ *is constant, and the support of* Z_2 *given* Z_1 *contains an open set, then* $g_0(W,q_\tau)$ *is locally identified if* $\Pr(\mathrm{rank}(\Gamma(Z_1))=d_x)=1$.

The hypotheses of this result are somewhat stronger than the corresponding result in Newey and Powell (2003), involving joint conditional normality of U and X rather than just conditional normality of X. The condition $\mathrm{rank}(\Gamma(Z_1))=d_x$ is identical to that in Newey and Powell (2003), being the nonparametric analog of the necessary and sufficient conditions for identification under conditional normality in a linear model. The matrix $\Gamma(z_1)$ is the analog of the coefficients of the excluded instruments in the reduced form for the right-hand side variables. An order condition for identification in this case is that $d_2\geqslant d_x$, just as in a linear model.

4. Estimation and consistency

The estimators we consider are nonparametric minimum distance estimators based on minimization of the sum of squares of the nonparametric regression of $\rho(Y,W,\theta)$ on Z. Let $\hat{E}[\cdot|Z]$ be a conditional expectation estimator and

$$\hat{R}(\theta)=\sum_{i=1}^{n}\{\hat{E}[\rho(Y,W,\theta)|Z_i]\}^2.$$

For Θ_n some finite dimensional set of functions, to be further specified below, the estimator $\hat{\theta}$ solves

$$\hat{R}(\hat{\theta})\leqslant\inf_{\theta\in\Theta_n}\hat{R}(\theta)+\mathrm{o_p}(1).$$

The form of this estimator as an approximate infimum is necessitated by the discontinuity of the objective function.

To make this estimator operational we have to specify the set Θ_n over which maximization takes place and the conditional expectation estimator $\hat{E}[\cdot|Z]$. We consider series approximations for both. For the structural function consider approximating $\theta_0(w)$ as

$$\theta_0(w) \cong p^J(w)'\gamma^J,$$

where $p^J(w) = (p_{1J}(w), \ldots, p_{JJ}(w))'$ is a sequence of "basis" functions, and γ^J is a corresponding vector of coefficients. For any vector a let $|a| = (a'a)^{1/2}$ denote the Euclidean norm. We assume that for the support $\mathscr{W}$ of w,

Assumption 1. $\mathscr{W}$ is compact and there is a constant C such that

$$\Theta_n = \{p^{\hat{J}_n}(w)'\gamma^{\hat{J}_n} : \sup_{w\in\mathscr{W}} |p(w)'\gamma^{\hat{J}_n}| \leqslant C, \sup_{w,\tilde{w}\in\mathscr{W}} |[p(w) - p(\tilde{w})]'\gamma^{\hat{J}_n}|/|w - \tilde{w}| \leqslant C\}.$$

Also, there exists $p^{\hat{J}_n}(\cdot)'\gamma^{\hat{J}_n} \in \Theta_n$ such that

$$\sup_{w\in\mathscr{W}} |\theta_0(w) - p^{\hat{J}_n}(w)'\bar{\gamma}^{\hat{J}_n}| \xrightarrow{p} 0.$$

Thus, we impose the requirements that linear combinations of the p^J functions can uniformly approximate any function, and certain bounds are placed on linear combinations of the coefficients. The bounds correspond to imposing the restriction that the functions belong to the set

$$\Theta = \{\theta(w) : |\theta(w)| \leqslant C, |\theta(w) - \theta(\tilde{w})| \leqslant C|w - \tilde{w}| \ \forall w, \tilde{w} \in \mathscr{W}\}.$$

By the Arzela theorem, the closure of this set in the norm $\|\theta\| = \sup_{w\in\mathscr{W}}|\theta(w)|$ is compact in $\|\theta\|$; e.g. see Ibragimov and Has'minskii (1981, p. 371, Theorem 17).

We use a series estimator for $\hat{E}[\cdot|Z_i]$. Let $q^K(z) = (q_{1K}(z), \ldots, q_{KK}(z))'$ be a vector of approximating functions for functions of z, such as power series or splines. We will assume that these approximating functions satisfy the following completeness condition:

Assumption 2. For any $m(z)$ such that $E[m(Z)^2] < \infty$, there exists π^K such that

$$\lim_{K\longrightarrow\infty} E[\{m(Z) - q^K(Z)'\pi^K\}^2] = 0.$$

Let $Q = \sum_{i=1}^n q^K(Z_i)q^K(Z_i)'/n$ denote the sample second moment matrix for $q^K(Z_i)$. The predicted value from regressing the residuals $\rho(Y_i, W_i, p^J(\cdot)'\gamma^J)$ on $q^K(Z_i)$ is given by

$$\hat{\pi}_i(\gamma^J) = q^K(Z_i)'Q^- \sum_{j=1}^n q^K(Z_j)\rho(Y_j, W_j, p^J(\cdot)'\gamma^J)/n,$$

where Q^- is any generalized inverse.

With this conditional expectation estimator and the compactness restriction given above, the estimator is given by

$$\hat{\theta}(w) = p^J(w)'\hat{\gamma}^J,$$

$$\sum_{i=1}^n \hat{\pi}_i(\hat{\gamma}^J)^2 \leqslant \inf_{\|\gamma^J\| \leqslant C_J} \sum_{i=1}^n \hat{\pi}_i(\gamma)^2 + o_p(n).$$

Theorem 4.1. *If Assumptions* 1 *and* 2 *are satisfied*, $E[\rho(Y, W, \theta)|Z] = 0$ *if and only if* $\theta = \theta_0$, *and* $f(y|X, Z) \leqslant B(X, Z)$ *with* $E[B(X, Z)] < \infty$, *then* $\sup_{w\in\mathscr{W}}|\hat{g}(w) - g_0(w)| \xrightarrow{p} 0$.

It would also be interesting to have convergence rates and asymptotic normality results. Because identification and estimation is based on conditional moment restrictions we expect that estimation will suffer from an ill-posed inverse problem analogous to that pointed out by Newey and Powell (2003). Consequently, as in Severini and Tripathi (2003), when X and Z have a joint normal distribution and weak restrictions are placed on $\theta_0(W)$ the convergence rate should only be some power of $1/\ln(n)$. As in Darolles et al. (2000) or Hall and Horowitz (2003), the convergence rate may be faster with other distributions or further restrictions on $\theta_0(W)$.

Acknowledgment

This research was supported by the NSF under Grant SES - 0136869, the Alfred P. Sloan Foundation, and the Center for Advanced Study in the Behavioral Sciences. This paper has been presented at seminars at UCL and Harvard-MIT.

Appendix. Proofs

Throughout this Appendix, C will denote a generic constant that may be different in different uses.

Proof of Theorem 3.1. Suppose that the conclusion does not hold. Then there exists a sequence $\theta_k, (k = 1, 2, \ldots)$ such that $\|\theta_k - \theta_0\|_2 \to 0$, $E[\rho(Y, W, \theta_k)|Z] = 0$, and

$$\begin{aligned} 0 &= E[a(Z)\rho(Y, W, \theta_k)] = E[a(Z)\rho(Y, W, \theta_0)] + E[a(Z)D(V)\{\theta_k(V) - \theta_0(V)\}] + \mathrm{o}(\|\theta_k - \theta_0\|_2) \\ &= E[a(Z)D(V)\{\theta_k(V) - \theta_0(V)\}] + \mathrm{o}(\|\theta_k - \theta_0\|_2). \end{aligned}$$

Dividing through by $\|\theta_k - \theta_0\|$, it follows that for $\Delta_k(V) = \{\theta_k(V) - \theta_0(V)\}/\|\theta_k - \theta_0\|_2$,

$$E[a(Z)D(V)\Delta_k(V)] \to 0.$$

Also, by compactness of the unit ball in the weak topology in the Hilbert space $\{\theta(V) : \|\theta\|_2 < \infty\}$, there exists $\Delta(V)$ such that for any $b(V)$ with $\|b\|_2 < \infty$, $E[b(V)\Delta_k(V)] \to E[b(V)\Delta(V)]$. Then, for $b(V) = E[a(Z)|V]D(V)$, it follows that

$$0 = E[b(V)\Delta(V)] = E[a(Z)D(V)\Delta(V)] = E[a(Z)E[D(V)\Delta(V)|Z]].$$

Since this equation holds for all bounded $a(Z)$ it follows that $E[D(V)\Delta(V)|Z] = 0$, contradicting condition (ii). □

Proof of Theorem 3.2. Note that for θ_k with $\|\theta_k - \theta\| \longrightarrow 0$ and $V = (Y, W, Z)$

$$\begin{aligned} &E[a(Z)\rho(Y, W, \theta_k)] - E[a(Z)\rho(Y, W, \theta_0)] \\ &\quad = E\left[a(Z)\left\{\int_{-\infty}^{\theta_k(W)} f(y|W, Z)\,\mathrm{d}y - \int_{-\infty}^{\theta_0(W)} f(y|W, Z)\,\mathrm{d}y\right\}\right] \\ &\quad = E[a(Z)f(\bar{\theta}_k(W)|W, Z)\{\theta_k(W) - \theta_0(W)\}] = E[a(Z)D(V)\{\theta_k(W) - \theta_0(W)\}] + R_k, \end{aligned}$$

$$R_k = E[a(Z)\{f(\bar{\theta}_k(W)|W, Z) - f(\theta_0(W)|W, Z)\}\{g_k(W) - g_0(W)\}].$$

By the Lipschitz condition on f, we have

$$\begin{aligned} |R_k| &\leqslant CE[|\bar{\theta}_k(W) - \theta_0(W)||\theta_k(W) - \theta_0(W)|] \\ &\leqslant C\|\theta_k - \theta_0\|^2 = \mathrm{o}(\|\theta_k - \theta_0\|). \end{aligned}$$

It follows that assumption (i) of Theorem 3.1 are satisfied with $D(V) = f(\theta_0(W)|W, Z)$, so the conclusion follows by the conclusion of Theorem 3.1. □

Proof of Theorem 3.3. Let $f(u|x,z)=f_{U|X,Z}(u|x,z)$ denote the conditional density of U given X and Z. Then $f_{Y|X,Z}(y|x,z) = |\partial g_0(x, g_0^{-1}(x,y))/\partial u|^{-1} f(g_0^{-1}(x,y)|x,z)$. Therefore,

$$f_{Y|X,Z}(g_0(x,q_\tau)|x,z) = |\partial g_0(x,q_\tau)/\partial u|^{-1} f(q_\tau|x,z) = D(x,z),$$

for $D(x,z)$ in Theorem 3.2. Note that $f(q_\tau|x,z)f(x|z)=f_{(X,U)|Z}(x,q_\tau|z)$, and that for $\tilde{s}(x,z_1) = |\partial g_0(x,q_\tau)/\partial u|^{-1}s(x,z_1)$ and $\tilde{t}(z) = t(z)/E[D(x,z)|z]$, it follows from the hypotheses that

$$\tilde{f}(x|z) = D(x,z)f(x|z)/E[D(x,z)|z] = \tilde{s}(x,z_1)\tilde{t}(z)\exp\{\mu(z)'\zeta(x,z_1)\}$$

is a conditional density satisfying the exponential family hypotheses of Newey and Powell (2003). Let $\tilde{E}[\cdot|Z]$ denote the conditional expectation for this conditional density. Note that for $V=(X,Z)$

$$E[D(V)\Delta(V)|Z] = \tilde{E}[\Delta(V)|Z]E[D(V)|Z],$$

so by $E[D(V)|Z]>0$ with probability one, $E[D(V)\Delta(V)|Z]=0$ if and only if $\tilde{E}[\Delta(V)|Z]=0$. Then local identification follows by completeness of $\tilde{f}(x|z)$, which is shown in the proof of Theorem 2.2 of Newey and Powell (2003). □

Proof of Theorem 3.4. For simplicity we suppress the z_1 argument, let $z=z_2$, and let $x = x - \Psi$. Also partition $\Omega^{-1} = [\Omega^{ij}]_{ij=1,2}$ conformably with x,u. Then

$$f_{(X,U)|Z}(x,q_\tau|z) = C\ \det(\Omega)^{-1/2}\exp(-(x-\Gamma z, q_\tau)\Omega^{-1}(x-\Gamma z, q_\tau)'/2),$$

which has the form given in Theorem 3.3 with

$$t(z) = C\ \det(\Omega)^{-1/2}\exp\{-[(\Gamma z)'\Omega^{11}\Gamma z + q_\tau^2\Omega^{22} - 2q_\tau\Omega^{21}\Gamma z]/2\}$$
$$s(x) = \exp\{-[x'\Omega^{11}x + 2q_\tau\Omega^{21}x]/2\}, \tau(x) = \Omega^{11}x, \mu(z) = \Gamma z.$$

Not that when $\operatorname{rank}(\Gamma)=d_x$, then $\mu(z)$ maps open z sets into open sets, so that the conclusion follows by Theorem 3.3. □

The following results are used in the consistency proofs.

Lemma A1. *Suppose* (i) $L(\theta)$ *has a unique minimum on* Θ *at* θ_0; (ii) $L(\theta)$ *is continuous*, Θ *is compact, and* $\sup_{\theta\in\Theta}|\hat{L}(\theta)-L(\theta)| \xrightarrow{p} 0$; (iii) $\hat{\Theta}$ *are subsets of* Θ *and there exists* $\tilde{\theta}\in\hat{\Theta}$ *such that* $\tilde{\theta} \xrightarrow{p} \theta$. *Then if* $\hat{L}(\hat{\theta}) \leqslant \inf_{\theta\in\hat{\Theta}}\hat{L}(\theta) + o_p(1)$, *it follows that* $\hat{\theta} \xrightarrow{p} \theta_0$.

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

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

For $\tilde{\theta}$ satisfying condition (iii), $\hat{L}(\hat{\theta}) \leqslant \hat{L}(\tilde{\theta}) + \Delta/3$ with probability approaching one (w.p.a.1). Then by the uniform convergence hypothesis in condition (ii), $L(\hat{\theta}) < L(\tilde{\theta}) + 2\Delta/3$ w.p.a.1. By $\tilde{\theta} \xrightarrow{p} \theta$ and continuity of $L(\theta)$, $L(\tilde{\theta}) < L(\theta_0) + \Delta/3$ w.p.a.1. Then summing up and subtracting $L(\tilde{\theta})$ from both sides gives $L(\tilde{\theta}) < L(\theta_0) + \Delta$ w.p.a.1. By the definition 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. □

Let $\rho_i(\theta) = \rho(Y_i, W_i, \theta)$, $\Delta_i(\tilde{\theta},\theta) = |\rho_i(\tilde{\theta}) - \rho_i(\theta)|$, $\|\theta\| = \sup_{w\in\mathcal{W}}|\theta(w)|$, and $d_i(\theta,\delta) = \sup_{\tilde{\theta}:\|\tilde{\theta}-\theta\|<\delta}\Delta_i(\tilde{\theta},\theta)$.

Lemma A2. *If the hypotheses of Theorem* 4.1 *are satisfied then there is a constant C such that for all $\theta \in \Theta$ and $\delta > 0$,*

$$E[d_i(\theta, \delta)] \leqslant C\delta.$$

Proof. For any $\tilde{\theta}$ with $\|\tilde{\theta} - \theta\| < \delta$ it follows that

$$\Delta_i(\tilde{\theta}, \theta) \leqslant 1(\theta(W_i) - \delta \leqslant Y_i \leqslant \theta(W_i) + \delta).$$

Therefore,

$$E[d_i(\theta, \delta)] \leqslant E[1(\theta(W_i) - \delta \leqslant Y_i \leqslant \theta(W_i) + \delta)]$$
$$= E\left[\int_{\theta(W_i)-\delta}^{\theta(W_i)+\delta} f(y|X_i, Z_i)dy\right] \leqslant 2\delta E[B(X_i, Z_i)]. \qquad \square$$

Lemma A3. *If the hypotheses of Theorem* 4.1 *are satisfied then*

$$\sup_{(\tilde{\theta},\theta)\in\Theta\times\Theta}\left|\sum_{i=1}^{n} \Delta_i(\tilde{\theta}, \theta)/n - E[\Delta_i(\tilde{\theta}, \theta)]\right| \xrightarrow{p} 0.$$

Proof. Let

$$D_i(\tilde{\theta}, \theta, \delta) = \sup_{\|\tilde{\gamma}-\theta\| \leqslant \delta, \|\gamma-\theta\| \leqslant \delta} |\Delta_i(\tilde{\gamma}, \gamma) - \Delta_i(\tilde{\theta}, \theta)|.$$

By the triangle inequality and Lemma A2 there is a constant C such that for all $(\tilde{\theta}, \theta) \in \Theta \times \Theta$,

$$E[D_i(\tilde{\theta}, \theta, \delta)] \leqslant E[d_i(\tilde{\theta}, \delta) + d_i(\theta, \delta)] \leqslant C\delta.$$

Consider any $\varepsilon > 0$. Choose δ such that for $C\delta$ from Lemma A2, $C\delta < \varepsilon/3$. Let $\mathcal{N}_{\tilde{\theta},\theta} = \{(\tilde{\gamma}, \gamma) : \|\tilde{\gamma} - \tilde{\theta}\| < \delta, \|\gamma - \theta\| < \delta\}$. Then $\vee_{(\tilde{\theta},\theta)\in\Theta\times\Theta}\mathcal{N}_{\tilde{\theta},\theta}$ is an open cover of $\Theta \times \Theta$, so by compactness of $\Theta \times \Theta$ in the product topology, there exists a finite sub cover $\vee_{j=1}^{J}\mathcal{N}_{\tilde{\theta}_j,\theta_j}$. Then,

$$\sup_{(\tilde{\theta},\theta)\in\Theta\times\Theta}\left|\sum_{i=1} \Delta_i(\tilde{\theta}, \theta)/n - E[\Delta_i(\tilde{\theta}, \theta)]\right|$$
$$\leqslant \max_{j=1,\dots,J}\left|\sum_{i=1}^{n} \Delta_i(\tilde{\theta}_j, \theta_j)/n - E[\Delta_i(\tilde{\theta}_j, \theta_j)]\right| + \max_{j=1,\dots,J}\left\{\sum_{i=1}^{n} D_i(\tilde{\theta}_j, \theta_j, \delta)/n + E[D_i(\tilde{\theta}_j, \theta_j, \delta)]\right\}$$
$$\leqslant o_p(1) + \max_{j=1,\dots,J}\left|\sum_{i=1}^{n} D_i(\tilde{\theta}_j, \theta_j, \delta)/n - E[D_i(\tilde{\theta}_j, \theta_j, \delta)]\right| + 2C\delta \leqslant o_p(1) + 2\varepsilon/3,$$

where the second inequality follows by the law of large numbers and the triangle inequality. Therefore, with probability approaching one,

$$\sup_{(\tilde{\theta},\theta)\in\Theta\times\Theta}\left|\sum_{i=1} \Delta_i(\tilde{\theta}, \theta)/n - E[\Delta_i(\tilde{\theta}, \theta)]\right| < \varepsilon. \qquad \square$$

Proof of Theorem 4.1. Define $\hat{\pi}_i(\theta) = q^K(Z_i)'Q^{-}\sum_{j=1}^{n}q^K(Z_j)\rho_j(\theta)/n$, $\pi_i(\theta) = E[\rho_i(\theta)|Z_i]$, and $\hat{L}(\theta) = \sum_{i=1}^{n}\hat{\pi}_i(\theta)^2/n$. It follows by Newey (1991, Proof of Corollary 4.2) that for each θ,

$$\hat{L}(\theta) \xrightarrow{p} E[\pi_i(\theta)^2].$$

Now, note that since $\Delta_i(\tilde{\theta}, \theta)$ can take on only the values 0 and 1, so that $\Delta_i(\tilde{\theta}, \theta) = |\rho_i(\tilde{\theta}) - \rho_i(\theta)|^2$. Then for any $\theta, \tilde{\theta} \in \Theta$, it follows by $\hat{\pi}_i(\theta)$ being a least squares projection that

$$\sum_{i=1}^{n} |\hat{\pi}_i(\tilde{\theta}) - \hat{\pi}_i(\theta)|^2 \leqslant \sum_{i=1}^{n} \Delta_i(\tilde{\theta}, \theta), \sum_{i=1}^{n} \hat{\pi}_i(\theta)^2 \leqslant \sum_{i=1}^{n} \rho_i(\theta)^2 \leqslant Cn.$$

It follows that

$$\begin{aligned} n|\hat{L}(\tilde{\theta}) - \hat{L}(\theta)| &\leqslant \sum_{i=1}^{n} |\hat{\pi}_i(\tilde{\theta}) - \hat{\pi}_i(\theta)|^2 + 2\sum_{i=1}^{n} |\hat{\pi}_i(\theta)||\hat{\pi}_i(\tilde{\theta}) - \hat{\pi}_i(\theta)|^2/n \\ &\leqslant \sum_{i=1}^{n} \Delta_i(\tilde{\theta}, \theta) + 2(Cn)^{1/2} \left(\sum_{i=1}^{n} \Delta_i(\tilde{\theta}, \theta) \right)^{1/2}. \end{aligned}$$

By the triangle inequality and Lemma A3, for any $\delta_n \longrightarrow 0$,

$$\begin{aligned} \sup_{\|\tilde{\theta}-\theta\| \leqslant \delta_n} \sum_{i=1}^{n} \Delta_i(\tilde{\theta}, \theta)/n &\leqslant \sup_{\tilde{\theta}, \theta \in \Theta} \left| \sum_{i=1}^{n} \Delta_i(\tilde{\theta}, \theta)/n - E[\Delta_i(\tilde{\theta}, \theta)] \right| + \sup_{\|\tilde{\theta}-\theta\| \leqslant \delta_n} E[\Delta_i(\tilde{\theta}, \theta)] \\ &\leqslant o_p(1) + \sup_{\theta} E[d_i(\theta, \delta_n)] \leqslant o_p(1) + C\delta_n \xrightarrow{p} 0. \end{aligned}$$

Therefore, it follows that $\sup_{\|\tilde{\theta}-\theta\| \leqslant \delta_n} |\hat{L}(\tilde{\theta}) - \hat{L}(\theta)| \xrightarrow{p} 0$. It then follows that Assumption 3 of Newey (1991) is satisfied, with $\Delta_n(\varepsilon, \eta) = \sup_{\|\tilde{\theta}-\theta\| \leqslant \delta} |\hat{L}(\tilde{\theta}) - \hat{L}(\theta)|$ for δ small enough, so that $\sup_{\theta \in \Theta} |\hat{L}(\theta) - L(\theta)| \xrightarrow{p} 0$ by Theorem 2.1 of Newey (1991). The conclusion now follows by Lemma A1. □

References

Blundell, R., Powell, J.L., 2003. Endogeneity in semiparametric and nonparametric regression models. in: Dewatripont, M., Hansen, L.P., Turnovsky, S.J. (Eds.), Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, vol. II, Cambridge University Press, Cambridge, pp. 312–357.

Chernozhukov, V., Hansen, C., 2005. An IV model of quantile treatment effects. Econometrica 73, 245–261.

Chesher, A., 2003. Identification in nonseparable models. Econometrica 71, 1405–1441.

Darolles, S., Florens, J.-P., Renault, E., 2000. Nonparametric instrumental regression. Manuscript, GREMAQ, University of Toulouse l.

Hall, P., Horowitz, J., 2003. Nonparametric methods for inference in the presence of instrumental variables. Working Paper, Northwestern University.

Hendricks, W., Koenker, R., 2002. Hierarchical spline models for conditional quantiles and the demand for electricity. Journal of the American Statistical Association 87, 58–69.

Ibragimov, I.A., Has'minskii, R.Z., 1981. Statistical Estimation: Asymptotic Theory. Springer, New York.

Imbens, G.W., Newey, W.K., 2003. Identification and estimation of triangular simultaneous equations models without additivity. Working paper, MIT.

Lehman, E.L., 1974. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

Matzkin, R.L., 2003. Nonparametric estimation of nonadditive random functions. Econometrica 71, 1339–1375.

Newey, W.K., 1991. Uniform convergence in probability and stochastic equicontinuity. Econometrica 59, 1161–1167.

Newey, W.K., Powell, J.L., 2003. Instrumental variables estimation for nonparametric models. Econometrica 71, 1565–1578.
Rothenberg, T.J., 1971. Identification in parametric models. Econometrica 39, 577–591.
Severini, T., Tripathi, G., 2003. Some identification issues in nonparametric linear models with endogenous regressors. Working Paper, Department of Economics, University of Wisconsin.

Further Reading

Lehman, E.L., 1959. Testing Statistical Hypotheses. Wiley, New York.
Newey, W.K., Powell, J.L., Vella, F., 1999. Nonparametric estimation of triangular simultaneous equations models. Econometrica 67, 565–603.