Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity*

Guido W. Imbens
UC Berkeley, and NBER

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

First Draft: March 2001
This Draft: October 2003

Abstract

This paper investigates identification and estimation in a nonparametric structural model with instrumental variables and non-additive errors. We allow for non-additive errors because the endogeneity of choices that is often motivated by unobserved heterogeneity in marginal returns requires objective functions that are non-additive. We also allow for a structural disturbance of arbitrary dimension. We formulate independence and monotonicity conditions that are sufficient for identification of a number of objects of interest, including quantile and average effects. We consider a two-step approach to identification and estimation. The first step is the conditional distribution function of the endogenous regressor given the instrument, which is used as a generalized control function. In the second step the control function is used as a regressor along with the endogenous variable. We propose series estimators of both first and second steps. For inference we establish rates of convergence, asymptotic normality, and give a consistent asymptotic variance estimator.

JEL Classification: C13, C30

Keywords: *Simultaneous equations models, Instrumental Variables, Additivity, Nonlinear Models, Nonparametric Estimation, Series Estimation*

*This research was partially completed while the second author was a fellow at the Center for Advanced Study in the Behavioral Sciences during the year 2000/2001. The NSF provided partial financial support through grants SES 0136789 (Imbens) and SES 0136869 (Newey). We are grateful for comments by Susan Athey, Lanier Benkard, Gary Chamberlain, Jim Heckman, Aviv Nevo, Ariel Pakes, Jim Powell and participants at seminars at Stanford University, University College London, Harvard University, and Northwestern University.

1 Introduction

Structural simultaneous equations models have long been of great interest to econometricians. Recently interest has focused on nonparametric identification and estimation. In this paper we consider a two-equation triangular simultaneous equations model, with a single right-hand side endogenous variable having a reduced form, where both equations are nonseparable in disturbances. This nonseparability is a key feature of many economic models. We assume that the disturbances are independent of instruments and that the reduced form is monotonic in its disturbance.

We give identification and estimation results when the structural equation disturbance has a general form, i.e. can be a vector of any dimension. This general disturbance case allows for individual heterogeneity and other random effects in a completely flexible way. We relate these results to other, more conventional, cases where the disturbance is a scalar or a pair of random variables.

Specifically we consider identification of quantiles of the structural outcome for a fixed value of the endogenous variables. Such quantiles correspond to the value of the structural function at quantiles of the disturbance when the disturbance is a scalar and otherwise can be used to characterize how endogenous variables affect structural outcomes. Differences of these quantiles over values of the endogenous regressors correspond to quantile treatment effects as in Lehman (1974). We also consider identification and estimation of averages over disturbances of features of the structural function. Such averages have long been of interest, because they summarize structural effects for a whole population, whereas quantile effects only apply to a small segment of the population. An early example of these averages is Chamberlain's (1984) average over individual effects of binary choice probabilities. Many more examples are found in the large literature on (binary) treatment effects, where differences of these averages correspond to the average treatment effect (e.g., Heckman, 1990; Imbens, 2003). These averages are also important in the current literature on structural estimation in nonlinear models, and were considered by Blundell and Powell (2002), Altonji and Matzkin (2001), and Wooldridge (2002). Furthermore, we give identification results for average derivatives and some specific economic policies such as the imposition of an upper or lower limit on the endogenous regressor.

We employ a multi-step approach to identification and estimation. The first step is to obtain the conditional distribution function of the endogenous regressor given the instruments. Evaluating this conditional distribution function at the observed values gives a generalized "control function." The second-step consists of obtaining the conditional quantiles or expectations of the outcome of interest given the endogenous variable and the control function. Various structural effects are then recovered by averaging over the control function or the endogenous variable and control function together.

An interesting feature of the model is that when the reduced form is linear the joint density of the endogenous regressor and the control function goes to zero as the control function goes to one or zero, even though the marginal density of the control function is uniform. Thus, averaging over the control function "upweights" the tails relative to the joint distribution. Consequently, estimators that average over the marginal distribution of the control function have a convergence rate that is no faster than the full-dimensional rate. Also, the rate of convergence will be affected by how fast the density goes to zero on the boundary, which is related to the r-squared of the reduced form in the Gaussian case. Estimators of the average derivative, and other averages over the joint distribution, do not suffer from this "upweighting," and so will converge faster.

We consider series estimation for both steps. Series estimators are theoretically useful for the second step because one can obtain convergence rates that allow for the density to go to zero on the boundary. Series estimators are also useful for computation. We derive convergence rates and show asymptotic normality of these estimators.

We extend Blundell and Powell (2002) by using a generalized control function, that allows for a nonseparable reduced form, by considering structural objects that are averages over the joint distribution of disturbances and endogenous variables, and by explicit consideration of quantile effects. All of this work extends Newey, Powell, and Vella (1999) and Pinske (2000b) by allowing for nonseparable disturbances in the structural equation. A different kind of nonparametric structural model, based on conditional mean restrictions, has been considered by Newey and Powell (1988, 2003), Das (2000), and Darrolles, Florens, and Renault (2003). With one or two structural disturbances, our model is a special case of that of Roehrig (1988), although our identification results are not, because we do not require differentiability. Imbens and Angrist (1994) and Angrist, Graddy, and Imbens (2000) also allow for nonseparable disturbances but focus on the interpretations of the limit of linear instrumental variable estimators. Das (2001) also allows for nonseparable disturbances, but considers a single index setting with monotonicity. Chesher (2003) and Ma and Koenker (2003) consider local identification and estimation of a nonseparable model with two structural disturbances. One of our identification results, which was given in Imbens and Newey (2001), is related to Chesher's (2003) results, as further explained below. Chesher (2002) also considers identification under index restrictions with multiple disturbances. Chernozhukov and Hansen (2002) and Chernozhukov, Imbens and Newey (2003) drop the control function assumption when the structural equation is strictly monotonic in one disturbance.

In Section 2 of the paper we present and motivate our models. Section 3 considers identification. Section 4 describes the estimators and Section 5 gives some large sample theory. A small Monte Carlo study is presented in Section 6.

2 The Model

The model we consider has a single right-hand side endogenous variable X_1 and a vector of exogenous variables $Z = (Z_1, Z_2)$. The first equation is a reduced form for X_1,

$$X_1 = h(Z, \eta), \tag{2.1}$$

where η is a scalar disturbance. The second equation is the structural equation for the outcome of interest Y,

$$Y = g(X_1, Z_1, \varepsilon), \tag{2.2}$$

where ε is a general disturbance vector. We are primarily interested in the relation between $X = (X_1, Z_1)$ and Y, as well as more generally in the effect that policies which change the distribution of X have on the distribution of Y. Endogeneity arises from statistical dependence between ε and η. The instrument Z will be assumed to be independent of the disturbances η and ε. We assume X_1 and Y are scalars, and allow Z to be a vector, although many of the results in the paper can be generalized to triangular systems of equations.

An economic example may help explain the model. For simplicity suppose Z_1 is absent, so that $X_1 = X$. Let Y denote some outcome such as firm revenue or individual lifetime earnings and let X be chosen by the individual agent. Here $g(x, e)$ is the (educational) production function and ε represent inputs at most partially observed by agents or firms, with e being a possible value for ε. The agent optimally chooses X by maximizing the expected outcome minus the costs associated with x given her information set. The information set consists of a scalar noisy signal η of the unobserved input ε and a cost shifter Z.[1] The cost function is $c(x, z)$. Then X would be obtained as the solution to the individual choice problem

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

leading to $X = h(Z, \eta)$. Thus, this economic example leads to a triangular system of the type we consider.

When X is schooling and Y is earnings this example corresponds to models for educational choices with heterogenous returns such as the one used by Card (2001) and Das (2001). When

[1] Although we do not do so in the present example, we could allow the cost to depend on the signal η, if, for example financial aid was partly tied to test scores.

X is an input and Y is output, this example is a non-additive extension of a classical problem in the estimation of production functions, e.g., Mundlak (1963). Note the importance of allowing the production function $g(x, e)$ to be non-additive in e (and thus allowing the marginal returns $\frac{\partial g}{\partial x}(x, \varepsilon)$ to vary with the unobserved heterogeneity). If the objective function $g(x, e)$ were additive in e, so that $g(x, \varepsilon) = g_0(x) + \varepsilon$ the optimal level of x would be $\text{argmax}_x\{g_0(x) + \mathbb{E}[\varepsilon|\eta] - c(x, Z)\}$. In that case the solution would depend on Z but not on η, and thus X would be exogenous. Hence in these models nonadditivity is important for generating endogeneity of choices.

A scalar reduced form disturbance η (and the resulting identifiability of η) is essential to our results, because we use a "control function" approach to identification and estimation, where conditioning on η removes the endogeneity of X_1.[2] In comparison, many of the objects we identify and estimate allow for ε to have any dimension. Thus, individual heterogeneity and other sources of of uncertainty can affect the structural response function in a completely flexible way. If we assume that ε has a specific dimension, the interpretation of $g(x, e)$ might be sensitive to the correctness of that assumption. We try to avoid such interpretation issues by allowing ε to have any dimension, although we will show some links to models with specified dimension for ε. On the other hand, by being agnostic about the dimension of ε, we do lose the ability to identify effects corresponding to particular elements of ε, which might be useful in some contexts.

For example, suppose ε is a scalar and the conditional CDF $F_{\varepsilon|\eta}(e|t)$ of ε given η is strictly monotone in e on the support of ε. Then $V = F_{\varepsilon|\eta}(\varepsilon|\eta)$ is uniformly distributed independently of η and for the inverse $F^{-1}(v, t)$ of $F_{\varepsilon|\eta}(e|t)$ in its first argument, we have $\varepsilon = F^{-1}(V, \eta)$. It follows that

$$g(X, \varepsilon) = g(X, F^{-1}(V, \eta)) = \tilde{g}(X, \eta, V).$$

Thus, even though the true model has just one disturbance, a model with two disturbances, one of which is η, also is correct. However, if ε is the structural disturbance (e.g., unobserved ability in the education example) then V and its quantiles have no useful interpretation. Conversely, if $\tilde{g}(X, \eta, V)$ is structural, then assuming scalar ε is potentially a misspecification. We avoid such interpretation issues when we allow the dimension of ε to be unspecified and unknown. Of course, we also then lose the ability to specify the disturbances, which could be useful when they have interpretations as individual or market heterogeneity.

There are many interesting characteristics of the structural response function g and distributions that we will be able to identify and estimate. One of these is the τ^{th} quantile of the

[2]For scalar X we need scalar η. In a systems generalization we would need η to have the same dimension as X.

distribution of $g(x, \varepsilon)$, over the marginal distribution of ε for a fixed x. We will refer to this as the quantile structural function (QSF), and denote it by $q_Y(\tau, x)$. This object describes how the τ^{th} quantile of the outcome Y with x fixed depends on the value of the endogenous regressor. If ε is a scalar and $g(x, e)$ is increasing in e then $q_Y(\tau, x) = g(x, q_\varepsilon(\tau))$, where $q_\varepsilon(\tau)$ is the τ^{th} quantile of ε, i.e. the QSF is the value of the structural response function at the τ^{th} quantile for ε. The QSF generalizes this object to the case where ε can have arbitrary dimension. The difference $q_Y(\tau, \tilde{x}) - q_Y(\tau, \bar{x})$ is the quantile treatment effect of Lehman (1974) and Abadie, Angrist and Imbens (2002). Distributional objects were also mentioned by Blundell and Powell (2002).

Another object of interest is the average structural function (ASF) of Blundell Powell (2002).

$$\mu(x) = \int g(x, \varepsilon) F_\varepsilon(d\varepsilon). \tag{2.3}$$

As in the Chamberlain (1984) binary choice panel data model, $\mu(x)$ summarizes structural effects for a whole population, whereas quantile effects only apply to a small segment. Also, $\mu(\tilde{x}) - \mu(\bar{x})$ corresponds to the average treatment effect which is the subject of a large literature on program evaluation, (e.g. Heckman, 1990; Imbens, 2003). For instance, in a linear random coefficients model where $Y = \alpha(\varepsilon) + X'\beta(\varepsilon)$, the ASF is $\mathbb{E}[\alpha(\varepsilon)] + x'\mathbb{E}[\beta(\varepsilon)]$, and varying x gives the average structural response to changes in x. See Blundell and Powell (2002), Altonji and Matzkin (2001), and Wooldridge (2002) for further motivation.

An additional object of interest is the average derivative

$$\delta = \mathbb{E}[\partial g(X, \varepsilon)/\partial x].$$

This object is analogous to the average derivative studied in Stoker (1986) and Powell, Stock and Stoker (1989) in the context of exogenous regressors. This parameter summarizes the marginal effect of x on g over the population of X and ε. In a linear random coefficients model $Y = \alpha(\varepsilon) + X'\beta(\varepsilon)$, the average derivative is $\delta = \mathbb{E}[\beta(\varepsilon)]$, equal to the derivative of the ASF. This is not true in general. If the structural function satisfies a single index restriction, with $g(x, \varepsilon) = \tilde{g}(x'\beta_0, \varepsilon)$, then δ will be proportional to β_0.

There are also many specific policies that can be of interest. For instance, consider the effect of imposing an upper limit $\bar{x}$ on the choice variable X in the economic models described above. Then, for a single peaked objective function it follows that the optimal choice will be $\ell(X) = \min\{X, \bar{x}\}$. Assuming there are no general equilibrium effects, the average difference of the outcome with and without the constraint will be $\mathbb{E}[g(\ell(X), \varepsilon) - Y]$. In what follows we will consider identification and estimation of such a policy effect, as well as the objects we have described in the previous paragraphs.

For some policies and effects the structural function $g(x, e)$ itself may be the object of interest. If the dimension of ε is unspecified then it is hard to see how $g(x, e)$ could be obtained. The function $g(x, e)$ can be obtained if the dimension of e is one or two. If ε is a scalar and $g(x, e)$ is monotonic increasing in e then $q_Y(\tau, x) = g(x, q_\varepsilon(\tau))$, as discussed above. Also, if $\varepsilon = (\eta, v)$ for some scalar v that is independent of η, and g is monotonic in v, then the conditional quantile of Y given $X = x$ and $\eta = t$ will be $g(x, t, q_v(\tau))$, where t denotes a possible value of η. This two disturbance case is that considered by Chesher (2003) and Ma and Koenker (2003). Here, identification and estimation results for g will correspond to identification and estimation results for this conditional quantile function. Identification conditions for g in both the one and two disturbance case are given in the next Section.

3 Identification

In this section we give precise identification results. The first assumption we make is that the instrument is independent of the disturbances.

ASSUMPTION 3.1: *(Independence) The disturbances* (ε, η) *are jointly independent of* Z.

Here we require full independence as in Roehrig (1988) and Imbens and Angrist (1994), rather than the weaker mean-independence as in Newey and Powell (1988, 2003), Newey, Powell and Vella (1999) and Darolles, Florens and Renault (2001). In the economic example of Section 2 this assumption could be plausible if the value of the instrument was chosen at a more aggregate level rather than at the level of the agents themselves. State or county level regulations could serve as such instruments, as would natural variation in economic environment conditions, in combination with random location of agents. For independence to be plausible in economic models with optimizing agents it is also important that the relation between the outcome of interest and the regressor, $g(x, \varepsilon)$, is distinct from the objective function that is maximized by the economic agent $(g(x, \varepsilon) - c(x, z)$ in the economic example from the previous section), as pointed out in Athey and Stern (1998). To make the instrument correlated with the endogenous regressor it should enter the latter (e.g., through the cost function), but to make the independence assumption plausible the instrument should not enter the former.

The second assumption requires monotonicity in η of the reduced form for X_1.

ASSUMPTION 3.2: *(Monotonicity of Endogenous Regressor in the Unobserved Component) With probability one* $h(Z, t)$ *is strictly monotone in* t.

This assumption is trivially satisfied if $h(z, t)$ is additive in t, but allows for general forms of non-additive relations. Matzkin (2003) considers nonparametric estimation of $h(z, t)$ under

Assumptions 3.1 and 3.2 in a single equation exogenous regressor framework and Pinkse (2000b) gives a multivariate version. Das (2001) uses a stochastic version of this assumption to identify parameters in single index models with a single endogenous regressor.

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; Athey, 2002). For example, in the setting of the educational production function in Section 2, assume that $g(x, e)$ is twice continuously differentiable. Suppose that: (i) The educational production function is strictly increasing in ability e and education x; (ii) the marginal return to formal education is strictly increasing in ability and decreasing in education, so that $\partial g/\partial e > 0$, $\partial g/\partial x > 0$, $\partial^2 g/\partial x \partial e > 0$, and $\partial^2 g/\partial x^2 < 0$(this would be implied by a Cobb-Douglas production function); (iii) the cost function and the marginal cost are increasing in education, so that $\partial c/\partial x > 0$, $\partial^2 c/\partial x^2 > 0$ and (iv) the signal η and ability ε are affiliated. Under those conditions the decision rule $h(Z, t)$ is monotone in the signal t.[3]

Under these two condition the reduced form disturbance η is identified up to a one-to-one transformation. Let $F_\eta(t)$ denote the CDF of η.

THEOREM 1: *(Identification of η) If Assumptions 3.1 and 3.2 are satisfied then* $F_{X_1|Z}(X_1|Z) = F_\eta(\eta)$.

All of our results are proved in the Appendices. This result states that the one-to-one function $\bar{\eta} = F_\eta(\eta)$ of η is identified.

An important function of the data is the conditional quantile function $Q_{Y|X\bar{\eta}}(\tau, X, \bar{\eta})$ of Y given X and $\bar{\eta} = F_{X_1|Z}(X_1|Z)$, hereafter denoted as CQF. All of the structural objects we have considered can be obtained from the CQF, under certain additional conditions. Intuitively, after conditioning on $\bar{\eta}$, variation in X_1 in the data is statistically independent of variation in ε, so that we can separate out the effect of X_1 on $g(X, \varepsilon)$ from the effect of ε, leading to identification of structural effects. This approach is essentially a nonparametric generalization of the control function approach (e.g., Heckman and Robb, 1984; Newey, Powell and Vella, 1999; Blundell and Powell, 2000), with $\bar{\eta} = F_{X_1|Z}(X_1|Z)$ being a generalized control function.

In order for this control function approach to identify structural effects, there must be sufficient variation in X_1 after conditioning on Z_1 and $\bar{\eta}$. This requires that X_1 vary with Z_2, which is like the rank condition in linear simultaneous equations models. For identification of many effects, including average derivatives and limit policies, a little variation in X_1 with Z_2 will suffice, as discussed below. For identification of the ASF and QSF, we will need a strong

[3] Of course in this case one may wish to exploit these restrictions on the production function, as in, for example, Matzkin, 1993.

condition, that the support of $\bar{\eta} = F_{X_1|Z}(X_1|Z_1, Z_2)$ conditional on $X = (X_1, Z_1)$ is equal to the marginal support of $\bar{\eta}$. Using the form of $\bar{\eta}$, a precise statement of this condition is:

ASSUMPTION 3.3: *(Full Instrument Effect) For all (x, z_1) in some set that has probability one, the support of $F_{X_1|Z}(x|z_1, Z_2)$ equals the support of $\bar{\eta}$.*

This condition requires that the excluded instrument Z_2 must move $F_{X_1|Z}(x|z_1, Z_2)$ over the whole range of its possible values for every fixed value of x and z_1. In addition to a rank condition, with Z_2 affecting the distribution of X_1, it entails a full range condition, where Z_2 must cause variation over the whole support of $\bar{\eta}$.

An example may help clarify Assumption 3.3. Suppose $z = z_2$ is a scalar and that the reduced form is $X = \delta Z + \eta$, where η is continuously distributed with CDF $G(u)$. Then

$$F_{X|Z}(x|z) = G(x - \delta z).$$

Assume that the support of $F_{X|Z}(X|Z)$ is $[0, 1]$. Then a necessary condition for Assumption 3.3 is that $\delta \neq 0$, because otherwise $F_{X|Z}(x|Z)$ would be a constant. Together $\delta \neq 0$ and the support of Z being the entire real line are sufficient. This example illustrates that Assumption 3.3 embodies two types of conditions, one being a generalization of a rank condition and the other being a full support condition for the instrument.

To show identification of the QSF and ASF we give explicit formulae in terms of identified objects. These formulae can then be used to construct estimators. For the QSF we have

$$q_Y(\tau, x) = G^{-1}(\tau, x), \tag{3.4}$$

where

$$\begin{aligned} G(y, x) &\stackrel{def}{=} \Pr(g(x, \varepsilon) \leq y) = \int \Pr(g(x, \varepsilon) \leq y|\eta) F_\eta(d\eta) \\ &= \int \Pr(Y \leq y|X = x, \eta) F_\eta(d\eta) = \int Q_Y^{-1}(y, x, \eta) F_\eta(d\eta), \end{aligned} \tag{3.5}$$

where the second equality follows by iterated expectations, the third by independence of Z and η and by the fact that conditional on η, X is function of Z. Thus the QSF is the inverse of the integral over the marginal distribution of η of the inverse of the CQF, i.e. of the conditional CDF of Y given X and η.

For continuous Y the ASF is identified from the QSF as $\mu(x) = \int_0^1 q_Y(\tau, x) d\tau$. Alternatively, using the other representation of the ASF in (2.3) and following Blundell and Powell (2002),

$$\begin{aligned} \mu(x) &= \int g(x, \varepsilon) F_\varepsilon(d\varepsilon) = \int \mathbb{E}[g(x, \varepsilon)|\eta] F_\eta(d\eta) \\ &= \int \mathbb{E}[g(X, \varepsilon)|X = x, \eta] F_\eta(d\eta) = \int \mathbb{E}[Y|X = x, \eta] F_\eta(d\eta). \end{aligned} \tag{3.6}$$

Thus, the ASF is the integral over the marginal distribution of η of $\mathbb{E}[Y|X,\eta]$.

The following theorem gives a precise statement of these results.

THEOREM 2: *(Identification of the QSF and ASF) If Assumptions 3.1, 3.2, and 3.3 hold then the QSF is identified. If, in addition,* $\mathbb{E}[|Y|]$ *and* $\int\int \mathbb{E}[|Y||X,\eta]F_X(dX)F_\eta(d\eta)$ *are finite then the ASF is identified.*

As mentioned in Section 2, identification of the QSF also implies identification of the structural function in the one disturbance case.

COROLLARY 3: *If* ε *is a scalar,* $g(x,\varepsilon)$ *is strictly monotonic increasing in* ε*, and Assumptions 3.1, 3.2, and 3.3 are satisfied then* $g(x, q_\varepsilon(\tau))$ *is identified for every* $0 < \tau < 1$.

The role of Assumption 3.3 is to allow us to integrate over the marginal distribution of $\bar{\eta}$ holding X fixed as in the previous formulae. If Assumption 3.3 holds for particular values of X_1 and Z_1 we may be able to identify the QSF and ASF at particular values. If Assumption 3.3 is not satisfied, we may be able to derive bounds on the QSF, and on the ASF if Y is bounded, using the approach by Manski (1990, 1995). Also, Assumption 3.3 is not required to hold for identification of other objects, including the average derivative and the limit policy.

For the average derivative parameter δ we require that X is continuously distributed given η, so that the derivative of $\beta(x,\eta) = \mathbb{E}[Y|X=x,\eta]$ with respect to x can be a well defined object. Note that by independence of Z and η,

$$\beta(X,\eta) = \int g(X,\varepsilon)F(d\varepsilon|X,\eta) = \int g(X,\varepsilon)F(d\varepsilon|\eta), \tag{3.7}$$

which is identified on the support of (X,η) by the same argument for the identification of the CQF. Assuming that we can differentiate under the integral, it then follows that $\partial\beta(X,\eta)/\partial x = \int g_x(X,\varepsilon)F(d\varepsilon|\eta)$. Then

$$\delta = \mathbb{E}[g_x(X,\varepsilon)] = \mathbb{E}\left[\int g_x(X,\varepsilon)F(d\varepsilon|X,\eta)\right] = \mathbb{E}\left[\int g_x(X,\varepsilon)F(d\varepsilon|\eta)\right] = \mathbb{E}\left[\frac{\partial}{\partial x}\beta(X,\eta)\right], \tag{3.8}$$

where the third equality follows by independence of Z and η.

For the limit policy, we require that the support of $X = h(Z,\eta)$ given η include $\ell(X) = \min(X,\bar{x})$. For this it is sufficient that $\bar{x}$ is in the support of $h(Z,\eta)$ given η for all η. Then

$$\mathbb{E}[g(\ell(X),Z_1,\varepsilon)] = \mathbb{E}[\mathbb{E}[g(\ell(X),Z_1,\varepsilon)|X,\eta]] = \mathbb{E}\left[\int g(\ell(X),Z_1,\varepsilon)F(d\varepsilon|\eta)\right] = \mathbb{E}[\beta(\ell(X),Z_1,\eta)]. \tag{3.9}$$

In neither of these examples is it necessary that the conditional support of X given η contains all possible values of X_1 in order to achieve identification. We make this discussion precise in the following result:

THEOREM 4: *(Identification of the average derivative and limit policy). Suppose that Assumptions 3.1 and 3.2 are satisfied. If (i) X_1 has a continuous conditional distribution given η, (ii) with probability one $g(x, \varepsilon)$ is continuously differentiable in x at $x = X$; (iii) for all x and some $\Delta > 0$, $\mathbb{E}[\int \sup_{\|x-X\|\leq\Delta} \|g_x(x, \varepsilon)\| F(d\varepsilon|\eta)]$ exists, then $\delta = \mathbb{E}[g_x(X, \varepsilon)]$ is identified. If with probability one there is some point in the support of X_1 given η that is less than or equal to $\bar{x}$ then $\mathbb{E}[g(\ell(X_1), Z_1, \varepsilon)]$ is identified.*

Analogous identification results can be formulated for expectations of other linear transformations of $g(x, \varepsilon)$. Let $h(x)$ denote a function of x and $T(h(\cdot), x)$ be a transformation that is linear in h. Then, by linearity of T the order of integration and transformation can be interchanged to obtain, from equation (3.7),

$$T(\beta(\cdot, t), x) = T(\int g(\cdot, \varepsilon) F_{\varepsilon|\bar{\eta}}(d\varepsilon|t), x) = \int T(g(\cdot, \varepsilon), x) F_{\varepsilon|\bar{\eta}}(d\varepsilon|t) = \int T(g(\cdot, \varepsilon), x) F_{\varepsilon|X,\bar{\eta}}(d\varepsilon|x, t).$$

Taking expectations of both sides we find that

$$\mathbb{E}[T(\beta(\cdot, \bar{\eta}), X)] = \mathbb{E}[E[T(g(\cdot, \varepsilon), X)|X, \bar{\eta}]] = \mathbb{E}[T(\beta(\cdot, \bar{\eta}), X)].$$

This formula leads to the following general identification result:

THEOREM 5: *(Identification of expectations of linear functions): Suppose that Assumptions 3.1 and 3.2 are satisfied, that $T(\beta(\cdot, \eta), X)$ is a well defined random variable, $\mathbb{E}[T(\beta(\cdot, \eta), X)]$ exists, and $T(\int g(\cdot, \varepsilon) F_{\varepsilon|\bar{\eta}}(d\varepsilon|\eta), X) = \int T(g(\cdot, \varepsilon), X) F_{\varepsilon|\bar{\eta}}(d\varepsilon|\eta)$. Then $\mathbb{E}[T(g(\cdot, \varepsilon), X)]$ is identified.*

Theorem 4 is a special case of this result with $T(h(\cdot), x) = \partial h(x)/\partial x$ and $T(h(\cdot), x) = h(\ell(x), z_1)$. Additional examples include integrals of g over elements of x, expectations of g evaluated at functions of x other than $\ell(x)$, and other linear functionals of g.

In addition to helping identify other structural objects, the CQF can be of interest in its own right, being equal to the structural function in the two disturbance case mentioned in Section 2.

COROLLARY 6: *If $\varepsilon = (\eta, v)$, v is independent of η, $g(x, \eta, v)$ is strictly monotonic in v, and Assumptions 3.1 and 3.2 are satisfied, then $Q_{Y|X\bar{\eta}}(\tau, X, \eta) = g(X, \eta, q_v(\tau))$, where $q_v(\tau)$ is the τ^{th} quantile of v.*

This result shows that the function $g(X, \eta, q_v(\tau))$ is identified on the support of (X, η). However, identification of this function does not mean that structural effects of changing x while holding η and v fixed, like the partial derivative with respect to x, are identified. For this purpose it is necessary that the support of (X, η) include subsets where x varies while holding z_1 and η fixed. This would entail that Z_2 affects X_1, which again is a kind of "rank condition."

Corollary 6 was given in an early version of this paper.[4] In independent work Chesher (2001, 2003) gave an identification result for the structural effect $\partial g(x, t, q_v(\tau))/\partial x$ at a point. Neither result implies the other, since we do not assume g is differentiable. The approaches to identification are related though. We identify g using a generalized control function. Chesher (2003) identifies $\partial g/\partial x$ with a certain function of quantile derivatives. In Appendix A we show that for $X_1 = X$, Z scalar, and $(X, F_{X|Z}(X|Z))$ a one-to-one function of (X, Z), that his function of quantile derivatives is $\partial Q_{Y|X\bar{\eta}}(\tau, X, \eta))/\partial x$, *without Assumptions 3.1, 3.2, or equations (2.1), (2.2) being satisfied.* Thus, the generalized control function method and Chesher's (2003) approach lead to the same conditional quantile derivative, even when the model is incorrect. He uses it to identify $\partial g/\partial x$ in the two disturbance model. We use the control function to identify g without requiring differentiability in the two disturbance model (in Corollary 6) and also to identify structural effects when ε has one or arbitrary dimension (in the other results of this Section).

A parametric version of the CQF may be of some interest. One version is a second stage linear quantile regression of Y on (X, U), where U is the residual from a first stage linear quantile regression of X on Z. When Z is scalar and the quantile regression of Y on (X, Z) and X on Z are linear then the coefficient of X in the second stage regression is Chesher's quantile derivative, as shown in Appendix A. Here U acts as a control function, which is an additional regressors, making this quantile estimation different than that of Amemiya (1982), because it includes an additional right-hand side variable. By including additional terms in the second stage regression, one could also estimate nonlinear models, such as the location scale model of Koenker and Ma (2002). Further consideration of this control function approach to controlling for endogeneity in parametric quantile regression is beyond the scope of this paper.

4 Estimation

We follow a multistep approach to estimation from i.i.d. data $(Y_i, X_{1i}, Z_i), (i = 1, ..., n)$. The first step is estimation of the control function $\bar{\eta} = F_{X_1|Z}(X_1|Z)$. From an estimator $\hat{F}_{X_1|Z}(x|z)$

[4]The two disturbance model and the result of Corollary 6 were formulated and presented in the spring of 2001.

we can constructs estimates of the control function observations as

$$\hat{\eta}_i = \hat{F}_{X_1|Z}(X_{1i}|Z_i).$$

These estimates can then be used to construct an estimator $\hat{F}_{Y|X,\eta}(y|x,t)$ of the conditional distribution function of Y given X and η or an estimator $\hat{\beta}(x,t)$ of the conditional mean. Estimators of objects of interest can then be formed by plugging these estimators into the formulae of Section 3. Assuming that η is continuously distributed (so that it is $U(0,1)$ by $\eta = F_\eta(\eta)$), estimators of the QSF and ASF are given by

$$\begin{aligned} \hat{q}_Y(\tau,x) &= \hat{G}^{-1}(y,x);\ \hat{G}(y,x) = \int_0^1 \hat{F}_{Y|X,\eta}(y|x,t)dt \\ \hat{\mu}(x) &= \int_0^1 \hat{\beta}(x,t)dt. \end{aligned}$$

Alternatively, the integrals over t could be estimated by simulation. Estimators of the average derivative and the upper limit effect can be constructed by plugging in the formulae and replacing the expectation over (X,η) with a sample average, as in

$$\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\hat{\beta}(X_i,\hat{\eta}_i)}{\partial x},\ \widehat{E[g(\ell(X),\varepsilon)-Y]} = \sum_{i=1}^{n}[\hat{\beta}(\ell(X_i),\hat{\eta}_i) - Y_i]/n.$$

We consider a power series estimator of $\beta(x,t)$. We use power series for the second step (e.g. rather than kernels) because theory for them can allow densities to go to zero on boundaries, which is an important feature of this model. We use them for the first step because they lead to simple conditions. Almost any other series estimator could be used for the first step also, including splines and Fourier series.

To describe the first step estimation of $\bar{\eta} = F_{X_1|Z}(X_1|Z)$, suppose that z is $r \times 1$ and let $q_\ell(z) = \Pi_{j=1}^r (z_j)^{\lambda_\ell(j)}, (\ell = 1, 2, ...)$, denote multivariate power series terms, where $\lambda_\ell(j) \geq 0$, the order of the term is $\sum_{j=1}^r \lambda_\ell(j)$, and we assume that the order is increasing in L and that all terms of given order are included before increasing the order. Examples include power series or spline functions. Also, let $q^L(z) = (q_1(z), ..., q_L(z))'$ and $\hat{M} = \sum_{i=1}^n q^L(Z_i)q^L(Z_i)'/n$. 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 $X_{1i} \leq x$ on functions of Z_i. It has the form $\tilde{\eta}_i = \tilde{F}(X_{1i} \mid Z_i)$, where

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

where A^- denotes any generalized inverse of the matrix A. As is well known, the predicted values $\tilde{F}(X_{1i}|Z_i)$ will be invariant to the choice of generalized inverse, which is important here because we will allow for $\hat{M}$ to be singular, even asymptotically. One feature of this estimator is that it is not necessarily bounded between 0 and 1. We impose that restriction 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

$$\hat{\eta}_i = \tau(\tilde{F}(X_{1i}|Z_i)).$$

To describe the power series estimator in the second step, let $w = (x, t)$ denote the entire $s \times 1$ vector of regressors in $\beta(X, \bar{\eta}) = \mathbb{E}[Y|X, \bar{\eta}]$. Let $p_k(w) = \Pi_{j=1}^{s}(w_j)^{\lambda(k,j)}$, $(k = 1, 2...)$, be multivariate power series with order increasing in K and all terms of given order included before increasing the order. Let $p^K(w) = (p_1(w), ..., p_K(w))'$, $\hat{w}_i = (X_i, \hat{\eta}_i)$, and $\xi(x)$ be a trimming function of x that is either zero or one, and $\hat{P} = \sum_{i=1}^{n} \xi(X_{1i})p^K(\hat{w}_i)p^K(\hat{w}_i)'/n$. A nonparametric estimator of the ACR $\beta(w) = \mathbb{E}[Y|W = w]$ is then

$$\hat{\beta}(w) = p^K(w)'\hat{\gamma}; \hat{\gamma} = \hat{P}^{-1}\sum_{j=1}^{n} \xi(X_{1j})p^K(\hat{w}_j)Y_j/n.$$

This estimator can be used as described above to estimate the ASF, average derivative, or input limit response. The trimming function is present to keep X_{1i} away from the boundary of its support which prevents the joint density of X_{1i} and $\bar{\eta}_i$ from becoming unbounded. We could let $\xi(x)$ depend on data, but for simplicity do not.

An important issue for practice is the choice of number of terms. We have tried cross-validation on the second stage as a method for this.

5 Convergence Rates for the ASF

To derive large sample properties of the estimator it is essential to impose some conditions. The first assumption imposes smoothness conditions on the control function $F_{X_1|Z}(x|z)$. Let $\mathcal{X}$ and $\mathcal{Z}$ denote the support of X_{1i} and Z_i, respectively.

ASSUMPTION 5.1: *Z_i has compact support and $F_{X|Z}(x|z)$ is continuously differentiable of order d_1 on the support with derivatives uniformly bounded in x and z.*

This condition implies an approximation rate of $K^{-d_1/r}$ for the CDF that is uniform in both its arguments; see Lorentz (1986). The following result gives a convergence rate for the first step:

LEMMA 7: *If Assumption 5.1 is satisfied,*

$$\mathbb{E}\left[\sum_{i=1}^{n}(\hat{\eta}_i - \bar{\eta}_i)^2/n\right] = O(L/n + L^{1-2d_1/r}).$$

The two terms in this rate result are variance (L/n) and squared bias $(L^{1-2d_1/r})$ terms respectively. In comparison with previous results for series estimators, this convergence result has $L^{1-2d_1/r}$ for the squared bias term as a rate rather than $L^{-2d_1/r}$. The extra L arises from the predicted values $\hat{\eta}_i$ being based on regressions with the dependent variables varying over the observations.

To obtain a convergence rate for the average structural function, it is necessary to account for the fact that the joint pdf of $(X, \bar{\eta})$ behaves differently in the tails that does the marginal density of $\bar{\eta}$, which is uniform. In the linear reduced form case $X = Z + \eta$ it is easy to see why this is so. For $f_Z(z)$ and $f_\eta(t)$ the marginal pdf's of Z and η, the joint pdf of $(X, \bar{\eta})$ is

$$f_{X,\bar{\eta}}(x,t) = f_Z(x - F_\eta^{-1}(t)), 0 < t < 1.$$

This allows for x to be unbounded, but this feature is accounted for by the trimming function $\xi(x_1)$. The more difficult problem is that as t goes to 0 or 1 holding x fixed, the joint pdf goes to zero. Thus, in the ASF estimator $\hat{\mu}(x) = \int_0^1 \hat{\beta}(x,t)dt$, the tails of the $\bar{\eta}$ distribution are being upweighted relative to the joint distribution. Consequently, we expect that, unlike other settings, integration of $E[Y|X, \bar{\eta}]$ over the marginal distribution of $\bar{\eta}$ will not increase the convergence rate. Indeed, with the upweighting one might expect that the convergence rate would involve a uniform rate for the $\bar{\eta}$ argument.

To obtain convergence rates for series estimators it is necessary to restrict the rate at which the density goes to zero as $\bar{\eta}$ approaches zero or one. The next condition fulfills this purpose:

ASSUMPTION 5.2: *$\mathcal{X}$ is a Cartesian product of compact intervals, $p^K(w) = p^{K_x}(x) \otimes p^{K_\eta}(t)$, and there exist constants $C, \alpha > 0$ such that*

$$\inf_{x_1 \in \mathcal{X}} f_{X,\bar{\eta}}(x,t) \geq C[t(1-t)]^{\alpha}.$$

This condition will be satisfied with a linear, Gaussian reduced form with an α that is related to the R^2, as shown in the following result.

LEMMA 8: *If $X = Z + \eta$ where Z and η are normally distributed and independent, then for $R^2 = Var(Z)/[Var(X)]$ and $\bar{\alpha} = (1 - R^2)/R^2$, for any $B, \delta > 0$ there exists C such that for all $|x| \leq B$, $t \in [0,1]$,*

$$C[t(1-t)]^{\bar{\alpha}-\delta} \geq f_{X,\bar{\eta}}(x,t) \geq C^{-1}[t(1-t)]^{\bar{\alpha}+\delta}$$

Here we see that the rate at which the joint density vanishes increases as the R^2 falls. We conjecture that R^2 will also affect the attainable convergence rate for the ASF. The faster the density vanishes the harder it will be to estimate $\beta(w)$ in the tails for $\bar{\eta}$. We will see below that lower R^2 does slow down the rate of convergence of the series estimator.

The next condition imposes smoothness of $\beta(w)$, in order to obtain an approximation rate for the second step.

Assumption 5.3: *$\beta(w)$ is continuously differentiable of order d* on $\mathcal{X} \times [0,1]$.

We also bound the conditional variance of Y, as is often done for series estimators

[illegible]

With these conditions in place we can obtain a convergence rate bound for the second-step estimator.

Theorem 9: *If Assumptions 5.1 - 5.4 are satisfied and $K^2 K_\eta^{\alpha+2}(L/n + L^{1-2d_1/r}) \to 0$ then*

$$\int \left[\hat{\beta}(w) - \beta(w)\right]^2 dF(w) = O_p(K/n + K^{-2d/s} + L/n + L^{1-2d_1/r})$$

$$\sup_{w \in W} |\hat{\beta}(w) - \beta(w)| = O_p(K_\eta^\alpha K[K/n + K^{-2d/s} + L/n + L^{1-2d_1/r}]^{1/2}).$$

The choice of L minimizing this expression is proportional to $n^{1/(2d+s-1)}$ and $n^{r/2d_1}$ respectively. For this choice of K_η and L the rate hypothesis and the convergence rate are given by

$$
\begin{aligned}
n^{[(1+\alpha+s)/d]+(r/2d_1)-1} &\longrightarrow 0, \\
\int [\hat{\mu}(x) - \mu(x)]^2 F_X(dx) &= O_p(n^{-2(d-\alpha-1)/(2d+s-1)} + n^{[(1+\alpha)/d]+(r/2d_1)-1}).
\end{aligned}
$$

The rate hypothesis means that $[(1 + \alpha + s)/d] + (r/2d_1) < 1$, which requires that $\beta(w)$ have more than $1 + \alpha + s$ derivatives and that $F_{X_1|Z}(x_1|z)$ have more than $r/2$ derivatives. In the Gaussian example of Lemma 8, $d - s > 1 + \alpha$ is equivalent to $d - s > 1/R^2$. In microeconometric applications, where reduced form r-squareds are low, this condition would require many derivatives to exist.

We emphasize that these results only provide a bound on the convergence rate of the estimators. Some improvements may be possible, although currently it seems not to be known whether series estimators can attain optimal uniform rates, see deJong (2003). It would be good to know what the best attainable rate is, but that does not yet seem to be known when the density goes to zero at the boundary, and derivation of such rates is beyond the scope of this paper.

For comparison purposes consider the additive disturbance model

$$Y = g(X) + \varepsilon,\ X = Z + \eta, Z \sim N(0, 1), \eta \sim N(0, (1 - R^2)/R^2),$$

where X, Z, ε, and η are scalars and we normalize $E[\varepsilon] = 0$. Here the ASF is $g(X)$. Under regularity conditions like those give above the estimator will converge at a rate that is a power of n, but slower than the optimal one-dimensional rate. In contrast, the estimator of Newey, Powell, and Vella (1999), which imposes additivity, does converge at the optimal one-dimensional rate. Also, the estimator of Severini and Tripathi (2003), which only uses the conditional mean restriction $E[\varepsilon|Z] = 0$, converges at a rate that is slower than any power of n. Thus, the convergence rate we have obtained here is intermediate between that indicate that estimators of $g(X)$ would converge at a rate that is slower than any power of n, due to an ill-posed inverse problem. Thus, our estimator has rates of convergence that are intermediate between those of a control function estimator where additivity is imposed and those of an estimator of the conditional mean model.

6 Extensions

It is also possible to show asymptotic normality of the ASF estimator at a point x and to show root-n consistency and asymptotic normality of the average derivative and estimators of other policy parameters. These results are important because they allow us to do inference and show that the slow convergence rate of the ASF does not carry over to other important estimators. By averaging over the joint distribution of X and $\bar{\eta}$ we avoid the upweighting problem associated with the ASF. In future work we will also give a Monte Carlo example.

Appendix

A Relationship to Chesher (2003)

We consider the relationship to Chesher (2003) when $z = z_2$ is a scalar, using his notation. Let $Q_{Y|X\bar{\eta}}(\tau, x, t)$, $Q_{Y|XZ}(\tau, x, z)$, and $Q_{X|Z}(t, z)$ be conditional quantile functions of Y given X and $\bar{\eta} = F_{X|Z}(X|Z)$, Y given X and Z, and of X given Z, respectively. Also let ∇_a denote a partial derivative with respect to a variable a. The functional used by Chesher (2003) for identification of structural effects is

$$\pi(\tau, \eta, x) = \nabla_x Q_{Y|XZ}(\tau, x, z) + \frac{\nabla_z Q_{Y|XZ}(\tau, x, z)}{\nabla_z Q_{X|Z}(\eta, z)},$$

assuming that the expression on the right exists.

THEOREM A1: *If $Q_{Y|XZ}(\tau, x, z)$ and $F_{X|Z}(x|z)$ are continuously differentiable in a neighborhood of (x_0, z_0), $\nabla_z F_{X|Z}(x_0|z_0) \neq 0$, $\nabla_x F_{X|Z}(x_0|z_0) \neq 0$, and $(X, F_{X|Z}(X|Z))$ is a one-to-one transformation of (X, Z), then for $\eta = F_{X|Z}(X|Z)$ and $\eta_0 = F_{X|Z}(x_0|z_0)$,*

$$\pi(\tau_0, \eta_0, x_0) = \nabla_x Q_{Y|X\bar{\eta}}(\tau, x_0, \eta_0).$$

Thus we find that Chesher's π is equal to the derivative of the CQF with respect to x, when $(X, F_{X|Z}(X|Z))$ is a one-to-one transformation of (X, Z) and the derivative conditions of this result are satisfied. This result holds without equations (2.1) or (2.3) being satisfied and without any independence assumption. Thus, in a specification robust sense, our generalized control function approach and Chesher's approach lead to the same functional of the data.

A linear quantile model provides further insight. Suppose that the conditional quantile functions are linear, with

$$Q_{Y|XZ}(\tau, X, Z) = \beta(\tau)X + \gamma(\tau)Z + \alpha_1(\tau), Q_{X|Z}(\eta, Z) = \gamma(\eta)Z + \alpha_2(\eta).$$

Then $\pi = \beta(\tau) + [\gamma(\tau)/\gamma(\eta)]$. Define $U = X - \gamma(\eta)Z - \alpha_2(\eta)$, the residual from the quantile regression for X. If $\gamma(\eta) \neq 0$ then (X, U) is a one-to-one transformation of (X, Z), so that

$$\begin{aligned} Q_{Y|XU}(\tau, X, U) &= Q_{Y|XZ}(\tau, X, Z) = Q_{Y|XZ}(\tau, X, [X - U - \alpha_2(\eta)]/\gamma(\eta)) \\ &= \beta(\tau)X + \gamma(\tau)[X - U - \alpha_2(\eta)]/\gamma(\eta) + \alpha_1(\eta) \\ &= \pi X + [-\gamma(\tau)/\gamma(\eta)]U + \alpha_1(\eta) - \alpha_2(\eta)\gamma(\tau)/\gamma(\eta). \end{aligned}$$

Thus, the quantile regression of Y on X and U is also linear, and the coefficient of X is π, so that in a linear quantile model π can also be interpreted as coming from a control function approach, where the residual U is the control function.

In both the nonparametric and linear models, the relationship of the Chesher's (2003) approach to the corresponding control function approach is analogous to the relationship of indirect least squares to two-stage least squares. Like indirect least squares, π is obtained by combining coefficients from two regressions, while like two-stage least squares, the control function obtains the coefficient from a second stage regression. Thus, one might refer to the Chesher (2003) approach as an *indirect control function* method. This indirect method avoids inference difficulties associated with two-step methods and only requires that $F_{X|Z}(x,z)$ be one-to-one locally, but is awkward to use when averaging over the control function.

PROOF OF THEOREM A1: Let $(X, k(\eta, X))$ denote the inverse of $(X, F_{X|Z}(X|Z))$, so that $Z = k(\eta, X)$. It then follows by (X, η) and (X, Z) being one-to-one transformations of each other that

$$Q_{Y|X\eta}(\tau, X, \eta) = Q_{Y|XZ}(\tau, X, Z) = Q_{Y|XZ}(\tau, X, k(\eta, X)).$$

Also, by the inverse function theorem, $Q_{X|Z}(\eta, z)$ is differentiable at (η_0, z_0) and $k(\eta, x)$ is differentiable at (η_0, x_0) with

$$\nabla_x k(\eta_0, x_0) = -\nabla_x F_{X|Z}(x_0, z_0)/\nabla_z F_{X|Z}(x_0, z_0) = 1/\nabla_z Q_{X|Z}(\eta_0, z_0).$$

Then by the chain rule

$$\begin{aligned}
\nabla_x Q_{Y|X\eta}(\tau, x_0, \eta_0) &= \frac{\partial}{\partial x} Q_{Y|XZ}(\tau, x, k(\eta_0, x))|_{x=x_0} \\
&= \nabla_x Q_{Y|XZ}(\tau, x_0, z_0) + \nabla_z Q_{Y|XZ}(\tau, x_0, z_0)\nabla_x k(\eta_0, x_0) \\
&= \nabla_x Q_{Y|XZ}(\tau, x_0, z_0) + \nabla_z Q_{Y|XZ}(\tau, x_0, z_0)/\nabla_z Q_{X|Z}(\eta_0, z_0). Q.E.D.
\end{aligned}$$

B Proofs of Identification

PROOF OF THEOREM 1: Let $h^{-1}(z, x)$ denote the inverse function for $h(z, \eta)$ in its second argument, which exists by Assumption 3.2. Then,

$$\begin{aligned}
F_{X_1|Z}(x|z) &= Pr(X_1 \leq x|Z = z) = Pr(h(z, \eta) \leq x|Z = z) = Pr(\eta \leq h^{-1}(z, x)|Z = z) \\
&= Pr(\eta \leq h^{-1}(z, x)) = F_\eta(h^{-1}(z, x)),
\end{aligned}$$

where the third equality follows by Assumption 3.2 and the fourth by Assumption 3.1. The conclusion then follows by Assumption 3.1, which implies $\eta = h^{-1}(Z, X_1)$. Q.E.D.

Proof of Theorem 2: By Theorem 1, η is identified. By Assumption 3.3 the support of η conditional on $X = x$ is the support of $F(x|z_1, Z_2)$ and is equal to the support of η. Therefore, $\Pr(g(x, \varepsilon) \leq y|X = x, \eta)$ is well defined on the Cartesian product of the support for X and η, so that $\int \Pr(Y \leq y|X, \eta)F_\eta(d\eta)$ is well defined. Also, conditional on η, $X_1 = (h(Z, \eta), Z_1)$ is a function of only Z, so that X_1 and ε are independent. Then the conclusion for the QSF follows from in eq. (3.5), which gives the QSF as an explicit functional of the identified object $\Pr(Y \leq y|X = x, \eta)$. For the ASF, it follows similarly that, by $\mathbb{E}[|Y|]$ finite, both $\mathbb{E}[|Y||X, \eta]$ and $\mathbb{E}[Y|X, \eta]$ are well defined on the Cartesian product of the support of X and η, and $|\mathbb{E}[Y|X, \eta]| \leq \mathbb{E}[|Y||X, \eta]$. Then by the Fubini theorem and the integrability hypothesis in the statement of the Theorem it follows that $\int \mathbb{E}[Y|X, \eta]F_\eta(d\eta)$ exists with probability one. It then follows from equation (3.7) that the ASF is an explicit function of the data distribution, so that it is identified. Q.E.D.

Proof of Corollary 3: By $g(x, \varepsilon)$ strictly monotonic increasing it follows, as noted in the text that for each x the τ^{th} quantile of $g(x, \varepsilon)$ is $g(x, q_\varepsilon(\tau))$, so that the conclusion follows from Theorem 3. Q.E.D.

Proof of Theorem 4: By the fact that $g(x, \varepsilon)$ continuously differentiable and the integrability condition, it follows that $\int g(x, \varepsilon)F(d\varepsilon|\eta)$ is differentiable with probability one at $x = X$ and $\partial \int g(x, \varepsilon)F(d\varepsilon|\eta)/\partial x|_{x=X} = \int g_x(X, \varepsilon)F(d\varepsilon|\eta)$. Then by equation (3.7) it follows that $\mathbb{E}[Y|X = x, \eta]$ is differentiable at $x = X$, and by equation (3.8) that the average derivative is an explicit functional of the data distribution, an so is identified. For the limit policy, by the assumption about $\bar{x}$ it follows that $\beta(x, t)$ is well defined, with probability one, at $(x, t) = (\ell(X), \eta)$, so that the conclusion follows as in equation (3.9) Q.E.D.

Proof of Theorem 5: By independence of (ε, η) and Z we have $\beta(X, \eta) = E[Y|X, \eta] = \int g(X, \varepsilon)F(d\varepsilon|\eta)$. Then by $T(\int g(\cdot, \varepsilon)F(d\varepsilon|\eta), X) = \int T(g(\cdot, \varepsilon), X)F(d\varepsilon|\eta)$ and iterated expectations,

$$\begin{aligned}\mathbb{E}[T(\beta(\cdot, \eta), X)] &= \mathbb{E}[T(\int g(\cdot, \varepsilon)F(d\varepsilon|\eta), X)] = \mathbb{E}[\int T(g(\cdot, \varepsilon), X)F(d\varepsilon|\eta)] \\ &= \mathbb{E}[\mathbb{E}[T(g(\cdot, \varepsilon), X)|\eta, Z]] = \mathbb{E}[T(g(\cdot, \varepsilon), X)].\end{aligned}$$

Since $\mathbb{E}[T(g(\cdot, \varepsilon), X)]$ is equal to an explicit function of the data distribution, it is identified. Q.E.D.

Proof of Corollary 6: Since η is identified and we can normalize $\eta = \bar{\eta}$ by Theorem 1, it follows that the τ^{th} conditional quantile $Q_{Y|X\bar{\eta}}(\tau, X, \eta)$ of Y given (X, η) is identified. By $Y = g(X, \eta, v)$ and $g(X, \eta, v)$ strictly monotonic in v and by invariance of quantiles to monotonic transformations, it follows that the τ^{th} conditional quantile of $g(X, \eta, v)$ is $g(X, \eta, q_v(\tau))$. Q.E.D.

C Proofs of Consistency

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. Also, let CM refer to the following well known result: *If* $\mathbb{E}[|Y_n||Z_n] = O_p(r_n)$ *then* $Y_n = O_p(r_n)$.

Before proving Theorem 4, we prove a some preliminary results. Let $q_i = q^L(Z_i)$, $v_{ij} = 1(X_{1j} \le X_{1i}) - F_{X_1|Z}(X_{1i}|Z_j)$.

LEMMA C1: *For* $Z = (Z_1, ..., Z_n)$ *and* $L \times 1$ *vectors of functions* $b_i(Z)$, $(i = 1, ..., n)$, *if* $\sum_{i=1}^n b_i(Z)'\hat{Q}b_i(Z)/n = O_p(r_n)$ *then*

$$\sum_{i=1}^n \{b_i(Z)' \sum_{j=1}^n q_j v_{ij}/\sqrt{n}\}^2/n = O_p(r_n).$$

Proof: Note that $|v_{ij}| \le 1$. Consider $j \neq k$ and suppose without loss of generality that $j \neq i$ (otherwise reverse the role of j and k because we cannot have $i = j$ and $i = k$). By independence of the observations,

$$\begin{aligned} \mathbb{E}[v_{ij}v_{ik}|Z] &= \mathbb{E}[\mathbb{E}[v_{ij}v_{ik}|Z, X_i, X_k]|Z] = \mathbb{E}[v_{ik}\mathbb{E}[v_{ij}|Z, X_i, X_k]|Z] = \mathbb{E}[v_{ik}\mathbb{E}[v_{ij}|Z_j, Z_i, X_i]|Z] \\ &= \mathbb{E}[v_{ik}\{\mathbb{E}[1(X_{1j} \le X_{1i})|Z_j, Z_i, X_i] - F_{X_1|Z}(X_{1i}|Z_j)\}|Z] = 0. \end{aligned}$$

Therefore, it follows that

$$\mathbb{E}[\sum_{i=1}^n \{b_i(Z)' \sum_{j=1}^n q_j v_{ij}/\sqrt{n}\}^2/n|Z] \le \sum_{i=1}^n b_i(Z)'\{\sum_{j,k=1}^n q_j \mathbb{E}[v_{ij}v_{ik}|Z]q_k'/n\}b_i(Z)/n =$$

$$\sum_{i=1}^n b_i(Z)'\{\sum_{j=1}^n q_j \mathbb{E}[v_{ij}^2|Z]q_j'/n\}b_i(Z)/n \le \sum_{i=1}^n b_i(Z)'\hat{Q}b_i(Z)/n,$$

so the conclusion follows by CM. *Q.E.D.*

LEMMA C2: *(Lorentz, 1986, p. 90, Theorem 8). If Assumption 5.1 is satisfied then there exists* C *such that for each* x *there is* $\gamma(x)$ *with* $\sup_{z \in Z} |F_{X_1|Z}(x|z) - q^L(z)'\gamma(x)| \le CK^{-d_1/r}$.

LEMMA C3: *If Assumption 5.2 is satisfied then for each K there exists a nonsingular constant matrix B such that $\tilde{p}^K(w) = Bp^K(w)$ satisfies $E[\tilde{p}^K(w_i)\tilde{p}^K(w_i)'] = I_K$, $sup_{w\in\mathcal{W}}\|\tilde{p}^K(w)\| \le CK_\eta^\alpha K$, $sup_{w\in\mathcal{W}}\|\partial\tilde{p}^K(w)/\partial t\| \le CK_\eta^{\alpha+2}K$, and $sup_{[0,1]}\|\tilde{p}^{K_\eta}(t)\| \le CK_\eta^{1+\alpha}$.*

Proof: For $u \in [0,1]$, let $P_j^\alpha(u)$ be the j^{th} orthonormal polynomial with respect to the weight $u^\alpha(1-u)^\alpha$. Denote $\mathcal{X} = \Pi_{\ell=1}^{s-1}[\underline{x}_\ell, \bar{x}_\ell]$. By the fact that the order of the power series is increasing and that all terms of a given order are included before a term of higher order, for each k and $\lambda(k,\ell)$ with $p_k(w) = \Pi_{\ell=1}^s w_\ell^{\lambda(k,\ell)}$, there exists $b_{kj}, (j \le k)$, such that

$$\sum_{j=1}^k b_{kj}p_j(w) = \Pi_{\ell=1}^{s-1}P^0_{\lambda(k,\ell)}([x_\ell - \underline{x}_\ell]/[\bar{x}_\ell - \underline{x}_\ell])P^\alpha_{\lambda(k,s)}(t).$$

Let B_k denote a $K\times 1$ vector $B_k = (b_{k1}, ..., b_{kk}, 0')'$, $b_{kk} \ne 0$where 0 is a $K-k$ dimensional zero vector and let $\bar{B}$ be the $K \times K$ matrix with k^{th} row B_k'. Then $\bar{B}$ is a lower triangular matrix with nonzero diagonal elements and so is nonsingular. As shown in Andrews (1991) there is C such that $|P_j^\alpha(u)| \le C(j^{\alpha+1/2}+1) \le Cj^{\alpha+1/2}$ and $|dP_j^\alpha(u)/du| \le Cj^{\alpha+5/2}$ for all $u \in [0,1]$ and $j \in \{1,2,...\}$. Then for $\bar{p}^K(w) = \bar{B}p^K(w)$, it follows that $|\bar{p}_k(w)| \le C\lambda(k,s)^{\alpha+1/2}\Pi_{\ell=1}^{s-1}\lambda(k,\ell)^{1/2}$, so that $\|\bar{p}^K(w)\| \le CK_\eta^\alpha K$, and $\sup_{w\in\mathcal{W}}\|\partial\bar{p}^K(w)/\partial t\| \le CK_\eta^{\alpha+2}K$. Then by Assumption 5.2, it follows that $\Omega_K = E[\bar{p}^K(w_i)\bar{p}^K(w_i)'] \ge CI_K$. Let $\tilde{B} = \Omega_K^{-1/2}$, and define $\tilde{p}^K(w) = \tilde{B}\bar{p}^K(w)$. Then $\|\tilde{p}^K(w)\| = \sqrt{\tilde{p}^K(w)'\tilde{p}^K(w)} \le \sqrt{\bar{p}^K(w)'\Omega^{-1}\bar{p}^K(w)} \le C\|\bar{p}^K(w)\|$ and an analogous inequality holds for $\|\partial\tilde{p}^K(w)/\partial t\|$, giving the conclusion. Q.E.D.

Henceforth define $\zeta = CK_\eta^\alpha K$ and $\zeta_1 = CK_\eta^{\alpha+2}K$. Also, since the estimator is invariant to nonsingular linear transformations of $p^K(w)$, we can assume that the conclusion of Lemma C3 is satisfied with $p^K(w)$ replacing $\tilde{p}^K(w)$.

Proof of Theorem 4: Let $\delta_{ij} = F_{X_1|Z}(X_{1i}|Z_j) - q_j'\gamma^L(X_{1i})$, with $|\delta_{ij}| \le L^{-d_1/r}$ by Lemma C2. Then for $\tilde{\eta}_i = \tilde{F}(X_{1i}|Z_i)$

$$\tilde{\eta}_i - \bar{\eta}_i = \Delta_i^I + \Delta_i^{II} + \Delta_i^{III},$$

where

$$\Delta_i^I = q_i'\hat{Q}^-\sum_{j=1}^n q_jv_{ij}/n, \Delta_i^{II} = q_i'\hat{Q}^-\sum_{j=1}^n q_j\delta_{ij}/n, \Delta_i^{III} = -\delta_{ii}.$$

Note that $|\Delta_i^{III}| \le CL^{-d_1/r}$ by Lemma C2. 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

$\sum_i q_i'\hat{Q}^- q_i = tr(\hat{Q}^-\hat{Q}) \leq CL$. By CS and Assumption 5.1,

$$\begin{aligned}\sum_{i=1}^n (\Delta_i^{II})^2/n &\leq \sum_{i=1}^n (q_i'\hat{Q}^- q_i \sum_{j=1}^n \delta_{ij}^2/n)/n \leq C\sum_{i=1}^n (q_i'\hat{Q}^- q_i) L^{-2d_1}/n \\ &= CL^{-2d_1/r} tr(\hat{Q}^-\hat{Q}) \leq CL^{1-2d_1/r}.\end{aligned}$$

Note that for $b_i(Z) = q_i'\hat{Q}^-/\sqrt{n}$ we have

$$\sum_{i=1}^n b_i(Z)'\hat{Q}b_i(Z)/n = tr(\hat{Q}\hat{Q}^-\hat{Q}\hat{Q}^-)/n = tr(\hat{Q}\hat{Q}^-)/n \leq CL/n = O_p(L/n),$$

so it follows by Lemma A1 that $\sum_{i=1}^n (\Delta_i^I)^2/n = O_p(L/n)$. The conclusion then follows by T and by $|\tau(\tilde{\eta}) - \tau(\eta)| \leq |\tilde{\eta} - \eta|$, which gives $\sum_i (\hat{\eta}_i - \bar{\eta}_i)^2/n \leq \sum_i (\tilde{\eta}_i - \bar{\eta}_i)^2/n$. Q.E.D.

Before proving other results we give some useful lemmas. For these results let $p_i = p^K(w_i)$, $\hat{p}_i = p^K(\hat{w}_i)$, $p = [p_1, ..., p_n]$, $\hat{p} = [\hat{p}_1, ..., \hat{p}_n]$, $\hat{P} = \hat{p}'\hat{p}/n$, $\tilde{P} = p'p/n$, $P = \mathbb{E}[p_i p_i']$. Note that in the statement of these results we allow $\hat{\eta}_i$ and $\bar{\eta}_i$ to be vectors. Also, as in Newey (1997) it can be shown that without loss of generality we can set $P = I_K$.

Lemma C4: *If Assumptions 3.1 - 3.2 are satisfied then* $\mathbb{E}[Y|X,Z] = \beta(X,\bar{\eta})$.

Proof: By the proof of Theorem 1, $\bar{\eta} = F_{X_1|Z}(X_1|Z)$ is a function of X_1 and Z that is invertible in X_1 with inverse $X_1 = h(Z,\bar{\eta})$. By independence of Z and (ε,η), ε is independent of Z conditional on η, so that

$$\begin{aligned}\mathbb{E}[Y|X,Z] &= \mathbb{E}[Y|X,Z,\eta] = \mathbb{E}[g(X,\varepsilon)|X,Z,\eta] = \mathbb{E}[g(h(Z,\eta),\varepsilon)|\eta,Z] \\ &= \int g(h(Z,\eta),e)F_{\varepsilon|\eta}(de|\eta) = \beta(X,\eta),\end{aligned}$$

at $\bar{\eta} = F_{X_1|Z}(X_1|Z)$. Q.E.D.

Let $u_i = Y_i - \beta(X_i,\eta_i)$, and let $u = (u_1,\dots,u_n)'$.

Lemma C5: *If* $\sum_i \|\hat{\eta}_i - \eta_i\|^2/n = O_p(\Delta_n^2)$ *and Assumptions 5.1 - 5.4 are satisfied then*

$$\begin{aligned}(i),\ \|\tilde{P} - P\| &= O_p(\zeta\sqrt{K/n}); (ii)\ \|p'u/n\| = O_p(\sqrt{K/n}), (iii)\ \|\hat{p} - p\|^2/n = O_p(\zeta_1^2\Delta_n^2), \\ (iv),\ \|\hat{P} - \tilde{P}\| &= O_p(\zeta_1^2\Delta_n^2 + \sqrt{K}\zeta_1\Delta_n); (v)\ \|(\hat{p} - p)'u/n\| = O_p(\zeta_1\Delta_n/\sqrt{n}).\end{aligned}$$

Proof: The first two results follow as in eq. A.1 and p. 162 of Newey (1997). For *(iii)* a mean value expansion gives $\hat{p}_i = p_i + [\partial p^K(\tilde{w}_i)/\partial\eta](\hat{\eta}_i - \bar{\eta}_i)$, where $\tilde{w}_i = (x_i,\tilde{\eta}_i)$ and $\tilde{\eta}_i$ lies in between $\hat{\eta}_i$ and $\bar{\eta}_i$. Since $\hat{\eta}_i$ and $\bar{\eta}_i$ lie in $[0,1]$, it follows that $\tilde{\eta}_i \in [0,1]$ so that by Lemma C3, $\|\partial p^K(\tilde{w}_i)/\partial v\| \le C\zeta_1$. Then by CS, $\|\hat{p}_i - p_i\| \le C\zeta_1|\hat{\eta}_i - \bar{\eta}_i|$. Summing up gives

$$\|\hat{p} - p\|^2/n = \sum_{i=1}^{n} \|\hat{p}_i - p_i\|^2/n = O_p(\zeta_1^2\Delta_n^2). \tag{C.1}$$

For *(iv)*, by Lemma C3, $\sum_{i=1}^{n} \|p_i\|^2/n = O_p(\mathbb{E}[\|p_i\|^2]) = tr(I_K) = K$. Then by T, CS, and M,

$$\begin{aligned} \|\hat{P} - \tilde{P}\| &\le \sum_{i=1}^{n} \|\hat{p}_i\hat{p}_i' - p_ip_i'\|/n \le \sum_{i=1}^{n} \|\hat{p}_i - p_i\|^2/n + 2(\sum_{i=1}^{n} \|\hat{p}_i - p_i\|^2/n)^{1/2}(\sum_{i=1}^{n} \|p_i\|^2/n)^{1/2}. \\ &= O_p(\zeta_1^2\Delta_n^2 + \sqrt{K}\zeta_1\Delta_n). \end{aligned}$$

Finally, for *(v)*, for $Z = (Z_1, ..., Z_n)$ and $X = (X_1, ..., X_n)$, it follows from Lemma C4, Assumption 5.4, and independence of the observations that $\mathbb{E}[uu'|X,Z] \le CI_n$, so that by p and $\hat{p}$ depending only on Z and X,

$$\begin{aligned} \mathbb{E}[\|(\hat{p} - p)'u/n\|^2|X,Z] &= tr\{(\hat{p} - p)'\mathbb{E}[uu'|X,Z](\hat{p} - p)/n^2\} \\ &\le C\|\hat{p} - p\|^2/n^2 = O_p(\zeta_1^2\Delta_n^2/n). \end{aligned}$$

Q.E.D.

LEMMA C6: *If Assumptions 5.1-5.4 are satisfied and* $K\zeta_1^2\Delta_n^2 \longrightarrow 0$, *then w.p.a.1,* $\lambda_{\min}(\hat{P}) \ge C$, $\lambda_{\min}(\tilde{P}) \ge C$.

Proof: By Lemma C3 and $\zeta^2 K/n \le CK\zeta_1^2 L/n$, we have $\|\hat{P} - \tilde{P}\| \xrightarrow{p} 0$ and $\|\tilde{P} - P\| \xrightarrow{p} 0$, so the conclusion follows as on p. 162 of Newey (1997). Q.E.D.

Let $\beta = (\beta(w_1), \ldots, \beta(w_n))'$, and $\hat{\beta} = (\beta(\hat{w}_1), \ldots, \beta(\hat{w}_n))'$.

LEMMA C7: *If* $\sum_i \|\hat{\eta}_i - \eta_i\|^2/n = O_p(\Delta_n^2)$, *Assumptions 5.1 - 5.4 are satisfied,* $\sqrt{K}\zeta_1\Delta_n \to 0$, *and* $K\zeta^2/n \to 0$ *then for* $\tilde{\alpha} = \hat{P}^{-1}\hat{p}'\hat{\beta}/n, \bar{\alpha} = \hat{P}^{-1}\hat{p}'\beta/n$,

$$(i)\ \|\hat{\alpha} - \bar{\alpha}\| = O_p(\sqrt{K/n}), (ii)\ \|\tilde{\alpha} - \bar{\alpha}\| = O_p(\Delta_n), (iii)\ \|\tilde{\alpha} - \alpha^K\| = O_p(K^{-d/s}).$$

Proof: For (i)

$$\begin{aligned} \mathbb{E}[\|\hat{P}^{1/2}(\hat{\alpha} - \bar{\alpha})\|^2|X,Z] &= \mathbb{E}[u'\hat{p}\hat{P}^{-1}\hat{p}'u/n^2|X,Z] = tr\{\hat{P}^{-1/2}\hat{p}'\mathbb{E}[uu'|X,Z]\hat{p}\hat{P}^{-1/2}\}/n^2 \\ &\le Ctr\{\hat{p}\hat{P}^{-1}\hat{p}'\}/n^2 \le Ctr(I_K)/n = CK/n. \end{aligned}$$

Since by Lemma C6, $\lambda_{\min}(\hat{P}) \ge C$ w.p.a.1, this implies that $\mathbb{E}[\|\hat{\alpha} - \bar{\alpha}\|^2|X,Z] \le CK/n$. Similarly, for *(ii)*,

$$\|\hat{P}^{1/2}(\tilde{\alpha} - \bar{\alpha})\|^2 \le C(\hat{\beta} - \beta)'\hat{p}\hat{P}^{-1}\hat{p}'(\hat{\beta} - \beta)/n^2 \le C\|\hat{\beta} - \beta\|^2/n = O_p(\Delta_n^2),$$

which follows from $\beta(w)$ being Lipschitz in η, so that also $\|\tilde{\alpha}-\bar{\alpha}\|^2 = O_p(\Delta_n^2)$. Finally for *(iii)*,

$$\begin{aligned} \|\hat{P}^{1/2}(\tilde{\alpha}-\alpha^K)\|^2 &= \|\tilde{\alpha}-\hat{P}^{-1}\hat{p}'\hat{p}\alpha^K/n\|^2 \le C(\hat{\beta}-\hat{p}'\alpha^K)'\hat{p}\hat{P}^{-1}\hat{p}'(\hat{\beta}-\hat{p}'\alpha^K)/n^2 \\ &\le \|\hat{\beta}-\hat{p}\alpha^K\|^2/n \le C\sup_{w\in\mathcal{W}}|\beta_0(w)-p^K(w)'\alpha^K|^2 = O_p(K^{-2d/s}), \end{aligned}$$

so that $\|\hat{P}^{1/2}(\tilde{\alpha}-\alpha^K)\|^2 = O_p(K^{-2d/s})$. Q.E.D.

Proof of Theorem 5: Note that by Theorem 4, for $\Delta_n^2 = L/n + L^{1-2d_1/r}$, we have $\sum_i \|\hat{\eta}_i - \eta_i\|^2/n = O_p(\Delta_n^2)$, so by $K\zeta^2/n \le CK\zeta_1^2 L/n$ the hypotheses of Lemma C7 are satisfied. Also by Lemma C7 and T, $\|\hat{\alpha}-\alpha^K\|^2 = O_p(K/n + K^{-2d/s} + \Delta_n^2)$. Then

$$\begin{aligned} \int [\hat{\beta}(w)-\beta(w)]^2 F_w(dw) &= \int [p^K(w)'(\hat{\alpha}-\alpha^K) + p^K(w)'\alpha^K - \beta(w)]^2 F_w(dw) \\ &\le C\|\hat{\alpha}-\alpha^K\|^2 + CK^{-2d} = O_p(K/n + K^{-2d/s} + \Delta_n^2). \end{aligned}$$

For the second part of Theorem 5,

$$\begin{aligned} \sup_{w\in\mathcal{W}} |\hat{\beta}(w)-\beta(w)| &= \sup_{w\in\mathcal{W}} |p^K(w)'(\hat{\alpha}-\alpha^K) + p^K(w)'\alpha^K - \beta(w)| \\ &= O_p(\zeta(K/n + K^{-2d/s} + \Delta_n^2)^{1/2}) + O_p(K^{-d/s}) \\ &= O_p(\zeta(K/n + K^{-2d/s} + L/n + L^{1-2d_1/r})^{1/2}). \end{aligned}$$

Q.E.D.

Proof of Theorem 6: Let $\bar{p} = \int_0^1 p^{K_\eta}(t)dt$ and note that by Lemma C3, $\bar{p}'\bar{p} \le CK_\eta^{2+2\alpha}$. Also,

$$\bar{p}(x) \overset{def}{=} \int_0^1 p^K(w)dt = p^{K_x}(x)\otimes\bar{p}. \tag{C.2}$$

As above, $\mathbb{E}[uu'|X,Z] \le CI_n$, so that by Fubini's Theorem,

$$\begin{aligned} \mathbb{E}[\int \{\bar{p}(x)'(\hat{\alpha}-\bar{\alpha})\}^2 F_X(dx)|X,Z] &= \int \{\bar{p}(x)'\hat{P}^{-1}\hat{p}'\mathbb{E}[uu'|X,Z]\hat{p}\hat{P}^{-1}\bar{p}(x)\}F_X(dx)/n^2 \\ &\le C\int \bar{p}(x)'\hat{P}^{-1}\bar{p}(x)F_X(dx)/n \le C\mathbb{E}[\bar{p}(X)'\bar{p}(X)]/n \\ &= C\{\mathbb{E}[p^{K_x}(X)'p^{K_x}(X)](\bar{p}'\bar{p})\}/n = K_xK_\eta^{2+2\alpha}/n. \end{aligned}$$

It then follows by CM that $\int\{\bar{p}(x)'(\hat{\alpha}-\bar{\alpha})\}^2F_X(dx) = O_p(K_xK_\eta^{2+2\alpha})/n)$.

$$\int \bar{p}(x)\bar{p}(x)'F_X(dx) = I_{K_x}\otimes\bar{p}\bar{p}' \le CI_K\bar{p}'\bar{p} \le CI_KK_\eta^{2+2a},$$

so that by Lemma C7 and T,

$$\begin{aligned} \int \{\bar{p}(x)'(\bar{\alpha}-\alpha^K)\}^2F_X(dx) &\le (\bar{\alpha}-\alpha^K)'\int \bar{p}(x)\bar{p}(x)'F_X(dx)(\bar{\alpha}-\alpha^K) \\ &\le CK_\eta^{2+2a}\|\bar{\alpha}-\alpha^K\|^2 = O_p(K_\eta^{2+2a}(K^{-2d/s}+\Delta_n^2)). \end{aligned}$$

Also, by CS,

$$\int \{\bar{p}(x)'\alpha^K - \mu(x)\}^2 F_X(dx) \leq \int \int_0^1 \{p^K(w)'\alpha - \beta(w)\}^2 d\eta F_X(dx) = O(K^{-2d/s}).$$

Then the conclusion follows by T and

$$\begin{aligned}\int [\hat{\mu}(x) - \mu(x)]^2 F_0(dx) &= \int \{\bar{p}(x)'(\hat{\alpha} - \alpha^K) + \bar{p}(x)'\alpha^K - \mu(x)\}^2 F_X(dx) \\ &= O_p(K_x/n + K_x^{-4d} + \Delta_n^2) + O_p(K_x^{-4d}). \quad Q.E.D.\end{aligned}$$

REFERENCES

Altonji, J., and R. Matzkin (2001), "Panel Data Estimators for Nonseparable Models with Endogenous Regressors", Department of Economics, Northwestern University.

Altonji, J., and H. Ichimura, (1997), "Estimating Derivatives in Nonseparable Models with Limited Dependent Variables," mimeo, Northwestern University.

Amemiya, T. (1982), "Two Stage Least Absolute Deviations Estimators," *Econometrica* 50, 689-712.

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): "The Interpretation of Instrumental Variable Estimators in Simultaneous Equations Models with An Application to the Demand for Fish," *Review of Economic Studies 67*, 499-527.

Athey, S. (2002), "Monotone Comparative Statics Under Uncertainty" *Quarterly Journal of Economics*, 187-223.

Athey, S., and P. Haile (2002), "Identification of Standard Auction Models", *Econometrica* 70, 2107-2140.

Athey, S., and S. Stern, (1998), "An Empirical Framework for Testing Theories About Complementarity in Organizational Design", NBER working paper 6600.

Bajari, P., and L. Benkard (2001), "Demand Estimation with Heterogenous Consumers and Unobserved Product Characteristics: A Hedonic Approach," unpublished paper, Department of Economics, Stanford University.

Blundell, R., and J.L. Powell (2000): "Endogeneity in Nonparametric and Semiparametric Regression Models," invited lecture, 2000 World Congress of the Econometric Society.

Brown, D., and R. Matzkin, (1996): "Estimation of Nonparametric Functions in Simultaneous Equations Models, with an Application to Consumer Demand," mimeo, Northwestern University.

Card, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, Vol. 69(5), 1127-1160.

Chamberlain, G., (1984): "Panel Data," in Griliches and Intrilligator (eds.), *Handbook of Econometrics*, Vol 2.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics* 34, 305-334.

Chesher, A. (2001), "Quantile Driven Identification of Structural Derivatives," Cemmap working paper CWP08/01.

Chesher, A. (2002), "Semiparametric Identification in Duration Models," Cemmap working paper CWP20/02.

Chesher, A. (2003), "Local Identification in Nonseparable Models," *Econometrica*, forthcoming.

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.

Doss, H. and R.D. Gill (1992): "An Elementary Approach to Weak Convergence for Quantile Processes, With Applications to Censored Survival Data," *Journal of the American Statistical Association* 87, 869-877.

Hausman, J.A. and W.K. Newey (1995), "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," *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.

Koenker, R. and L. Ma (2002): "Quantile Regression for the Structural Equation Model," preprint.

Lewbel, A., (2002); "Endogenous Selection or Treatment Model Estimation," unpublished working paper.

Lorentz, G., (1986), *Approximation of Functions*, New York: Chelsea Publishing Company.

Ma, L. and R. Koenker (2003): "Quantile Regression Methods for Recursive Structural Equation Models," working paper, University of Illinois.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80:2, 319-323.

Manski, C. (1995): *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge, MA.

Manski, C. (1997): "The Mixing Problem in Program Evaluation," *Review of Economic Studies* 64, 537-553.

Mark, S., and J. Robins, "Estimating the Causal Effect of Smoking Cessation in the Presence of Confounding Factors Using a Rank-Preserving Structural Failure Time Model," *Statistics in Medicine*, 12, 1605-1628.

Matzkin, R. (1993), "Restrictions of Economic Theory in Nonparametric Models" *Handbook of Econometrics*, Vol IV, Engle and McFadden (eds.)

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.

Mundlak, Y., (1963), "Estimation of Production Functions from a Combination of Cross-Section and Time-Series Data," in *Measurement in Economics, Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, C. Christ (ed.), 138-166.

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 (2003): "Nonparametric Instrumental Variables Estimation," *Econometrica*, forthcoming.

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, "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.

Robins, J. (1995): "An Analytic Method for Randomized Trials with Informative Censoring: Part 1, *Lifetime Data Analysis* 1, 241-254.

Roehrig, C. (1988): "Conditions for Identification in Nonparametric and Parametric Models", *Econometrica* 55, 875-891.

Schumaker (1981): *Spline Functions*, Wiley, New York.

Stoker, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica* 54, 1461-1481.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Variable Models: An Equivalence Result," *Econometrica* 70, 331-342.

Wooldridge, J. (2002): *Econometric Analysis of Cross Section and Panel Data* MIT Press.