

Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity*

Guido W. Imbens

Department of Economics
Agricultural and Resource Economics
UC Berkeley, and NBER

Whitney K. Newey

Department of Economics
M.I.T.

First Draft: March 2001
This Draft: February 2002

Abstract

This paper investigates identification and inference in a nonparametric structural model with instrumental variables without additivity. We use independence and monotonicity conditions for identification of the average structural function, as well as for identification of the structural response function. A two step series estimator is used, the first step estimating the conditional distribution of the endogenous regressor given the instrument. The second estimation step uses the estimated conditional distribution function as a regressor in a control function approach. Convergence rates are given.

JEL Classification:

Keywords: *Simultaneous equations models, Instrumental Variables, Additivity, Non-linear Models, Nonparametric Estimation*

*This research was partially completed while the second author was a fellow at the Center for Advanced Study in the Behavioral Sciences. The NSF provided partial financial support through grant SES 0136789 (Imbens). We are grateful for comments by Susan Athey, Lanier Benkard, Jim Heckman, and participants at seminars at Stanford University (March 2001) and University College London (May 2001).

1 Introduction

Simultaneous equations models are of great interest in econometrics, and identification and inference in such models in settings with instrumental variables has received considerable attention. Recently interest has focused on identification under weak assumptions without functional form or distributional restrictions (e.g., Roehrig 1988; Newey and Powell, 1988; Newey, Powell and Vella, 1999; Darolles, Florens and Renault, 2000; Pinkse, 2000b; Blundell and Powell, 2000; Heckman, 1990; Imbens and Angrist, 1994; Vytlačil, 2001; Das, 2000).

Most of the work on instrumental variables has maintained additive separability of the disturbances and the regression functions. In this paper we make two contributions. First, we present two identification results that do not require additive separability of the disturbances in either the first stage regression or the main outcome equation. Instead we consider three key assumptions beyond the independence between instrument and disturbances that underlies much of the work in this area: *(i)* the relation between the endogenous regressor and the instrument is monotone in the unobserved component; *(ii)* the relation between the outcome of interest and the endogenous regressor is monotone in the unobserved component; *(iii)* the instrument has sufficient power to move the endogenous regressor over its entire support. The first identification results states that given the first and the third of these assumptions (and the instrumental variables assumption), the average structural function, introduced by Blundell and Powell (2000) is identified. The second identification results states that under the first two assumptions (plus the instrumental variables assumption) both the entire relation between the endogenous regressor and the outcome of interest and the joint distribution of the disturbance and the endogenous regressor are identified on the joint support of the disturbance and the endogenous regressor. Together these allow us to estimate the effect of many policies of interest.

Our second contribution is the development of a framework for inference in these models. Using a series approach we develop a consistent estimator for the unknown func-

tions. The series estimator consists of two steps. The first step estimates the conditional distribution function of the endogenous regressor given the instrument. Evaluating this at the observed values gives an estimated regressor in a control function approach (e.g., Heckman and Robb, 1984; Newey, Powell and Vella, 1999). The second step regresses the outcome on the endogenous regressor and the estimated residual from the first step. This gives us the behavioral relation of interest. We can then also estimate the average structural function through averaging over the marginal distribution of the first stage residual. We give convergence rates for both the first and second step of the estimation procedure and for the averaging in the average structural function.

2 The Model

We consider the following triangular simultaneous equations model involving two equations:

$$Y = g(X, \eta, \nu), \tag{2.1}$$

$$X = h(Z, \eta). \tag{2.2}$$

The interest is in the effect of X on Y , as well as in the joint distribution of X and (η, ν) . The regressor X is endogenous, because η , the disturbance or unobserved component in the first equation and therefore correlated with X , enters in the second equation. The instrument Z will be assumed to be independent of the pair of disturbances (η, ν) .

A special case of this model arises when we replace the first equation with:

$$Y = g(X, \varepsilon), \tag{2.3}$$

with unrestricted dependence between ε and η . To see that this model is more restrictive, consider the model described by the equations (2.2) and (2.3). Define in this case $\nu = r(\varepsilon, \eta)$, where $r(\varepsilon, \eta) = F_{\varepsilon|\eta}(\varepsilon|\eta)$, with inverse $\varepsilon = r^{-1}(\nu, \eta)$. Then we can write

$$Y = g(X, \varepsilon) = g(X, r^{-1}(\nu, \eta)) = \tilde{g}(X, \eta, \nu).$$

The second model characterized by (2.2)-(2.3) is more restrictive than (2.1)-(2.2) because it restricts the way in which the two disturbances interact with X in the function that

determines Y . It corresponds closely, however to many models considered in economics, and in the examples we focus on this special case. The identification strategy in the general model is more transparent, however, and for the formal results we therefore focus on the general model. The following example provide some motivation for such models without additivity, and will be used to discuss some of the assumptions we consider later.

Example: (PRODUCTION FUNCTION)

Consider a production function, as a function of an input x and an unobserved component ε : $y = g(x, \varepsilon)$. Unlike the additive case with $g(x, \varepsilon) = g(x) + \varepsilon$, the marginal return to the input x , $\partial g / \partial x(x, \varepsilon)$, is allowed to differ by firm. The level of the input x is chosen by the firm, whereas ε is an input which is not under the control of the firm, and not observed by either the firm or the econometrician. The firm chooses the value of x by maximizing expected profits based on a noisy signal of ε , denoted by η . Profits are the difference between production times price (normalized to equal one), and costs, which depend on the value of the input and an observed cost shifter z :

$$\pi(x, z, \varepsilon) = g(x, \varepsilon) - c(x, z),$$

so that

$$X = \operatorname{argmax}_x E[\pi(x, Z, \varepsilon) | \eta, Z] = \operatorname{argmax}_x \left[E[g(x, \varepsilon) | \eta] - c(x, Z) \right].$$

Thus, $X = h(Z, \eta)$, so that equations (2.1) and (2.2) are satisfied. To illustrate the importance of relaxing the additivity of the production function in ε in this model, note that if $g(x, \varepsilon)$ were additive in ε , the optimal level of the input would be the solution to $\max_x g(x) - c(x, Z)$. In that case the optimal solution $X = h(Z, \eta)$ would not depend on η and all firms with the same level of the instrument Z would choose the same level of the input.

To motivate our interest in both the production function $g(x, \varepsilon)$ and $F_{X, \varepsilon}(x, \varepsilon)$, the joint distribution of the input and disturbance (X, ε) , consider the effect on average output of various interventions or policies that may be contemplated by policymakers.

Similar to the binary endogenous regressor case there is a variety of parameters of interest (e.g., Heckman and Vytlacil, 2000; Pearl, 2000). Here we discuss three examples of policies of interest and how their effect depends on $g(x, \varepsilon)$ and $F_{X,\varepsilon}(x, \varepsilon)$.

(I) Blundell and Powell (2000) focus on the identification and estimation of what they label the *average structural function* (ASF), the average of the structural function $g(x, \varepsilon)$ over the marginal distribution of ε . A policy maker may consider fixing the input at a particular level x , either at $x = x_0$ or $x = x_1$. Evaluating the average outcome at this level requires knowledge of the function

$$\mu(x) = E[g(x, \varepsilon)] = \int g(x, \varepsilon) dF_\varepsilon(\varepsilon), \quad (2.4)$$

at $x = x_0$ and $x = x_1$. Note that $\mu(x)$ is not equal to the conditional expectation of Y given $X = x$,

$$E[Y|X = x] = \int g(x, \varepsilon) dF_{\varepsilon|X}(\varepsilon|x),$$

because of the dependence between X and ε . If the production function is linear and additive, that is, $g(x, \varepsilon) = \beta_0 + \beta_1 \cdot x + \varepsilon$, the average structural function is $\beta_0 + \beta_1 \cdot x$, and so the effect of fixing the input at x_1 versus x_0 is $\beta_1 \cdot (x_1 - x_0)$. This slope coefficient β_1 is typically taken as the parameter of interest in linear simultaneous equations models.

(II) A second policy of interest is increasing for all units the value of the input by a small amount. The per-unit effect of such a policy on average output is

$$E \left[\frac{\partial g}{\partial x}(X, \varepsilon) \right] = \int \int \frac{\partial g}{\partial x}(x, \varepsilon) dF_{X,\varepsilon}(x, \varepsilon). \quad (2.5)$$

This average derivative parameter is similar to those studied in Stoker (1986) and Powell, Stock and Stoker (1989) in the context of exogenous regressors. Again, if the production function is linear and additive, that is, $g(x, \varepsilon) = \beta_0 + \beta_1 \cdot x + \varepsilon$, this effect can be expressed in terms of the coefficients of the linear model. The marginal effect of a unit increase in x would be β_1 , the coefficient on the input. Note that this average derivative cannot be inferred from the ASF $\mu(x)$. In particular, it is in general not equal to the expected value of the derivative of the ASF,

$$E \left[\frac{\partial \mu}{\partial x}(X) \right] = \int \frac{\partial \mu}{\partial x}(x) dF_X(x) = \int \int \frac{\partial g}{\partial x}(x) dF_\varepsilon(\varepsilon) dF_X(x),$$

unless X and ε are independent.

(III) A third policy of interest is imposing a minimum on the value of the input at \underline{x} . The average outcome under such a policy would be

$$E[g(\max(X, \underline{x}), \varepsilon)] = \int \int g(\max(x, \underline{x}), \varepsilon) dF_{x, \varepsilon}(x, \varepsilon). \quad (2.6)$$

An example of such a policy would be an increase in the compulsory schooling age, with the interest in the effect of such a policy on average earnings. Note that even in the context of standard additive linear simultaneous equations models knowledge of the regression coefficients would not be sufficient for the evaluation of such a policy—this would also require knowledge of the joint distribution of (X, ε) unless X is exogenous.

3 Identification

We consider the following four assumptions. First, the instrument is assumed to be independent of the disturbances.

Assumption 3.1 (INDEPENDENCE)

The disturbances (ν, η) are jointly independent of Z and of each other.

Note that as in, for example, Roehrig (1988) and Imbens and Angrist (1994), full independence is assumed, rather than the weaker mean-independence as in, for example, Newey, Powell and Vella (1999) and Darolles, Florens and Renault (2001). This is part of the price paid for relaxing the additivity assumption.

The second assumption requires the structural relation between the endogenous regressor and the instrument to be monotone in the unobserved disturbance.

Assumption 3.2 (MONOTONICITY OF ENDOGENOUS REGRESSOR IN THE UNOBSERVED COMPONENT)

The function $h(z, \eta)$ is strictly monotone in its second argument.

This assumption is trivially satisfied if this relation is additive in instrument and disturbance, but clearly allows for general forms of non-additive relations. Matzkin (1999)

considers nonparametric estimation of $h(z, \eta)$ under this assumption and independence of η and Z in a single equation framework. Pinkse (2000b) refers to a multivariate version of this as “weak separability”. Das (2001) considers a stochastic version of this assumption to identify parameters in single index models with a single endogenous regressor.

It is interesting to compare this assumption to the monotonicity assumption used in Imbens and Angrist (1994) and Vytlacil (2001). In terms of the current notation, Imbens-Angrist and Vytlacil focus on monotonicity of $h(z, \eta)$ in the observed component, the instrument z rather than in the unobserved component, the disturbance ε . With a binary x and binary instrument z weak monotonicity in z and weak monotonicity in η are equivalent. However, in the multivalued regressor case, e.g., Angrist and Imbens (1995), the two assumptions are distinct, with neither one implying the other.

Assumption 3.2 has only weak testable implications. A slightly weaker form, requiring $h(z, \varepsilon)$ to be monotone in η rather than strictly monotone has no testable implications at all. The testable implications for strict monotonicity version arise when Z and X are discrete. With both Z and X continuous, there are no testable implications.

Das (2001) discusses a number of examples where monotonicity of the decision rule is implied by conditions on the economic primitives using monotone comparative statics results (e.g., Milgrom and Shannon, 1994). In the same vein, consider the production function example introduced in Section 2, and assume that the production and cost function are twice continuously differentiable. Suppose that the production function is strictly increasing in the unobserved input ε , and that the return to the observed input is also increasing in the unobserved input, so that $\partial g / \partial \varepsilon > 0$ and $\partial^2 g / \partial x \partial \varepsilon > 0$. If in addition the signal η and the unobserved input ε are affiliated, the decision rule $h(z, \eta)$ is monotone in the signal.

The third assumption requires monotonicity of the production function in the second unobserved component.

Assumption 3.3 (MONOTONICITY OF THE OUTCOME IN THE UNOBSERVED COMPONENT)

The function $g(x, \eta, \nu)$ is strictly monotone in its third argument.

Finally, the fourth assumption requires the conditional support of X given η to be independent of the value of η .

Assumption 3.4 (SUPPORT)

The support of X given η does not depend on the value of η .

Assumption 3.4 is very strong. Given the deterministic relation between Z and X given η , this implies that by changing the value of the instrument, one can induce any value of the endogenous regressor. In the binary endogenous variable case this implies that by changing the value of Z , one can induce both values for the endogenous regressor, similar to the “identification-at-infinity” results in Chamberlain () and Heckman (1990). In the binary case that would immediately imply identification of the average outcome at both values of the endogenous regressor. Here this assumption in itself is not sufficient to identify the average structural function at all values of the regressor.

First we show a relation between the assumptions considered here and some easily interpretable assumption in the special case of model (2.2)-(2.3) where $Y = g(X, \varepsilon)$.

Lemma 1: (IMPLICATION OF MODEL (2.2)-(2.3))

If Model (2.2)-(2.3) holds, and $h(z, \eta)$ is strictly monotone in η , $g(x, \varepsilon)$ is strictly monotone in ε , and (η, ε) are jointly independent of z , then Model (2.1)-(2.2) and Assumptions 3.1-3.3 hold for $\nu = F_{\varepsilon|\eta}(\varepsilon|\eta)$.

The first key result is an extension of the results in Blundell and Powell (2000), allowing for a more flexible relation between the endogenous regressor and the instrument. Blundell and Powell (2000) do allow for a function $g(\cdot)$, but assume that $h(\cdot)$ is additive and linear.

Theorem 1: (IDENTIFICATION OF THE AVERAGE STRUCTURAL FUNCTION)

If Assumptions 3.1, 3.2 and 3.4 are satisfied, then $\mu(x)$ is identified from the joint distribution of (Y, X, Z) .

Proofs are given in the Appendix. This result shows that $\mu(x)$ is identified by first calculating $\eta = F_{X|Z}(X, Z)$, then regressing Y on X and η , and then averaging over the

marginal distribution of η . This approach is essentially a control function approach (e.g., Heckman and Robb, 1984; Newey, Powell and Vella, 1999; Blundell and Powell, 2000), with the disturbance η playing the role of the control function. Note that it is only in the last step where we average over the distribution of η , that we use the support condition. If this condition does not hold, we cannot integrate over the marginal distribution of η , at least not at all values of X , because we can only estimate $E[Y|X, \eta]$ at values (X, η) with positive density.

The second identification result uses the additional monotonicity assumption to identify, for some values of X and (η, ν) , the unit-level structural function.

Theorem 2: (IDENTIFICATION OF THE UNIT-LEVEL STRUCTURAL FUNCTION)

If Assumptions 3.1-3.3 are satisfied then the joint distribution of (X, η, ν) is identified, up to normalizations on the distributions of η and ν , and $g(x, \eta, \nu)$ is identified on the joint support of (X, η, ν) .

Note that we do not need a support condition for this theorem, but limit the identification of the structural function to the joint support of (X, η, ν) . For policy II this is sufficient, and it may also be sufficient for evaluations of policy III.

4 Estimation

First consider estimation of $\mu(x)$ under the identifying assumptions 3.1, 3.2, and 3.4. We consider a two-step nonparametric estimator. The first step is construction of a nonparametric estimator $\hat{\eta}_i = \hat{F}(x_i|z_i)$ of the control variable $\eta_i = F(x_i|z_i)$. The second step is construction of a nonparametric estimator $\hat{\beta}(x, \eta)$ of $\beta(x, \eta) = E[y|x, \eta]$. Finally, we estimate the ASF by integrating over the marginal distribution of η , which is normalized to a uniform distribution on the interval $[0, 1]$:

$$\hat{\mu}(x) = \int_0^1 \hat{\beta}(x, \eta) d\eta$$

In both steps we use series estimation.

To describe the first step, let $q_{\ell L}(z)$, ($\ell = 1, \dots, L; L = 1, 2, \dots$) denote approximating functions for the first step. Examples include power series or spline functions. Also, let $q^L(z) = (q_{1L}(z), \dots, q_{LL}(z))'$ and $\hat{Q} = \sum_{i=1}^n q^L(z_i)q^L(z_i)'/n$. A series estimator of the conditional CDF at a particular x and z can be obtained as the predicted value from regressing an indicator function for $x_i \leq x$ on functions of z_i . It has the form

$$\tilde{F}(x|z) = q^L(z)' \hat{Q}^- \sum_{j=1}^n q^L(z_j) 1(x_j \leq x) / n,$$

where A^- denotes any generalized inverse of the matrix A . As is well known, the predicted values $\tilde{F}(x_i|z_i)$ will be invariant to the choice of generalized inverse. Its use is important here because we will allow for \hat{Q} to be singular, even asymptotically.

One feature of this estimator is that it is not necessarily bounded between 0 and 1. To impose that bound by fixed trimming. Let $\tau(\eta) = 1(\eta > 0) \min\{\eta, 1\}$ be the CDF of a uniform distribution. Then our estimate of the control function is given by

$$\hat{\eta}_i = \tau(\tilde{F}(x_i|z_i)).$$

To describe the second step, let $w = (x, v)$ denote the entire vector of regressors in $E[y|x, \eta]$. Let $p_{kK}(w)$, ($k = 1, \dots, K; K = 1, 2, \dots$), be approximating functions of w , $p^K(w) = (p_{1K}(w), \dots, p_{KK}(w))'$, $\hat{w}_i = (x_i, \hat{\eta}_i)$, and $\hat{P} = \sum_{i=1}^n p^K(\hat{w}_i)p^K(\hat{w}_i)'/n$. A nonparametric estimator of $\beta(w) = E[y|w]$ is then

$$\hat{\beta}(w) = p^K(w)' \hat{P}^{-1} \sum_{j=1}^n p^K(\hat{w}_j) y_j / n.$$

Since by construction η_i is uniformly distributed on $[0, 1]$, we can use the estimator

$$\hat{\mu}(x) = \int_0^1 \hat{\beta}(x, v) dv. \quad (4.1)$$

This estimator often has a simple form, that can be obtained analytically. For example, for power series or splines, $\hat{\beta}(x, v)$ will be a linear combination of products of functions of x with functions of v , and the functions of v can easily be integrated analytically to calculate $\hat{\mu}(x)$. Averaging over the sample values \hat{v}_i gives an alternative estimator of the ASF,

$$\tilde{\mu}(x) = \sum_{j=1}^n \hat{\beta}(x, \hat{v}_j) / n \quad (4.2)$$

This has the form of the partial mean estimator of Newey (1994), except that series nonparametric regression is used rather than kernel estimation.

Second we consider estimation of $g(x, \eta, \nu)$ and the joint distribution of (X, η, ν) under identifying assumptions 3.1-3.3. The first step is the same as above, leading to an estimator for the disturbance η_i . Using this estimate $\hat{\eta}_i$, we construct an estimator for the conditional distribution function of Y given X and η , $\hat{F}_{Y|X,\eta}(y|x, \eta)$. The estimator for $g(x, \eta, \nu)$ is then the inverse $\hat{F}_{Y|X,\eta}^{-1}(\nu|x, \eta)$. In addition the estimator for ν_i is $\hat{\nu}_i = \hat{F}_{Y|X,\eta}(y_i|x_i, \hat{\eta}_i)$, which can then be used to construct an estimator for the joint distribution of (X, η, ν) .

we need to add a formal theorem for this estimator, which is essentially a version of the first stage with estimated regressors, combined with something that says it is ok to invert the thing

5 Large Sample Theory

We derive the large sample properties of this estimator first for the case where x and v are scalars, allowing z to be a vector. Also, we focus on the integral estimator (4.1). To derive large sample properties of the estimator it is essential to impose some conditions. The first assumption imposes an approximation rate for the first step regression that is uniform in both the arguments x and z of the conditional distribution function $F(x|z)$. Let \mathcal{X} and \mathcal{Z} denote the support of x_i and z_i , respectively.

Assumption 5.1: *There exists $d_1, C > 0$ such that for every L there is a $L \times 1$ vector $\gamma^L(x)$ satisfying*

$$\sup_{x \in \mathcal{X}, z \in \mathcal{Z}} |F(x|z) - q^L(z)' \gamma^L(x)| \leq CL^{-d_1}.$$

This condition imposes an approximation rate for the CDF that is uniform in both its arguments. It is well known that such rates exist when higher order derivatives are uniformly bounded and the support of z is compact. In particular, it will be satisfied for both splines and power series with $d_1 = s_F/r_z$, if $F(x|z)$ has continuous derivatives

up to order s_F , r_z is the dimension of z , and the spline order is at least s_F . **Lorentz reference?**

We do not impose any other requirement on the distribution of z . This distribution need not be continuous and could even have components that are discrete with infinite support. With this condition in place we obtain the following convergence rate for the sample mean-square error of \hat{v}_i .

Theorem 3: *If Assumption 1 is satisfied,*

$$\sum_{i=1}^n (\hat{v}_i - v_i)^2 / n = O_p(L/n + L^{1-2d_1}).$$

In comparison with previous results in the literature, this one has L^{1-2d_1} in the rate rather than L^{2d_1} . The source of the additional L term is the fact that we are looking at a rate that allows the dependent variable in the regression (i.e. $1(x_j \leq x)$) to vary with the observations.

The estimator of the ASF we have considered has the partial mean form. To obtain a convergence rate for this estimator, that accounts for the fact that it has been averaged over one component, we impose a particular structure on the second step approximating functions. We assume for simplicity that

$$p^K(w) = p_x^{K_x}(x) \otimes p_v^{K_v}(v). \quad (5.3)$$

This product form leads to easily interpretable conditions but can be generalized without affecting the following results. The next condition is a joint restriction on the approximating functions and the distribution of x_i and v_i . Let \mathcal{X} denote the support of x_i and $\lambda_{\min}(A)$ denote the smallest eigenvalue of a symmetric matrix A .

Assumption 5.2: *i) The joint density of (x, v) is bounded above and below by constant multiples of the product of marginal densities; ii) There is a constant C and $\zeta^x(K)$, $\zeta_0^v(K)$, $\zeta_1^v(K)$ such that for each K_x and K_v there exists B_x and B_v such that $\tilde{p}_x^{K_x}(x) = B_x p_x^{K_x}(x)$, $\tilde{p}_v^{K_v}(v) = B_v p_v^{K_v}(v)$, $\lambda_{\min}(E[\tilde{p}_x^{K_x}(x_i) \tilde{p}_x^{K_x}(x_i)']) \geq C$, $\lambda_{\min}(\int_0^1 \tilde{p}_v^{K_v}(v) \tilde{p}_v^{K_v}(v)' dv) \geq$*

$$C, \sup_{x \in \mathcal{X}} \|\tilde{p}_x^{K_x}(x)\| \leq C\zeta^x(K_x), \sup_{v \in [0,1]} \|\tilde{p}_v^{K_v}(v)\| \leq C\zeta_0^v(K_v), \sup_{v \in [0,1]} \|\partial \tilde{p}_v^{K_v}(v)/\partial v\| \leq C\zeta_1^v(K_v).$$

Part ii) of this condition is a normalization like that adopted by Newey (1997), applied separately to the x and v components. Part i) restricts the joint density in such a way that the separate normalizations are effective for the vector of joint approximating functions in equation (5.3). The size of the bounds are known for some important cases. For example, if the joint density of (x_i, v_i) is bounded below and above on a rectangle then this condition will be satisfied for splines and power series with

$$\begin{aligned} \zeta^x(K_x) &= \sqrt{K_x}, \zeta_0^v(K_v) = \sqrt{K_v}, \zeta_1^v(K_v) = K_v^{3/2}; \text{ splines.} \\ \zeta^x(K_x) &= K_x, \zeta_0^v(K_v) = K_v, \zeta_1^v(K_v) = K_v^3; \text{ power series.} \end{aligned}$$

To obtain a convergence rate, it is also important to specify a rate of approximation for $\beta(w)$. Such a rate is imposed in the following condition:

Assumption 5.3: $\beta(w)$ is Lipschitz in v and there exists $s, C > 0$ such that for every K_x and K_v there is a α^K with

$$\sup_{w \in \mathcal{W}} |\beta(w) - p^K(w)' \alpha^K| \leq C(\min\{K_x, K_v\})^{-s}.$$

The size of s is related to the smoothness of $\beta(w)$. Also, $K_x K_v$ is the total number of approximating functions. It is well known that this condition holds for polynomials and splines, where \mathcal{W} is a compact rectangle and s is the number of continuous derivatives that exist. In addition to these assumptions we also require that the following one, which is common in the series estimation literature.

Assumption 5.4: $\text{Var}(y_i|w_i)$ is bounded.

With these conditions in place we can obtain a convergence rate for the second-step estimator. Let $\Delta_n^2 = L/n + L^{1-2d_1}$ and $K = K_x K_v$.

Theorem 4: *If Assumptions 1 - 4 are satisfied, $K\zeta^x(K_x)^2\zeta_0^v(K_v)^2/n \rightarrow 0$ and $K\zeta^x(K_x)^2\zeta_1^v(K_v)^2\Delta_n^2 \rightarrow 0$, $K^2/n \rightarrow 0$, and K_x/K_v is bounded and bounded away from zero then*

$$\int [\hat{\mu}(x) - \mu(x)]^2 F_0(dx) = O_p(K_x/n + K_x^{-2s} + \zeta^x(K_x)^2\zeta_1^v(K_v)^2\Delta_n^2).$$

It is interesting to note that the convergence rate is the sum of two terms. The first $K_x/n + K_x^{-2s}$ is the convergence rate for a nonparametric estimator of $\mu(x)$ assuming the first stage is known. This term is what we would expect to find for a partial mean estimator which averages out over v , and is analogous to the convergence rates found for kernel estimators of partial means in Newey (1994). The other term accounts for the estimation of the control variable in the first step. It goes to zero slower than the first-step convergence rate.

For example, for splines of order at least s we obtain

$$\int [\hat{\mu}(x) - \mu(x)]^2 F_0(dx) = O_p(K_x/n + K_x^{-2s} + K_x K_v^3 \{L/n + L^{1-2d_1}\}).$$

6 Appendix: Proofs of Theorems 1 through 4

Proof of Lemma 1:

Given Model (2.1)-(2.2), define

$$\nu = \nu(Y, Z, \eta) = F_{Y|Z, \eta}(Y|Z, \eta).$$

By definition ν is independent of (Z, η) , and since η is independent of Z by assumption, Assumptions 3.1 and 3.2 are satisfied. To show that Assumption 3.3 also holds we demonstrate that ν , defined as a function of (Y, Z, η) can be written as a function of (Y, X, η) that is strictly monotone in Y . To see this, note that because $(\varepsilon, \eta) \perp Z$, it follows that $\varepsilon \perp Z|\eta$, and hence $\varepsilon \perp Z, h(Z, \eta)|\eta$, or

$$\varepsilon \perp (Z, X) \mid \eta.$$

Hence

$$\begin{aligned}
F_{Y|Z,\eta}(Y|z_0, \eta_0) &= Pr(Y \geq y|Z = z_0, \eta = \eta_0) \\
&= Pr(g(X, \varepsilon) \leq y|Z = z_0, \eta = \eta_0) \\
&= Pr(\varepsilon \leq g^{-1}(X, y)|Z = z_0, \eta = \eta_0) \\
&= Pr(\varepsilon \leq g^{-1}(h(z_0, \eta_0), y)|Z = z_0, \eta = \eta_0) \\
&= Pr(\varepsilon \leq g^{-1}(h(z_0, \eta_0), y)|\eta = \eta_0) \\
&= Pr(\varepsilon \leq g^{-1}(h(z_0, \eta_0), y)|X = h(z_0, \eta_0), \eta = \eta_0) \\
&= Pr(\varepsilon \leq g^{-1}(X, y)|X = h(z_0, \eta_0), \eta = \eta_0) \\
&= Pr(g(X, \varepsilon) \leq y|X = h(z_0, \eta_0), \eta = \eta_0) \\
&= Pr(Y \leq y|X = h(z_0, \eta_0), \eta = \eta_0) = F_{Y|X,\eta}(Y, h(z_0, \eta_0), \eta_0).
\end{aligned}$$

Hence, ν can be written as $\nu(Y, Z, \eta) = F_{Y|Z,\eta}(Y|Z, \eta)$ or as $\nu = \nu(Y, X, \eta) = F_{Y|X,\eta}(Y|X, \eta)$.

Since the last representation shows that ν can be written as a function of (Y, X, η) that is strictly monotone in Y , the proof is completed. Q.E.D

Proof of Theorem 1:

We normalize the marginal distribution of η to a uniform distribution on the interval $[0, 1]$. Then:

$$\begin{aligned}
F_{X|Z}(x_0|z_0) &= Pr(X \leq x_0|Z = z_0) = Pr(h(Z, \eta) \leq x_0|Z = z_0) \\
&= Pr(\eta \leq h^{-1}(Z, x_0)|Z = z_0) = Pr(\eta \leq h^{-1}(z_0, x_0)|Z) \\
&= F_\eta(h^{-1}(z_0, x_0)) = h^{-1}(z_0, x_0).
\end{aligned}$$

Since we can estimate the lefthand side, the conditional distribution function of X given Z , we can estimate the inverse $h^{-1}(z, x)$ of the function of interest, and hence $h(x, \eta)$ itself as $h(z, \eta)$ is invertible. As a by-product we get the value of $\eta = h^{-1}(Z, X)$.

Since ν is independent of X given η , we can derive

$$\mu(x, \eta) = E[g(x, \eta, \nu)|\eta] = E[g(X, \eta, \nu)|X = x, \eta] = E[Y|X = x, \eta],$$

for all values of (x, η) in the joint support of (X, η) , as well as the joint distribution of (X, η) . Hence we can calculate integrals of the type

$$\int \mu(x, \eta) \lambda(x, \eta) f_{\eta, x}(\eta, x) d\eta,$$

for any weight function $\lambda(x, \eta)$. The average structural function

$$\mu(x) = \int \mu(x, \eta) f_{\eta}(\eta) d\eta$$

corresponds to the choice

$$\lambda(x, \eta) = \frac{f_{\eta}(\eta)}{f_{\eta, x}(\eta, x)}.$$

Thus $\mu(x)$ is identified if the conditional density of η given $X = x$ is positive everywhere on the support of η so $\lambda(x, \eta)$ is finite at each pair (x, η) . Q.E.D

Proof of Theorem 2:

We normalize the marginal distributions of η and ν to uniform distributions on the interval $[0, 1]$. Theorem 1 shows that $h(z, \eta)$ is identified.

Next we follow the same procedure to estimate ν , since conditional on η , ν and X are independent:

$$\begin{aligned} F_{Y|X, \eta}(y_0, x_0, \eta_0) &= Pr(Y \leq y_0 | X = x_0, \eta = \eta_0) \\ &= Pr(g(X, \eta, \nu) \leq y_0 | X = x_0, \eta = \eta_0) \\ &= Pr(\nu \leq g^{-1}(X, \eta, y_0) | X = x_0, \eta = \eta_0) \\ &= Pr(\nu \leq g^{-1}(x_0, \eta_0, y_0) | X = x_0, \eta = \eta_0) \\ &= F_{\nu}(g^{-1}(x_0, \eta_0, y_0)) = g^{-1}(x_0, \eta_0, y_0). \end{aligned}$$

For all values (x_0, η_0) in the joint distribution of (X, η) we can estimate this conditional distribution function, and hence for all those values we can infer the inverse of the function $g(x, \eta, \nu)$ and thus the function itself. Q.E.D.

Throughout the remainder of the Appendix, C will denote a generic positive constant that may be different in different uses. Also, with probability approaching one will be

abbreviated as w.p.a.1, positive semi-definite as p.s.d., positive definite as p.d., $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, and $A^{1/2}$ will denote the minimum and maximum eigenvalues, and square root, of respectively of a symmetric matrix A . Let \sum_i denote $\sum_{i=1}^n$. Also, let CS, M, and T refer to the Cauchy-Schwartz, Markov, and triangle inequalities, respectively.

Proof of Theorem 3: Let $q_i = q^L(z_i)$, $\varepsilon_{ij} = 1(x_j \leq x_i) - F(x_i|z_j)$, and $\delta_{ij} = F(x_i|z_j) - q'_j \gamma^L(x_i)$. Then for $\tilde{v}_i = \tilde{F}(x_i|z_i)$ and $v_i = F(x_i|z_i)$,

$$\begin{aligned}\tilde{v}_i - v_i &= \Delta_i^I + \Delta_i^{II} + \Delta_i^{III}, \Delta_i^I = q'_i \hat{Q}^- \sum_{j=1}^n q_j \varepsilon_{ij} / n, \\ \Delta_i^{II} &= q'_i \hat{Q}^- \sum_{j=1}^n q_j \delta_{ij} / n, \Delta_i^{III} = -\delta_{ii}.\end{aligned}$$

(minus sign in front of δ_{ii})

Note that $|\Delta_i^{III}| \leq CL^{-d_1}$ by Assumption 5.1. Also, by \hat{Q} p.s.d. and symmetric there exists a diagonal matrix of eigenvalues Λ and an orthonormal matrix B such that $\hat{Q} = B\Lambda B'$. Let Λ^- denote the diagonal matrix of inverse of nonzero eigenvalues and zeros and $\hat{Q}^- = B\Lambda^- B'$. Then $tr(\hat{Q}^- \hat{Q}) \leq L$. By CS and Assumption 5.1,

$$\begin{aligned}\sum_{i=1}^n (\Delta_i^{III})^2 / n &\leq \sum_{i=1}^n (q'_i \hat{Q}^- q_i \sum_{j=1}^n \delta_{ij}^2 / n) / n \leq C \sum_{i=1}^n (q'_i \hat{Q}^- q_i) L^{-2\alpha_1} / n \\ &= CL^{-2\alpha_1} tr(\hat{Q}^- \hat{Q}) \leq CL^{1-2\alpha_1}.\end{aligned}$$

in second expression in first line, added a $/n$. Also, I could not figure out what happened to the term $\sum q_j \delta_{ij} \delta_{ii} q'_i$ in the first inequality sign

Furthermore, we have $E[\varepsilon_{ii}^2 | Z] \leq C$. isn't this ≤ 1 ?

Also

$$E[\varepsilon_{ij}^2 | x_i, Z] = E[\varepsilon_{ij}^2 | x_i, z_i, z_j] = E[1(x_j \leq x_i) | x_i, z_i, z_j] \{1 - 2F(x_i|z_j)\} + F(x_i|z_j)^2 \leq C.$$

Furthermore, note that for $i \neq j \neq k$, **only need $j \neq k$ here**

it follows by independence of the observations,

$$\begin{aligned}E[\varepsilon_{ij} \varepsilon_{ik} | Z] &= E[E[\varepsilon_{ij} \varepsilon_{ik} | Z, x_i, x_k] | Z] \\ &= E[\varepsilon_{ik} E[\varepsilon_{ij} | Z, x_i, x_k] | Z] = E[\varepsilon_{ik} E[\varepsilon_{ij} | z_j, z_i, x_i] | Z]\end{aligned}$$

$$\begin{aligned}
&= E[\varepsilon_{ik}E[\varepsilon_{ij}|z_j, z_i, x_i]|Z] = E[\varepsilon_{ik}\{E[1(x_j \leq x_i)|z_j, z_i, x_i] - F(x_i|z_j)\}|Z] \\
&= 0.
\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}
E[\sum_{i=1}^n (\Delta_i^I)^2/n|Z] &= \sum_{i=1}^n q'_i \hat{Q}^- (\sum_{j,k=1}^n q_j E[\varepsilon_{ij}\varepsilon_{ik}|Z] q'_k/n^2) \hat{Q}^- q_i/n \\
&\leq \sum_{i=1}^n q'_i \hat{Q}^- \hat{Q} \hat{Q}^- q_i/n^2 \leq CL/n.
\end{aligned}$$

The conclusion then follows by the triangle inequality and by $|\tau(\tilde{v}) - \tau(v)| \leq |\tilde{v} - v|$, which gives $\sum_i (\hat{v}_i - v_i)^2/n = \sum_i (\tau(\tilde{v}_i) - \tau(v_i))^2/n \leq \sum_i (\tilde{v}_i - v_i)^2/n$. Q.E.D..

Before proving other results we give some useful lemmas. For these results let $p_i = p^K(w_i)$, $\hat{p}_i = p^K(\hat{w}_i)$, $p = [p_1, \dots, p_n]$, $\hat{p} = [\hat{p}_1, \dots, \hat{p}_n]$, $\hat{P} = \hat{p}'\hat{p}/n$, $\tilde{P} = p'p/n$, $P = E[p_i p'_i]$. In the statement of these results we allow \hat{v}_i and v_i to be vectors.

Lemma A1: *In Assumption 5.2 it can be assumed without loss of generality that*

- (i) $E[\tilde{p}_x^{K_x}(x_i)\tilde{p}_x^{K_x}(x_i)'] = I_{K_x}$,
- (ii), $E[\tilde{p}_v^{K_v}(v_i)\tilde{p}_v^{K_v}(v_i)'] = I_{K_v}$,
- (iii), $E[\tilde{p}_v^{K_v}(v_i)] = e_1 = (1, 0, \dots, 0)'$,
- (iv), $E[\tilde{p}^K(w_i)\tilde{p}^K(w_i)'] \leq CI_K$, and
- (v), $\lambda_{\min}(E[\tilde{p}^K(w_i)\tilde{p}^K(w_i)']) \geq C$.

Proof: The first two conclusions can be shown as in the beginning of the Appendix of Newey (1997). For the third conclusion, note that by $c'p_v^{K_v}(v_i) = 1$ for some c , it follows that there is a \tilde{c} such that $\tilde{c}'\tilde{p}_v^{K_v}(v_i) = 1$. Note that $\tilde{c}'\tilde{c} = \tilde{c}'E[\tilde{p}_v^{K_v}(v_i)\tilde{p}_v^{K_v}(v_i)']\tilde{c} = E[\{\tilde{c}'\tilde{p}_v^{K_v}(v_i)\}^2] = 1$. Let \tilde{B} be an orthonormal matrix with c' as its first row. Then $\tilde{B}\tilde{p}_v^{K_v}(v_i)$ is an orthonormal basis satisfying all the conditions of Assumption 5.2, and satisfying (i) and (iii) in Lemma A.1. In particular, $\|\tilde{B}\tilde{p}_v^{K_v}(v)\| = \|\tilde{p}_v^{K_v}(v)\|$ and $\|\tilde{B}\partial\tilde{p}_v^{K_v}(v)/\partial v\| = \|\partial\tilde{p}_v^{K_v}(v)/\partial v\|$. Also, by the joint density bounded above by the product of the marginals it follows that

$$\begin{aligned}
E[\tilde{p}^K(w_i)\tilde{p}^K(w_i)'] &= E[\tilde{p}_x^{K_x}(x_i)\tilde{p}_x^{K_x}(x_i)' \otimes \tilde{p}_v^{K_v}(v_i)\tilde{p}_v^{K_v}(v_i)'] \\
&\leq CE[\tilde{p}_x^{K_x}(x_i)\tilde{p}_x^{K_x}(x_i)'] \otimes E[\tilde{p}_v^{K_v}(v_i)\tilde{p}_v^{K_v}(v_i)'] = CI_K.
\end{aligned}$$

Finally, it follows similarly that by the density bounded below by a product of the marginals that

$$E[\tilde{p}^K(w_i)\tilde{p}^K(w_i)'] \geq CE[\tilde{p}_x^{K_x}(x_i)\tilde{p}_x^{K_x}(x_i)'] \otimes E[\tilde{p}_v^{K_v}(v_i)\tilde{p}_v^{K_v}(v_i)'] = CI_K,$$

giving the conclusion (v). Q.E.D.

Lemma A2: *If $\sum_i \|\hat{v}_i - v_i\|^2/n = O_p(\Delta_n^2)$ and Assumption 5.2 is satisfied then*

$$\begin{aligned} \sum_i \|\hat{p}_i - p_i\|^2/n &= O_p(\zeta^x(K_x)^2 \zeta_1^v(K_v)^2 \Delta_n^2), \\ \|\hat{P} - \tilde{P}\| &= O_p(\zeta^x(K_x)^2 \zeta_1^v(K_v)^2 \Delta_n^2 + \sqrt{K} \zeta^x(K_x) \zeta_1^v(K_v) \Delta_n). \end{aligned} \quad (\text{A.1})$$

Proof: Because the estimator is invariant to nonsingular linear transformations of $p_x^{K_x}(x)$ and $p_v^{K_v}(v)$ we can let $p_x^{K_x}(x) = \tilde{p}_x^{K_x}(x)$ and $p_v^{K_v}(v) = \tilde{p}_v^{K_v}(v)$ in Assumption 5.2 and Lemma A1. Also, a mean value expansion gives $\hat{p}_i = p_i + [\partial p^K(\bar{w}_i)/\partial v](\hat{v}_i - v_i)$, where $\bar{w}_i = (x_i, \bar{v}_i)$ and \bar{v}_i lies in between \hat{v}_i and v_i . It follows that $\bar{v}_i \in [0, 1]$ so that $\|\partial p^K(\bar{w}_i)/\partial v\| \leq C\zeta^x(K_x)\zeta_1^v(K_v)$. Then by CS, $\|\hat{p}_i - p_i\| \leq C\zeta^x(K_x)\zeta_1^v(K_v)|\hat{v}_i - v_i|$. Summing up gives

$$\sum_{i=1}^n \|\hat{p}_i - p_i\|^2/n = O_p(\zeta^x(K_x)^2 \zeta_1^v(K_v)^2 \Delta_n^2). \quad (\text{A.2})$$

By Lemma A1, $\sum_{i=1}^n \|p_i\|^2/n = O_p(E[\|p_i\|^2]) = \text{tr}(I_K) = K$. Then by T, CS, and eq. (??),

$$\begin{aligned} \|\hat{P} - \tilde{P}\| &\leq \sum_{i=1}^n \|\hat{p}_i \hat{p}_i' - p_i p_i'\|/n \leq \sum_{i=1}^n \|\hat{p}_i - p_i\|^2/n + 2\left(\sum_{i=1}^n \|\hat{p}_i - p_i\|^2/n\right)^{1/2} \left(\sum_{i=1}^n \|p_i\|^2/n\right)^{1/2}. \\ &= O_p(\zeta^x(K_x)^2 \zeta_1^v(K_v)^2 \Delta_n^2 + \sqrt{K} \zeta^x(K_x) \zeta_1^v(K_v) \Delta_n). \text{ Q.E.D.} \end{aligned}$$

Lemma A3: *If $\sum_i \|\hat{v}_i - v_i\|^2/n = O_p(\Delta_n^2)$, Assumption 5.2 is satisfied, $\sqrt{K}\zeta^x(K_x)\zeta_1^v(K_v)\Delta_n \rightarrow 0$, and $K\zeta^x(K_x)^2\zeta_0^v(K_v)^2/n \rightarrow 0$ then w.p.a.1, $\lambda_{\min}(\hat{P}) \geq C$, $\lambda_{\min}(\tilde{P}) \geq C$, and $\lambda_{\min}(P) \geq C$.*

Proof: The last conclusion follows by Lemma A1. Also, by Lemma A2, $\|\hat{P} - \tilde{P}\| \xrightarrow{p} 0$ while as in Newey (1997) it follows that

$$\|\tilde{P} - P\| = O_p(\sqrt{K}\zeta^x(K_x)\zeta_0^v(K_v)/\sqrt{n}) \xrightarrow{p} 0. \quad (\text{A.3})$$

Therefore, by the triangle inequality, $\|\hat{P} - P\| \xrightarrow{p} 0$. The other conclusions then follow as in Newey (1997). Q.E.D.

Lemma A4: *If $\sum_i \|\hat{v}_i - v_i\|^2/n = O_p(\Delta_n^2)$ and Assumption 5.2 is satisfied then for $\varepsilon_i = y_i - \beta(w_i)$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$,*

$$\|(\hat{p} - p)' \varepsilon/n\| = O_p(\zeta^x(K_x) \zeta_1^v(K_v) \Delta_n), \|p' \varepsilon/n\| = O_p(\sqrt{K/n}).$$

Proof: Note that $E[\varepsilon_i^2] < \infty$ by $E[\varepsilon_i^2 | w_i] = \text{Var}(y_i | w_i)$ bounded. Then by T, CS, M, and Lemma A2,

$$\begin{aligned} \|(\hat{p} - p)' \varepsilon/n\| &\leq \sum_i \|\hat{p}_i - p_i\| |\varepsilon_i|/n \leq \left(\sum_i \|\hat{p}_i - p_i\|^2/n\right)^{1/2} \left(\sum_i \varepsilon_i^2/n\right)^{1/2} \\ &= O_p(\zeta^x(K_x) \zeta_1^v(K_v) \Delta_n) O_p(1) = O_p(\zeta^x(K_x) \zeta_1^v(K_v) \Delta_n). \end{aligned}$$

giving the first conclusion. Also, for $W = (w_1, \dots, w_n)$, note that as in Newey (1997), $E[\varepsilon \varepsilon' | W] \leq CI_n$. Then by iterated expectations and Lemma A1,

$$\begin{aligned} E[\|p' \varepsilon/n\|^2] &= E[\text{tr}(p' \varepsilon \varepsilon p)/n^2] = E[\text{tr}(p' E[\varepsilon \varepsilon' | W] p)/n^2] \\ &\leq \text{Ctr}(E[p' p])/n^2 = \text{Ctr}(E[p_i p_i'])/n \leq CK/n. \end{aligned} \tag{A.4}$$

The second conclusion then follows by M. Q.E.D.

Proof of Theorem 4: Let $\hat{\alpha} = \hat{P}^{-1} \hat{p}' y/n$ and

$$\bar{p}(x) = \int_0^1 p_x^{K_x}(x) \otimes p_v^{K_v}(v) dv = p_x^{K_x}(x) \otimes e_1,$$

where the last equality follows by Lemma A1. Then

$$\hat{\mu}(x) = \bar{p}(x)' \hat{\alpha}.$$

Note that by Lemma A1,

$$E[\bar{p}(x_i) \bar{p}(x_i)'] = I_{K_x} \otimes e_1 e_1' \leq I_K. \tag{A.5}$$

Also, for $\alpha = \alpha_K$ from Assumption 5.3, $\beta = (\beta(w_1), \dots, \beta(w_n))'$, and $\hat{\beta} = (\beta(\hat{w}_1), \dots, \beta(\hat{w}_n))'$,

$$\begin{aligned} \hat{\alpha} - \alpha &= \tilde{P}^{-1}p'\varepsilon/n + \hat{P}^{-1}(\tilde{P} - \hat{P})\tilde{P}^{-1}p'\varepsilon/n + \hat{P}^{-1}(\hat{p} - p)'\varepsilon/n \\ &\quad + \hat{P}^{-1}\hat{p}'(\beta - \hat{\beta})/n + \hat{P}^{-1}\hat{p}'(\hat{\beta} - \hat{p}'\alpha)/n. \end{aligned} \quad (\text{A.6})$$

Note that $\hat{p}\hat{P}^{-1}\hat{p}'/n$ is idempotent, so that $\hat{p}\hat{P}^{-1}\hat{p}'/n \leq I$. Also, by the smallest eigenvalue of \hat{P} bounded away from zero w.p.a.1, we have, for any vector a , $\|\hat{P}^{-1}a\| \leq C\|a\|$ and $\|\hat{P}^{-1}a\|^2 \leq Ca'\hat{P}^{-1}a$. Note that as in Newey (1997), $E[\varepsilon\varepsilon'|W] \leq CI_n$, so that by the Fubini Theorem,

$$\begin{aligned} E[\int \{\bar{p}(x)'\tilde{P}^{-1}p'\varepsilon/n\}^2 F_0(dx)|W] &= \int \{\bar{p}(x)'\tilde{P}^{-1}p'E[\varepsilon\varepsilon'|W]p\tilde{P}^{-1}\bar{p}(x)^2 F_0(dx)/n^2 \\ &\leq \int \bar{p}(x)'\tilde{P}^{-1}\bar{p}(x)F_0(dx)/n \leq CE[\bar{p}(x_i)'\bar{p}(x_i)]/n \\ &= CE[\tilde{p}_x^{K_x}(x_i)'\tilde{p}_x^{K_x}(x_i) \otimes e_1'e_1]/n = K_x/n. \end{aligned}$$

A well known implication of this inequality is that

$$\int \{\bar{p}(x)'\tilde{P}^{-1}p'\varepsilon/n\}^2 F_0(dx) = O_p(K_x/n). \quad (\text{A.7})$$

Next, by Lemma A3 there is C such that w.p.a.1, $\|\hat{P}^{-1}a\| \leq C\|a\|$ and $\|\hat{P}^{-1}a\|^2 \leq Ca'\hat{P}^{-1}a$ for all conformable vectors a . Then it follows by Lemmas A2 and A4, CS, eq. (A.5), and $K^2/n \rightarrow 0$ that

$$\begin{aligned} \int \{\bar{p}(x)'\hat{P}^{-1}(\tilde{P} - \hat{P})\tilde{P}^{-1}p'\varepsilon/n\}^2 F_0(dx) &= \|\hat{P}^{-1}(\tilde{P} - \hat{P})\tilde{P}^{-1}p'\varepsilon/n\|^2 \leq C\|\tilde{P} - \hat{P}\|^2 \|p'\varepsilon/n\|^2 \\ &= O_p((K^2/n)\zeta^x(K_x)^2\zeta_1^v(K_v)^2\Delta_n^2) \\ &= O_p(\zeta^x(K_x)^2\zeta_1^v(K_v)^2\Delta_n^2). \end{aligned}$$

replaced last expression

It follows similarly from Lemma A4 that

$$\begin{aligned} \int \{\bar{p}(x)'\hat{P}^{-1}(\hat{p} - p)'\varepsilon/n\}^2 F_0(dx) &= \|\hat{P}^{-1}(\hat{p} - p)'\varepsilon/n\|^2 \leq C\|(\hat{p} - p)'\varepsilon/n\|^2 \\ &= O_p(\zeta^x(K_x)^2\zeta_1^v(K_v)^2\Delta_n^2). \end{aligned} \quad (\text{A.9})$$

Furthermore, by $\beta(w)$ Lipschitz in v ,

$$\int \{\bar{p}(x)'\hat{P}^{-1}\hat{p}'(\beta - \hat{\beta})/n\}^2 F_0(dx) = \|\hat{P}^{-1}\hat{p}'(\beta - \hat{\beta})/n\|^2 \leq C(\beta - \hat{\beta})'\hat{p}\hat{P}^{-1}\hat{p}'(\beta - \hat{\beta})/n$$

$$\begin{aligned}
&\leq C \sum_{i=1}^n [\beta(w_i) - \beta(\hat{w}_i)]^2/n \leq C \sum_{i=1}^n (v_i - \hat{v}_i)^2/n \\
&= O_p(\Delta_n^2) = o_p(\zeta^x(K_x)^2 \zeta_1^v(K_v)^2 \Delta_n^2).
\end{aligned}$$

added last expression

Also by Assumption A3 we have for $\alpha = \alpha_K$, by $\min\{K_x, K_v\} \geq CK_x$,

$$\begin{aligned}
\int \{\bar{p}(x)' \hat{P}^{-1} \hat{p}'(\hat{\beta} - \hat{p}\alpha)/n\}^2 F_0(dx) &= \|\hat{P}^{-1} \hat{p}'(\hat{\beta} - \hat{p}\alpha)/n\|^2 \leq C(\hat{\beta} - \hat{p}\alpha)' \hat{p} \hat{P}^{-1} \hat{p}'(\hat{\beta} - \hat{p}\alpha) \\
&\leq C \sum_{i=1}^n \{\beta(\hat{w}_i) - p^K(\hat{w}_i)' \alpha\}^2/n \\
&\leq C \sup_{w \in \mathcal{W}} |\beta_0(w) - p^K(w)' \alpha|^2 = O(K_x^{-2s}).
\end{aligned}$$

It then follows from eq. (A.6), eqs. (A.7)-(A.11), and T that

$$\begin{aligned}
\int [\hat{\mu}(x) - \bar{p}(x)' \alpha]^2 F_0(dx) &= \int [\bar{p}(x)'(\hat{\alpha} - \alpha)]^2 F_0(dx) \\
&= O_p(K_x/n + K_x^{-2s} + \zeta^x(K_x)^2 \zeta_1^v(K_v)^2 \Delta_n^2).
\end{aligned}$$

Furthermore, by CS we also have

$$\int [\bar{p}(x)' \alpha - \mu(x)]^2 F_0(dx) \leq \int \int_0^1 [p^K(w)' \alpha - \beta(w)]^2 dv F_0(dx) \leq CK_x^{-2s}.$$

The conclusion then follows by T. Q.E.D.

REFERENCES

- Altonji, J., and R. Matzkin (2001), “Panel Data Estimators for Nonseparable Models with Endogenous Regressors”, Department of Economics, Northwestern University.
- Angrist, J., G.W. Imbens, and D. Rubin (1996): “Identification of Causal Effects Using Instrumental Variables,” *Journal of the American Statistical Association* 91, 444-472.
- Angrist, J., K. Graddy, and G.W. Imbens (2000): “Nonparametric Demand Analysis with an Application to the Demand for Fish,” *Review of Economic Studies*.
- Athey, S., and P. Haile (2000), “Identification of Standard Auction Models”

- Bajari, P., and L. Benkard (2001)
- Blundell, R., and J.L. Powell (2000): "Endogeneity in Nonparametric and Semiparametric Regression Models," invited lecture, 2000 World Congress of the Econometric Society.
- Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics* 34, 305-334.
- Darolles, S., J.-P., Florens, and E. Renault, (2001), "Nonparametric Instrumental Regression".
- Das, M. (2000): "Nonparametric Instrumental Variable Estimation with Discrete Endogenous Regressors," Working Paper, Department of Economics, Columbia University.
- Das, M. (2001): "Monotone Comparative Statics and the Estimation of Behavioral Parameters," Working Paper, Department of Economics, Columbia University.
- Hausman, J.A. and W.K. Newey (1995), "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," with J.A. Hausman, *Econometrica* 63, 1445-1476.
- Heckman, J. (1990): "Varieties of Selection Bias," *American Economic Review, Papers and Proceedings* 80.
- Heckman, J., and E. Vytlacil, (2000), "Local Instrumental Variables", Chapter 1, in Hsiao, Morimune, and Powell, (eds.) *Nonlinear Statistical Modelling*, Cambridge University Press, Cambridge.
- Imbens, G.W. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62, 467-476.
- Matzkin, R. (1999), "Nonparametric Estimation of Nonadditive Random Functions", Department of Economics, Northwestern University.

- Milgrom, P., and C. Shannon, (1994), "Monotone Comparative Statics," *Econometrica*, 58, 1255-1312.
- Newey, W.K. (1994), "Kernel Estimation of Partial Means and a Variance Estimator", *Econometric Theory* 10, 233-253.
- Newey, W.K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79, 147-168.
- Newey, W.K. and J.L. Powell (1988): "Nonparametric Instrumental Variables Estimation," Working paper, Princeton University.
- Newey, W.K., J.L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica* 67, 565-603.
- Pearl, J. (2000), *Causality*, Cambridge University Press, Cambridge, MA.
- Pinkse, J., (2000a): "Nonparametric Two-step Regression Functions when Regressors and Error are Dependent," *Canadian Journal of Statistics* 28, 289-300.
- Pinkse, J. (2000b): "Nonparametric Regression Estimation Using Weak Separability", University of British Columbia.
- Powell, J., J. Stock, and T. Stoker, *Econometrica*.
- Roehrig, C. (1988): "Conditions for Identification in Nonparametric and Parametric Models", *Econometrica* 55, 875-891.
- Stoker, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica* 54, 1461-1481.
- Vytlacil, E. (2001): "Independence, Monotonicity, and Latent Variable Models: An Equivalence Result," *Econometrica*, forthcoming.