69

working paper
department
of economics

Nonparametric Estimation Of Triangular Simultaneous Equations Models

Whitney Newey

No. 98-16 October 1998

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139

WORKING PAPER
DEPARTMENT
OF ECONOMICS

Nonparametric Estimation Of Triangular Simultaneous Equations Models

Whitney Newey

No. 98-16 October 1998

MASSACHUSETTS
INSTITUTE OF
TECHNOLOGY

50 MEMORIAL DRIVE
CAMBRIDGE, MASS. 02142

Nonparametric Estimation of Triangular Simultaneous Equations Models[1]

Whitney K. Newey
Department of Economics
MIT, E52-262D
Cambridge, MA 02139
wnewey@mit.edu

James L. Powell
Department of Economics
University of California at Berkeley

Francis Vella
Department of Economics
Rutgers University
New Brunswick, NJ 08903
Vella@fas-econ.rutgers.edu

March 1995

Revised, Febuary 1998

Abstract

This paper presents a simple two-step nonparametric estimator for a triangular simultaneous equation model. Our approach employs series approximations that exploit the additive structure of the model. The first step comprises the nonparametric estimation of the reduced form and the corresponding residuals. The second step is the estimation of the primary equation via nonparametric regression with the reduced form residuals included as a regressor. We derive consistency and asymptotic normality results for our estimator, including optimal convergence rates. Finally we present an empirical example, based on the relationship between the hourly wage rate and annual hours worked, which illustrates the utility of our approach.

Keywords: Nonparametric Estimation, Simultaneous Equations, Series Estimation, Two-Step Estimators.

[1] The NSF provided financial support for the work of Newey and Powell on this paper. An early version of this paper was presented at an NBER conference on applications of semiparametric estimation, October 1993, at Northwestern. We are grateful to participants at this conference and workshops at Harvard/MIT and Tilburg for helpful comments. We are also grateful to the referees for their extensive comments that have helped both the content and style of this paper.

1. Introduction

Structural estimation is important in econometrics, because it is needed to account correctly for endogeneity that comes from individual choice or market equilibrium. Often structural models do not have tight functional form specifications, so that it is useful to consider nonparametric structural models and their estimation. This paper proposes and analyzes one approach to nonparametric structural modeling and estimation. This approach is different than standard nonparametric regression, because the object of estimation is the structural model and not just a conditional expectation.

Nonparametric structural models have been previously considered in Roehrig (1988), Newey and Powell (1989) and Vella (1991). Roehrig (1988) gives identification results for a system of equations when the errors are independent of the instruments. Newey and Powell (1989) consider identification and estimation under the weaker condition that the disturbance has conditional mean zero given the instruments. The results of this paper are complementary to these previous results. The model we consider is more restrictive in some ways than Newey and Powell (1989), but is easier to estimate.

The model we consider is a triangular nonparametric simultaneous equations model where

$$(1.1)\qquad \begin{aligned} y &= g_0(x,z_1) + \varepsilon \\ x &= \Pi_0(z) + u \end{aligned}, \qquad E[\varepsilon|u,z] = E[\varepsilon|u], \quad E[u|z] = 0,$$

x is a $d_x \times 1$ vector of endogenous variables, z is a $d_1 \times 1$ vector of instrumental variables that includes z_1 as a $d_{11} \times 1$ subvector, $\Pi_0(z)$ is a $d_x \times 1$ vector of functions of the instruments z, and u is a $d_x \times 1$ vector of disturbances. Equation (1.1) generalizes the limited information simultaneous equations model to allow for a nonparametric relationship $g_0(x,z_1)$ between the variables y and (x,z_1) and a nonparametric reduced form $\Pi_0(z)$.

The conditional mean restriction in equation (1.1) is one conditional mean version of the usual orthogonality condition for a linear model.[2] Equation (1.1) is a more general assumption than requiring that (ε,u) be independent of z and that $E[u] = 0$. This added generality may be important for some econometric models, because it allows for conditional heteroskedasticity in the disturbances. For example, in some separable demand models y can be purchases of a commodity and x expenditures on a subgroup of commodities. Endogeneity results from expenditure on a commodity subset being a choice variable. Also, heteroskedasticity often results from individual heterogeneity in demand functions, as pointed out by Brown and Walker (1989). Assumption (1.1) allows for both endogeneity and heteroskedasticity.

An alternative model, considered by Newey and Powell (1989), requires only that $E[\varepsilon|z] = 0$. Strictly speaking, neither model is stronger than the other, because equation (1.1) does not imply that $E[\varepsilon|z] = 0$.[3] The additive separability of x into a reduced form and a residual u satisfying equation (1.1) is a strong restriction. One benefit of such a condition is that it leads to a straightforward estimation approach, as will be further discussed below. In this sense, the model of equation (1.1) is a convenient one for applications where its restrictions are palatable. In contrast, estimation is very difficult if only the conditional mean assumption $E[\varepsilon|z] = 0$ is satisfied, as discussed in Newey and Powell (1989).[4]

We propose a two-step nonparametric estimator of this model. The first step consists of the construction of a residual vector $\hat{u}$ from the nonparametric regression of x on z. The second step is the nonparametric regression of y on x, z_1, and

[2] If $\Pi_0(z)$ were linear in z and the conditional expectations were replaced by population regressions, then the third condition implies that z is orthogonal to u, and the second condition that z is orthogonal to ε.

[3] If equation (1.1) is satisfied, u is independent of z, and $E[\varepsilon] = 0$, then $E[\varepsilon|z] = E[E[\varepsilon|u,z]|z] = E[E[\varepsilon|u]|z] = E[E[\varepsilon|u]] = E[\varepsilon] = 0$.

[4] The mapping from the reduced form $E[y|z]$ to the structure $g_0(x,z_1)$ turns out to be discontinuous, making it difficult to construct a consistent estimator.

$\hat{u}$. Two-step nonparametric kernel estimation has previously been considered in Ahn (1994). Our results concern series estimators, which are convenient for imposing the additive structure implied by equation (1.1). We derive optimal mean-square convergence rates and asymptotic normality results that account for the first step estimation and allow for $\sqrt{n}$-consistency of mean-square continuous functionals of the estimator.

Section 2 of the paper considers identification, presenting straightforward sufficient conditions. Section 3 describes the estimator and Section 4 derives convergence rates. Section 5 gives conditions for asymptotic normality and inference, including consistent standard error formulae. Section 6 describes an extension of the model and estimators to semiparametric models. Section 7 contains an empirical example and Monte Carlo results illustrating the utility of our approach.

2. Identification

An implication of our model is that for $\lambda_0(u) = E[\varepsilon|u]$,

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

$$= g_0(x,z_1) + \lambda_0(u) = h_0(w), \quad w = (x', z_1', u')'.$$

Thus, the function of interest, $g_0(x,z_1)$, is one component of an additive regression of y on (x,z_1) and u. Equation (2.1) also implies that $E[\varepsilon|u,z] = E[y-g_0(x,z_1)|u,z] = E[\lambda_0(u)+\{y-E[y|x,z]\}|u,z] = \lambda_0(u)$ only depends on u, which implies $E[\varepsilon|u,z] = E[\varepsilon|u]$, as specified in equation (1.1). Thus, the additive structure of equation (2.1) is equivalent to the conditional mean restriction of equation (1.1). Furthermore, u is identified, so that the identification of g_0 under equation (1.1) is the same as the identification of this additive component in equation (2.1).

To analyze identification of $g_0(x,z_1)$, note that a conditional expectation is unique with probability one, so that any other additive function $\bar{g}(x,z_1)+\bar{\lambda}(u)$

satisfying equation (2.1) must have $\mathrm{Prob}(\bar{g}(x,z_1)+\bar{\lambda}(u) = g_0(x,z_1)+\lambda_0(u)) = 1$. Therefore, identification is equivalent to equality of conditional expectations implying equality of the additive components, up to a constant. Equivalently, working with the difference of two conditional expectations, identification is equivalent to the statement that a zero additive function must have only constant components. To be precise we have

Theorem 2.1: $g_0(x,z_1)$ *is identified, up to an additive constant, if and only if* $\mathrm{Prob}(\delta(x,z_1) + \gamma(u) = 0) = 1$ *implies there is a constant* c_g *with* $\mathrm{Prob}(\delta(x,z_1) = c_g) = 1$.

There is a straightforward interpretation of this result that leads to a simple sufficient condition for identification. Suppose that identification fails. Then by Theorem 2.1, there are $\delta(x,z_1)$ and $\gamma(u)$ with $\delta(x,z_1) + \gamma(u) = 0$ and $\delta(x,z_1)$ nonconstant. Intuitively, this implies a functional relationship between u and (x,z_1), a degeneracy in the joint distribution of these two random variables. For instance, if $\gamma(u)$ were a one-to-one function of a subvector u_1 of u, then $u_1 = \gamma^{-1}(\delta(x,z_1))$. More generally, $\delta(x,z_1) + \gamma(u) = 0$ implies an exact relationship between the random vectors (x,z_1) and u. Consequently, the absence of an exact relationship will imply identification.

To formalize these ideas it is helpful to be precise about what we mean by existence of a functional relationship.

Definition: There is a functional relationship between (x,z_1) *and* u *if and only if there exists* $h(x,z_1,u)$ *and a set* $\mathcal{U}$ *such that* $\mathrm{Prob}(\mathcal{U}) > 0$, $\mathrm{Prob}(h(x,z_1,u) = 0) = 1$ *and* $\mathrm{Prob}(h(x,z_1,\bar{u}) = 0) < 1$ *for all fixed* $\bar{u} \in \mathcal{U}$.

This condition says that the pair of vectors (x,z_1) and u solve a nontrivial implicit equation. Thus, the functional relationship of this definition is an implicit one. As previously noted, nonidentification implies existence of such an implicit relationship. Therefore, the contrapositive statement leads to the following sufficiency result for identification.

Theorem 2.2: If there is no functional relationship between (x,z_1) and u then $g_0(x,z_1)$ is identified up to an additive constant.

Although it is a sufficient condition, nonexistence of a functional relationship between (x,z_1) and u is not a necessary condition for identification. By Theorem 2.1, it is nonexistence of an *additive* functional relationship that is necessary and sufficient condition. Thus, identification may still occur when there is an exact, *nonadditive* functional relationship.

The additive structure in equation (2.1) is so strong that the model may be identified even when the usual order condition is not satisfied, i.e. even though z has smaller dimension than (x,z_1). For example, suppose x is two dimensional, z_1 is not present, z is a scalar, $\Pi(z) = (z,e^z)$, and $g_0(x)$ and $\lambda_0(u)$ are restricted to be differentiable. In this example there is only one instrument although there are two endogenous variables. The reduced form $x = \Pi(z) + u$ implies a nonlinear, nonadditive relationship between x and u of the form $x_2-u_2 = \exp(x_1-u_1)$. Consider an additive function $\delta(x) + \gamma(u)$ satisfying

$$(2.2) \qquad 0 = \delta(x) + \gamma(u) = \delta(x_1,\exp(x_1-u_1)+u_2) + \gamma(u_1,u_2) = 0.$$

The restriction that $g_0(x)$ and $\lambda_0(u)$ are differentiable means that $\delta(x)$ and $\gamma(u)$ must also be differentiable. Let numerical subscripts denote partial derivatives with respect to corresponding arguments. Then equation (2.2) implies

$$\delta_1(x) + \delta_2(x)\exp(x_1-u_1) = 0, \quad -\delta_2(x)\exp(x_1-u_1) + \gamma_1(u) = 0, \quad \delta_2(x) + \gamma_2(u) = 0.$$

Combining the last two equations gives

$$\gamma_1(u) = -\gamma_2(u)\exp(u_1-x_1).$$

Assuming that support of x_1 conditional on u contains more than one point with probability one, so that x_1 can vary for given u, it follows from this equation that with probability one $\gamma_1(u) = \gamma_2(u) = 0$. Then the second and first equations imply $\delta_1(x) = \delta_2(x) = 0$, implying that both δ and γ are constant functions. Thus, equation (2.2) implies that $\delta(x)$ is a constant function, and hence $g_0(x)$ is identified, up to a constant, by Theorem 2.1.

There is a simple sufficient condition for identification that generalizes the rank condition for identification in a linear simultaneous equations system. Suppose that z is partitioned as $z = (z_1,z_2)$. Let x, u, 1, and 2 subscripts denote differentiation with respect to x, u, z_1, and z_2 respectively.

Theorem 2.3: If $g(x,z_1)$, $\cdot\lambda(u)$, and $\Pi(z)$ are differentiable, the boundary of the support of (z,u) has zero probability, and with probability one $\mathrm{rank}(\Pi_2(z)) = d_x$, then $g_0(x,z_1)$ is identified.

If $\Pi(z)$ were linear in z then $\mathrm{rank}(\Pi_2(z)) = d_x$ is precisely the condition for identification of one equation of a linear simultaneous system, in terms of the reduced form coefficients. Thus, this condition is a nonlinear, nonparametric generalization of the rank condition.

It is interesting to note, as pointed out by a referee, that this condition leads to an explicit formula for the structure. Note that $h(x,z) = E[y|x,z] = g(x,z_1) + \lambda(x-\Pi(z))$. Differentiating gives

$$h_x(z,x) = g_x(x,z_1) + \lambda_u(u), \quad h_1(x,z_1) = g_1(x,z_1) - \Pi_1(z)'\lambda_u(u),$$

$$h_2(x,z) = -\Pi_2(z)'\lambda_u(u),$$

Then, if $\mathrm{rank}(\Pi_2(z)) = d_x$ for almost all z, multiplying $h_2(x,z)$ by $D(z) = [\Pi_2(z)\Pi_2(z)']^{-1}\Pi_2(z)$ and solving gives $\lambda_u(u) = -D(z)h_2(x,z)$. Plugging this result into the other terms gives

(2.3) $$g_1(x,z_1) = h_1(x,z) - \Pi_1(z)'D(z)h_2(x,z),$$

$$g_x(x,z_1) = h_x^{\cdot}(x,z) + D(z)h_2(x,z).$$

This formula gives an explicit equation for the derivatives of the structural function g in terms of the identified conditional expectations $\Pi(z)$ and $h(x,z)$.

So far we have only considered identification of $g_0(x,z_1)$ up to an additive constant. This is sufficient for many purposes, such as comparing $g_0(x,z_1)$ at different values of (x,z_1). In other cases, such as forecasting demand quantity, it is also desirable to know the level of $g_0(x,z_1)$, which can be identified if a location restriction is imposed on the distribution of ε. For example, suppose that it is assumed that $E[\varepsilon] = 0$, so that $E[y] = E[g_0(x,z_1)]$. Then from equation (2.2) it follows that if $\tau(u)$ is some function with $\int\tau(u)du = 1$,

(2.4) $$\int E[y|x,z_1,u]\tau(u)du - E[\int E[y|x,z_1,u]\tau(u)du] + E[y]$$

$$= g_0(x,z_1) - E[g_0(x,z_1)] + E[y] = g_0(x,z_1).$$

Thus, under the familiar restriction $E[\varepsilon] = 0$ the conditions for identification of $g_0(x,z_1)$ up to a constant, as in equation (2.2), also serve to identify the level of $g_0(x,z_1)$.

Another approach to identifying the constant term is to place restrictions on $\lambda_0(u)$. For example, the restriction $\lambda_0(0) = 0$ is a generalization of a condition that holds in the linear simultaneous equations model because, in the linear model where $g_0(x,z_1)$ and $\Pi(z)$ are linear and equation (1.1) holds with population regressions replacing conditional expectations, the regression of y on (x,z_1) and u will be linear in u. More generally, we will consider imposing the restriction $\lambda_0(\bar{u}) = \bar{\lambda}$ for some known $\bar{u}$ and $\bar{\lambda}$. In this case

(2.5) $$E[y|x,z_1,\bar{u}] - \bar{\lambda} = g_0(x,z_1) + \lambda_0(\bar{u}) - \bar{\lambda} = g_0(x,z_1).$$

Thus, under the restriction $\lambda_0(\bar{u}) = \bar{\lambda}$, there is a simple, explicit formula for the structural function in terms of h. The constant will be identified under other types of restrictions as well, but this will suffice for many purposes.

3. Estimation

Equation (2.1) can be estimated in two steps by combining estimation of the residual u with estimation of the additive regression in equation (2.1). The first step of this procedure is formation of nonparametrically estimated residuals $\hat{u}_i = w_i - \hat{\Pi}(z_i)$, $(i = 1, \ldots, n)$. The second step is estimation of the additive regression in equation (2.1), using the estimated residuals $\hat{u}_i$ in place of the true ones. An estimator of the structural function $g_0(x,z_1)$ can then be recovered by pulling out the component that depends on those variables.

Estimation of additive regression models has previously been considered in the literature, and we can apply those results for our second step estimation. There are at least two general approaches to estimation of an additive component of a nonparametric regression. One is to use an estimator that imposes additivity and then pull out the component of interest. This approach for kernel estimators has been considered by Breiman and Friedman (1985), Hastie and Tibshirani (1991), and others. Series estimators are particularly convenient for imposing additivity, by simply excluding interaction terms, as in Stone (1985) and Andrews and Whang (1990).

Another approach to estimating the additive component $g_0(x,z_1)$ of the conditional expectation is to use the fact that, for a function $\tau(u)$ such that $\int\tau(u)du = 1$,

$$\int E[y|x,z_1,u]\tau(u)du = g_0(x,z_1) + \int\lambda_0(u)\tau(u)du.$$

Then $g_0(x,z)$ can be estimated by integrating, over u, an unrestricted estimator of the conditional expectation. This partial means approach to estimating additive models has been proposed by Newey (1994), Tjostheim and Auestad (1994), and Linton and Nielsen (1995). It is computationally convenient for kernel estimators, although it is less asymptotically efficient than imposing additivity for estimation of $\sqrt{n}$-consistent functionals of additive models when the disturbance is homoskedastic.

In this paper we focus on series estimators because of their computational convenience and high efficiency in imposing additivity. To describe the two-step series estimator we initially consider the first step. For each positive integer L let $r^L(z) = (r_{1L}(z),\ldots,r_{LL}(z))'$ be a vector of approximating functions. In this paper we will consider in detail polynomial and spline approximating functions, that will be further discussed below. Let an i subscript index the observations, and let n be the total number of observations. Let $\hat{\Pi}(z)$ be the predicted value from a regression of x_i on $r_i = r^L(z_i)$,

$$(3.1) \qquad \hat{\Pi}(z) = r^L(z)'\hat{\gamma}, \quad \hat{\gamma} = (R'R)^{-1}R'(x_1,\ldots,x_n)', \quad R = [r_1,\ldots,r_n]'.$$

To form the second step, let $p^K(w) = (p_{1K}(w),\ldots,p_{KK}(w))'$ be a vector of approximating functions of $w = (x',z_1',u')'$ such that each $p_{kK}(w)$ depends either on (x,z_1) or on u, but not both. Exclusion of the interaction terms will mean that any linear combination of the approximating functions has the same additive structure as in equation (2.1), i.e. that additivity is imposed. Also, let $1(\mathcal{A})$ denote the indicator function for the event $\mathcal{A}$, and $\tau(w)$ denote a function of the form

$$(3.2) \qquad \tau(w) = \Pi_{j=1}^{d+d_x} 1(a_j \le w_j \le b_j)$$

where a_j and b_j are finite constants, w_j is the jth component of w, and $d = d_x + d_{11}$ is the dimension of (x,z_1). This $\tau(w)$ is a trimming function to be further discussed below. Let $\hat{u}_i = x_i - \hat{\Pi}(z_i)$, $\hat{w}_i = (x_i',z_{1i}',\hat{u}_i')'$, where an i subscript for w refers to the observation, and $\hat{\tau}_i = \tau(\hat{w}_i)$. The second step is obtained by regressing y_i on $\hat{p}_i = p^K(\hat{w}_i)$ for each observation where $\hat{\tau}_i = 1$, giving

$$(3.3)\qquad \hat{h}(w) = p^K(w)'\hat{\beta}, \quad \hat{\beta} = (\hat{P}'\hat{P})^{-1}\hat{P}'Y$$

$$\hat{P} = [\hat{\tau}_1\hat{p}_1,\ldots,\hat{\tau}_n\hat{p}_n]', \quad Y = (y_1,\ldots,y_n)'.$$

The trimming function $\tau(w)$ is convenient for the asymptotic theory. It can also be used to exclude large values of w that might lead to outliers in the case of polynomial regression. Although we assume that the trimming function is nonrandom, some results would extend to the case where the limits a_j and b_j are estimated. In particular, if these limits are estimated from independent data (e.g. a subsample that is not otherwise used in estimation), then the results described here will still hold.

The estimator $\hat{h}(w)$ can be used to construct an estimator of the structural function $g_0(x,z_1)$ and of $\lambda_0(u)$ by collecting those terms that depend only on (x,z_1) and those that depend only on u, respectively. Suppose that $p_{1K}(w) = 1$ is the constant, that $p_{kK}(w)$ depends only on (x,z_1) for the next K_g terms, and that the remaining terms depend only on u. Then the estimators can be formed as

$$(3.4)\qquad \hat{g}(x,z_1) = \hat{c}_g + \textstyle\sum_{j=2}^{K_g+1}\hat{\beta}_jp_j(x,z_1), \quad \hat{\lambda}(u) = \hat{c}_\lambda + \sum_{j=K_g+2}^{K}\hat{\beta}_jp_j(u), \quad \hat{c}_g + \hat{c}_\lambda = \hat{\beta}_1.$$

These estimators are uniquely defined except for the constant terms $\hat{c}_g$ and $\hat{c}_\lambda$.

To complete the definition of this estimator the constant terms should be specified. For many objects of estimation, such as predicting how $g(x,z_1)$ changes as x or z_1 shifts, the constants are not important. For example, if $g(x,z_1)$ represents log-demand and x includes the log of price, a constant is not needed for estimation of elasticities (i.e. derivatives of g). In other cases the constant is important. For example, knowing the level of demand is important for predicting tax receipts.

Because of the trimming we cannot use the most familiar condition that $E[\varepsilon] = 0$ to estimate the constants. However, the condition that $\lambda_0(\bar{u}) = \bar{\lambda}$, which generalizes the linear model, can easily be used to estimate the constants. Choosing $\hat{c}_\lambda = \bar{\lambda} - \sum_{j=K_g+2}^{K}\hat{\beta}_jp_j(\bar{u})$ and $\hat{c}_g = \hat{\beta}_1 - \hat{c}_\lambda$ leads to an estimator satisfying $\hat{\lambda}(\bar{u}) = \bar{\lambda}$ and equation (3.4). In this case it will follow that the estimator of the individual components are

$$(3.6) \qquad \hat{g}(x,z_1) = \hat{h}(x,z_1,\bar{u}) - \bar{\lambda}, \quad \hat{\lambda}(u) = \hat{h}(x,z_1,u) - \hat{h}(x,z_1,\bar{u}) + \bar{\lambda}.$$

We consider two specific types of series estimators corresponding to polynomial and spline approximating functions. To describe these let $\mu = (\mu_1,\ldots,\mu_{d_1})'$ denote a vector of nonnegative integers, i.e. a multi-index, with norm $|\mu| = \sum_{j=1}^{d_1}\mu_j$, and let $z^\mu \equiv \Pi_{j=1}^{d_1}(z_j)^{\mu_j}$. For a sequence $(\mu(\ell))_{\ell=1}^{\infty}$ of distinct such vectors, a power series approximation for the first stage is given by

$$r^L(z) = (z^{\mu(1)},\ldots,z^{\mu(L)})'.$$

To describe a power series in the second stage, let $(\mu(k))_{k=1}^{\infty}$ denote a sequence of multi-indices with dimension the same as w, and such that for each k, $w^{\mu(k)}$ depends only on (x,z_1) or u, but not both. This feature can be incorporated by making sure that the first d components of $\mu(k)$ are zero if any of the last d_x components are nonzero, and vice versa. Then for the second stage a power series approximation can be obtained as

$$p^K(w) = (w^{\mu(1)},\ldots,w^{\mu(K)})'.$$

For the asymptotic theory it will be assumed that each multi-index sequence is ordered with degree $|\mu|$ increasing in ℓ or k, and that all terms of a particular order are included before increasing the order.

The theory to follow uses orthonormal polynomials, which may also have computational advantages. For the first step, replacing each power z^μ by the product of univariate polynomials of the same order, where the univariate polynomials are orthonormal with respect to some distribution, may lead to reduced collinearity. The estimator will be numerically invariant to such a replacement, because $|\mu(\ell)|$ is monotonically increasing. An analogous replacement could also be used in the second step.

Regression splines are smooth piecewise polynomials with fixed knots (join points). They have been considered in the statistics literature by Agarwal and Studden (1980). They are different than smoothing splines. For regression splines the smoothing parameter is the number of knots. They have attractive features relative to power series in that they are less sensitive to outliers and to bad approximations over small regions. The theory here requires that the knots be placed in the support of z_i, which therefore must be known, and that the knots be evenly spaced. It should be possible to generalize results to allow the boundary of the support to be estimated and the knots not evenly spaced, but these generalizations lead to further technicalities that we leave to future research.

To describe regression splines it is convenient to assume that $a_j = -1$, $b_j = 1$, and that the support of z is a cartesian product of the interval $[-1,1]$. For a scalar c let $(c)_+ = 1(c > 0)\cdot c$. An m^{th} degree spline with $J-1$ evenly spaced knots on $[-1,1]$ is a linear combination of

$$p_{jJ}(u) = \begin{cases} u^{j-1} & 1 \leq j \leq m+1, \\ \{[u + 1 - 2(j-m-1)/J]\}_+^m, & m+2 \leq j \leq m+J \end{cases}.$$

In this paper we take m fixed but will allow J to increase with sample size. For a set of multi-indices $\{\mu(\ell)\}$, with $\mu_j(\ell) \leq m+J$ for each j and ℓ, the approximating functions for z will be products of univariate splines. In particular, if the number of knots for the jth component of z is J_j, then the approximating functions could be formed as

$$(3.7) \qquad r_{kK}(z) = \left\{ \prod_{j=1}^{d_1} p_{\mu_j(k),J_j}(z_j), \quad (k = 1, \ldots K), \right.$$

Throughout the paper it will be assumed that J_j depends on K in such a way that the ratio of numbers of knots for each pair of elements of z is bounded above and away from zero. Spline approximating functions in the second step could be formed in the analogous way, from products of univariate splines, with additivity imposed by only including terms that depend only on one of (x,z_1) or u.

The theory to follow uses B-splines, which are a linear transformations of the above functions that have lower multicollinearity. The low multicollinearity of B-splines and recursive formula for calculation also lead to computational advantages; e.g. see Powell (1981).

4. Convergence Rates

In this Section we derive mean-square error and uniform convergence rates. To obtain results it is important to impose some regularity conditions. Let $X = (x,z)$, $\eta = y - h_0(w)$, and $u = x - \Pi_0(z)$. Also, for a matrix D let $\|D\| = [\text{trace}(D'D)]^{1/2}$, for a random matrix Y, $\|Y\|_v = \{E[\|Y\|^v]\}^{1/v}$, $v < \infty$, and $\|Y\|_\infty$ the infimum of constants C such that $\text{Prob}(\|Y\| < C) = 1$.

Assumption 1: $\{(y_i,x_i,z_i)\}$, $(i = 1, 2, \ldots)$ is i.i.d. and $\text{Var}(x|z)$ and $\text{Var}(y|X)$ are bounded.

The bounded second conditional moment assumption is quite common in the series estimation literature (e.g. Stone, 1985). Relaxing it would lead to complications that we wish to avoid.

Next, we impose the following requirement. Let $\mathcal{W} = \{w : \tau(w) = 1\}$.

Assumption 2: z is continuously distributed with density that is bounded away from zero on its support, and the support of z is a cartesian product of compact, connected intervals. Also, w is continuously distributed and the density of w is bounded away from zero on $\mathcal{W}$, and $\mathcal{W}$ is contained in the interior of the support of w.

This assumption is useful for deriving the properties of series estimators like those considered here (e.g. see Newey, 1997). It allows us to bound below the eigenvalues of the second moment matrix of the approximating functions. Also, an identification condition like those discussed in Section 2 is embodied in this assumption. The density of w being bounded away from zero means that there is no functional relationship between (x,z_1) and u, so that by Theorem 2.2 identification holds. This assumption, along with smoothness conditions discussed below, imply that the dimension of z must be at least as large as the dimension of (x,z_1), a familiar order condition for identification.

To see why Assumption 2 implies the order condition, note that the density of w is bounded away from zero on an open set. Then, because w is a one-to-one function of x, z_1, and $\Pi(z)$, the density of $(z_1,\Pi(z))$ must be bounded away from zero on an open set. But $(z_1,\Pi(z))$ is a vector function of z that will be smooth under assumptions given below, so its range must be confined to a manifold of smaller dimension than $(z_1,\Pi(z))$ unless z has at least as many components as $(z_1,\Pi(z))$. Since the dimension of $\Pi(z)$ and x are the same, z must have as many components as (x,z_1).

Some discreteness in the components of z can be allowed for without affecting the results. In particular, Assumption 2 can be weakened to hold only for some component of the distribution of z. Also, one could allow some components of z to be discretely distributed, as long as they have finite support. On the other hand, it is not easy to extend our results to the case where there are discrete instruments that take on an infinite number of values.

Some smoothness conditions are useful for controlling the bias of the series estimators.

Assumption 3: $\Pi_0(z)$ is continuously differentiable of order s_1 on the support of z and $g_0(x,z_1)$ and $\lambda_0(u)$ are Lipschitz and continuously differentiable of order s on $\mathcal{W}$.

The derivative orders s_1 and s control the rate at which polynomials or splines approximate the function. In particular, the rate of approximation for $g_0(x,z_1)$ and $\lambda(u)$, and hence also for $h_0(w)$, will be $O(K^{-s/d})$, where d is the dimension of (x,z_1).

The last regularity condition restricts the rate of growth of the number of terms K and L. Recall that d_1 denotes the dimension of z.

Assumption 4: Either a) for power series, $(K^3 + K^2L)[(L/n)^{1/2}+L^{-s_1/d_1}) \longrightarrow 0$; or b) for splines, $(K^2 + KL^{1/2})[(L/n)^{1/2}+L^{-s_1/d_1}] \longrightarrow 0$.

For example, for splines, if K and L grow at the same rate then this condition requires that they grow slower than $n^{1/5}$, and that $s_1 > 2d_1$.

As a preliminary step it is helpful to derive convergence rates for the conditional expectation estimator $\hat{h}(w)$. Let F_0 denote the distribution function of w.

Lemma 4.1: If Assumptions 1-4 are satisfied then

$$\int \tau(w)[\hat{h}(w)-h_0(w)]^2 dF_0(w) = O_p(K/n + K^{-2s/d} + L/n + L^{-2s_1/d_1}).$$

Also, for $q = 1/2$ *for splines, and* $q = 1$ *for power series,*

$$\sup_{w\in \mathcal{W}}|\hat{h}(w) - h_0(w)| = O_p(K^q[(K/n)^{1/2} + K^{-s/d} + (L/n)^{1/2} + L^{-s_1/d_1}]).$$

This and all subsequent proofs are given in the Appendix.

This result leads to convergence rates for the structural estimator $\hat{g}(x,z_1)$ given in equation (3.4). We will treat the constant term a little differently for the mean square and uniform convergence rates, so it is helpful to state and discuss them separately. For mean square convergence we give a result for mean-square convergence of the de-meaned difference between the estimator and the truth. Let $\bar{\tau} = E[\tau(w)]$.

Theorem 4.2: If Assumptions 1-4 are satisfied then for $\hat{\Delta}(x,z_1) = \hat{g}(x,z_1)-g_0(x,z_1)$

$$\int\tau(w)[\hat{\Delta}(x,z_1)-\int\tau(w)\hat{\Delta}(x,z_1)F_0(dw)/\bar{\tau}]^2F_0(dw) = O_p(K/n + K^{-2s/d} + L/n + L^{-2s_1/d_1}).$$

The convergence rate is the sum of two terms, depending on the number L of approximating functions in the first step and the number K in the second step. Each of these two terms has a form analogous that derived in Stone (1985), which would attain the optimal mean-square convergence rate for estimating a conditional expectation if K and L were chosen appropriately. In particular, if K is chosen proportional to $n^{d/(d+2s)}$ and L proportional to $n^{d_1/(d_1+2s_1)}$, and Assumption 4 is satisfied, then the conclusion of Theorem 4.2 is

$$(4.1)\quad \int\tau(w)[\hat{\Delta}(x,z_1)-\int\tau(w)\hat{\Delta}(x,z_1)F_0(dw)/\bar{\tau}]^2F_0(dw) = O_p(\max\{n^{-2s/(d+2s)}, n^{-2s_1/(d_1+2s_1)}\}).$$

From Stone (1982) it follows that the convergence rates in this expression are optimal for their respective steps, i.e. $n^{-2s/(d+2s)}$ would be optimal for estimation of $g_0(x,z_1)$ if u did not have to be estimated in the first step. Thus, when K and L are chosen appropriately, the mean square error convergence rate of the second step estimator is the slower of the optimal rate for the first step and the optimal rate for the second step that would obtain if the first step did not have to be estimated.

An important feature of this result is that the second step convergence rate $n^{-2s/(d+2s)}$ is the optimal one for a function of a d-dimensional variable rather than the slower rate that would be optimal for estimation of a general function of w, where additivity was not imposed. This occurs because the additivity of $h_0(w)$ is used in estimation. It is essentially the exclusion of interaction terms that leads to this result, as discussed in Andrews and Whang (1990).

Although a full derivation of bounds on the rate of convergence of estimators of

$g_0(x,z_1)$ is beyond the scope of this paper, something can be said.[5] It is known from Stone (1982) that $n^{-2s/(d+2s)}$ is the best attainable mean-square error (MSE) rate for estimation of $g_0(x,z_1)$ *when* u *is known.* Since the best rate when u is unknown cannot be better than the best rate when u is known, $\hat{g}$ must attain the best rate when the rate is $n^{-2s/(d+2s)}$. Therefore, for the ranges of d_1, s_1, d, s where this rate is attained we know that it is the best one (and that the estimator has the best rate). Setting K and L to be proportional to the rates $n^{d/(d+2s)}$ and $n_1^{d_1/(d_1+2s_1)}$ that minimize each component of the mean-square error it follows that the optimal second stage rate will be attained for splines when $s/d \geq s_1/d_1$ and $s/d > 2$. For power series the side conditions of Assumption 4 are even stronger and would narrow the range of s/d and s_1/d_1 where the estimator could attain the best convergence rate for the second step.

The condition $s/d \geq s_1/d_1$ means that the second stage is at least as smooth (relative to its dimension) as the first stage. We do not know what the optimal convergence rate would be when the first stage is less smooth than the second stage, although in that case we know that the first stage convergence rate would dominate the mean-square error of our estimator if K and L are chosen to minimize the mean-square error in Theorem 4.1 or 4.2. Our estimator could conceivably be suboptimal in that case. Also, the side conditions in Assumption 4 are quite strong and rule out any possibility of optimality of our estimator when neither the first or the second stage is very smooth (e.g. when $s/d \leq 1$).

It is interesting to note that the convergence rates in equation (4.1) are only relevant for the function itself and not for derivatives. Because $\hat{u}$ is plugged into $\hat{h}(w)$, one might think that convergence rates for derivatives are also important, as in a Taylor expansion of $\hat{h}(\hat{u})$ around $\hat{h}(u)$. The reason that they are not is that $\hat{u}$ depends only on the conditioning variables x and z in equation (2.1), a feature that

[5] We thank a referee for pointing this out and for giving a slightly more general version of the following discussion of optimality.

allows us to work with a Taylor expansion of $h_0(w)$ rather than $\hat{h}(w)$ in the proofs.

Theorem 4.2 gives convergence rates that are net of the constant term. It would also be interesting to know if similar results hold when the constant term is included. That will depend on the identifying assumption for the constant term. With the trimming the usual restriction $E[\varepsilon] = 0$ does not identify the constant. Also, the restriction $\lambda_0(\bar{u}) = \bar{\lambda}$ is a pointwise one, so it does not lead to mean-square error convergence rates. It would be possible to obtain mean-square rates that include the constant term by specifying a restriction $E[\tau(w)\varepsilon] = 0$, but this condition does not seem to be well motivated in the model and so will not be considered.

It is straightforward to obtain uniform convergence rates for $\hat{g}(x,z_1)$ from the uniform convergence rates for $\hat{h}(w)$ in Lemma 4.1. In particular, note that $\hat{g}(x,z_1)$ can be recovered from $\hat{h}(w)$, up to the constant term, by just fixing u at some value (in $\mathcal{W}$), so that uniform rates follow immediately. Furthermore, as long as the restriction on $\lambda_0(u)$ used to identify the constant term is continuous in the supremum norm on $\mathcal{W}$, uniform rates for $\hat{g}$ will also follow from Lemma 4.I. In particular, when the identification restriction $\lambda_0(\bar{u}) = \bar{\lambda}$ is imposed leading to an estimator as in equation (3.6), uniform convergence rates for $\hat{g}$ follow immediately from those for $\hat{h}$. Specifically, for $\mathcal{W}_1$ equal to the coordinate projection of $\mathcal{W}$ on values of (x,z_1),

Theorem 4.3: If $\hat{g}(x,z_1) = \hat{h}(x,z_1,\bar{u}) - \bar{\lambda}$, $\lambda_0(\bar{u}) = \bar{\lambda}$, and Assumptions 1-4 are satisfied then for $q = 1/2$ for splines, and $q = 1$ for power series,

$$\sup_{(x,z_1)\in\mathcal{W}_1} |\hat{g}(x,z_1) - g_0(x,z_1)| = O_p(K^q[(K/n)^{1/2} + K^{-s/d} + (L/n)^{1/2} + L^{-s_1/d_1}]).$$

This uniform convergence rate cannot be made equal to Stone's (1982) bounds for either the first or second steps, due to the leading K^q term. Nevertheless, these uniform rates improve on some in the literature, e.g. on Cox (1988), as noted in Newey (1997). Apparently, it is not yet known whether series estimators can attain the optimal uniform convergence rates.

The uniform convergence rate for polynomial regression estimators is slower and the conditions on K and L more stringent for power series than splines. Nevertheless we retain the results for polynomials because they have some appealing practical features. They are often used for increased flexibility in functional form. Also polynomials do not require knot placement, and hence are simpler to use in practice.

We could also derive convergence rates for $\hat{\lambda}(u)$, and obtain convergence rates identical to those in Theorems 4.2 and 4.3. We omit these results for brevity.

5. Inference

Large sample confidence intervals are useful for inference. We focus on pointwise confidence intervals rather than uniform ones, as in much of the literature. For this purpose let h denote a possible true function $h_0(w)$, where the w argument is suppressed for convenience. That is, the symbol h represents the entire function. We consider all linear functionals of h, which turn out to include the value of the function $g_0(x,z_1)$ at a point. We construct confidence intervals using standard errors that account for the nonparametric first step estimation.

To describe the standard errors and associated confidence intervals it is useful to consider a general framework. For this purpose let h denote a possible true function $h_0(w)$, where the w argument is suppressed for convenience. That is, the symbol h represents the entire function. Let $a(h)$ be a particular number associated with h. As a function of h, $a(h)$ represents a mapping from possible functions of h to the real line, i.e. a functional. The object of interest will be assumed to be the value $\theta_0 = a(h_0)$ of the functional at h_0. In this Section we develop standard errors for the estimator $\hat{\theta} = a(\hat{h})$ of θ_0 and give asymptotic normality results that allow formation of large sample confidence intervals.

This framework is general enough to include many objects of interest, such as the value of g at a point. For example, under the restriction $\lambda_0(\bar{u}) = \bar{\lambda}$ discussed earlier, the value of $g_0(x,z_1)$ at a point $(\bar{x},\bar{z}_1)$ can be represented as

$$(5.1) \qquad g_0(\bar{x},\bar{z}_1) = a(h_0), \quad a(h) = h(\bar{x},\bar{z}_1,\bar{u}) - \bar{\lambda}.$$

A general inference procedure for linear functionals will apply to the functional of this equation, i.e. can be used in the construction of large sample confidence intervals for the value of g_0 at a point.

Another example is a weighted average derivative of $g_0(x,z_1)$. This object is particularly interesting when g_0 is an index model that depends on $w_1 = (x,z_1)$ only through a linear combination $w_1\gamma$, say $g_0(w_1) = t(w_1\gamma)$ where $t(q)$ is a differentiable function with derivative $t_q(q)$. Then for any $v(w)$ such that $v(w)t_q(w_1'\gamma)$ is integrable,

$$(5.2) \qquad \int v(w)[\partial h_0(w)/\partial w_1]dw = [\int t_q(w_1\gamma)v(w)dw]\gamma,$$

so that the weighted average derivative is proportional to γ. This is also a linear functional of h_0, where $a(h) = \int v(w)[\partial h(w)/\partial w_1]dw$. This functional may also be useful for summarizing the slope of $g_0(w_1)$ over some range, as discussed in Section 7. Unlike the value of $g_0(w_1)$ or $\partial g_0(w_1)/\partial w_1$ at a point, the weighted average derivative functional will be $\sqrt{n}$-consistent under regularity conditions discussed below.

We can derive standard errors for linear functionals of h, which is a class general enough to include many important examples, such as the value of g at a point or a weighted average derivative. The estimator $\hat{\theta} = a(\hat{h})$ of $\theta_0 = a(h_0)$ is a natural, "plug-in" estimator. For example, the value of g_0 at a point could be estimated by applying the functional in equation (5.1) to obtain $\hat{g}(\bar{x},\bar{z}_1) = a(\hat{h}) = \hat{h}(\bar{x},\bar{z}_1,\bar{u}) - \bar{\lambda}$. In general, linearity of $a(h)$ implies that

$$(5.3) \qquad \hat{\theta} = A\hat{\beta}, \quad A = (a(p_{1K}),\ldots,a(p_{KK})).$$

Because $\hat{\theta}$ is a linear combination of the two-step least squares coefficients $\hat{\beta}$, a natural estimator of the variance can be obtained by applying the formula for parametric two step estimators (see Newey 1984). As for other series estimators, such as in Newey (1997), this should lead to correct inference asymptotically, because it accounts properly for the variability of the estimator. Let $\otimes$ denote the Kronecker product and

$$(5.4) \qquad \hat{Q} = \hat{P}'\hat{P}/n, \quad \hat{\Sigma} = \textstyle\sum_{i=1}^{n}\hat{\tau}_i\hat{p}_i\hat{p}_i'[y_i-\hat{h}(\hat{w}_i)]^2/n, \quad \hat{Q}_1 = I_{d-k_1}\otimes(R'R/n),$$

$$\hat{\Sigma}_1 = \textstyle\sum_{i=1}^{n}(\hat{u}_i\hat{u}_i')\otimes(r_ir_i')/n, \quad \hat{H} = \sum_{i=1}^{n}\hat{\tau}_i[\partial\hat{h}(\hat{w}_i)/\partial u]'\otimes\hat{p}_ir_i'/n.$$

The variance estimate for $a(\hat{h})$ is then

$$(5.5) \qquad \hat{V} = A\hat{Q}^{-1}[\hat{\Sigma} + \hat{H}\hat{Q}_1^{-1}\hat{\Sigma}_1\hat{Q}_1^{-1}\hat{H}']\hat{Q}^{-1}A'.$$

This estimator is like that suggested by Newey (1984) for parametric two-step regressions. It is equal to a term $A\hat{Q}^{-1}\hat{\Sigma}\hat{Q}^{-1}A'$, that is the standard heteroskedasticity consistent variance estimator, plus an additional nonnegative definite term that accounts for the presence of $\hat{u}$. It has this additive form, where the variance of the two-step estimator is larger than the first, because the second step conditions on the first step dependent variables, making the second and first steps orthogonal.

We will show that under certain regularity conditions $\sqrt{n}\hat{V}^{-1/2}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,1)$. Thus doing inference as if $\hat{\theta}$ were distributed normally with mean θ_0 and variance $\hat{V}/n$ will be a correct large sample approximation. This extends large sample inference results of Andrews (1991) and Newey (1997) to account for an estimated first step. However, as in this previous work, such results do not specify the convergence rate of $a(\hat{h})$. That rate will depend on how fast $\hat{V}$ goes to infinity. To date there is still not a complete characterization of the convergence rate of a series estimator available in the literature, although it is known when $\sqrt{n}$-consistency occurs (see Newey, 1997). Those results can be extended to obtain $\sqrt{n}$-consistency of functionals of the two-step estimators developed here. Let $\|\cdot\|^{\tau} = (E[\tau(w)(\cdot)^2]/\bar{\tau})^{1/2}$ denote a trimmed mean-square norm and $\mathcal{P}$ denote the set of functions that can be approximated arbitrarily well in this norm by a linear combination of $p^K(w)$ for large enough K. It is well known, for

$p^K(w)$ corresponding to power series or splines, that $\mathcal{P}$ includes all additive functions in (x,z_1) and u where the individual components have finite $\|\cdot\|^\tau$ norm. The critical condition for $\sqrt{n}$-consistency is that there is $\nu(w) \in \mathcal{P}$ such that

$$(5.6) \qquad a(h) = E[\tau(w)\nu(w)h(w)],$$

for any $h(w) \in \mathcal{P}$. Essentially this means that $a(h)$ can be represented as an expected product of $h(w)$ with a function $\nu(w)$ that is in the same space as h. By the Riesz representation theorem, this is the same as the functional $a(h)$ being continuous with respect to the mean square error $E[\tau(w)(\cdot)^2]/\bar{\tau}$. In this case $\sqrt{n}(\hat{\theta} - \theta_0)$ will be asymptotically normal, and its asymptotic variance can be given an explicit expression.

To describe the asymptotic variance in the $\sqrt{n}$-consistent case, let $\rho(z) = E[\tau(w)\nu(w)\partial h_0(w)/\partial u' \,|\, z]$. The asymptotic variance of the estimator can then be expressed as:

$$(5.7) \qquad \bar{V} = E[\tau(w)\nu(w)\nu(w)'\mathrm{Var}(y|X)] + E[\rho(z)\mathrm{Var}(x|z)\rho(z)'].$$

For example, for the weighted average derivative in equation (5.2), if $v(w)$ is zero where $\tau(w)$ is zero, so $\tau(w)v(w) = v(w)$, then integration by parts shows that equation (5.6) is satisfied with $\nu(w) = -\mathrm{Proj}(f_0(w)^{-1}\partial v(w)/\partial x|\mathcal{P})$, where $f_0(w)$ is the density of w and $\mathrm{Proj}(\cdot|\mathcal{P})$ denotes the mean-square projection on the space of functions that are additive in w_1 and x. The presence of $\mathrm{Proj}(\cdot|\mathcal{P})$ is necessary to make $\nu(w)$ lie in the space spanned by $p^K(w)$ for large K, that is the set of functions that are additive in w_1 and x. Then

$$\rho(z) = -E[\tau(w)\{\mathrm{Proj}(f_0(w)^{-1}\partial v(w)/\partial x|\mathcal{P})\}\partial h_0(w)/\partial u' \,|\, z],$$

and the asymptotic variance will be $\bar{V}$ from equation (5.7).

To state precisely results on asymptotic normality it is helpful to introduce some regularity conditions. Recall that $\eta = y-h_0(w)$.

Assumption 5: $\sigma^2(X) = \mathrm{Var}(y|X)$ is bounded away from zero, $E[\eta^4|X]$ is bounded, and $E[\|u\|^4|X]$ is bounded. Also, $h_0(w)$ is twice continuously differentiable in u with bounded first and second derivatives.

This condition strengthens the bounded conditional second moment condition to boundedness of the fourth conditional moment and the conditional variance being bounded away from zero.

The next assumption is the key condition for $\sqrt{n}$-consistency of the estimator. It requires that the functional $a(h)$ have a representation as an expected product form, at least for the truth h_0 and the approximating functions.

Assumption 6: There exists $\nu(w)$ and β_K such that $E[\tau(w)\|\nu(w)\|^2] < \infty$, $a(h_0) = E[\tau(w)\nu(w)h_0(w)]$, $a(p_{kK}) = E[\tau(w)\nu(w)p_{kK}(w)]$, $E[\tau(w)\|\nu(w)-\beta_K p^K(w)\|^2] \longrightarrow 0$ as $K \longrightarrow \infty$.

This condition also requires that $\nu(w)$ can be approximated well by a linear combination $p^K(w)$ for large K, which is important for the precise form that $\nu(w)$ must take. For example, with the average derivative this condition leads $\nu(w)$ to be the projection of $f_0(w)^{-1}\partial v(w)/\partial x$ on functions that are additive in w_1 and u.

The next condition is complementary to Assumption 6. For $d_w = d + d_x$ and a $d_w \times 1$ vector μ of nonnegative integers let $|\mu| = \sum_{j=1}^{d_w}\mu_j$, $\partial^\mu h(w) = \partial^{|\mu|}h(w)/\partial w_1\cdots\partial w_{d_w}$. Also let δ denote a nonnegative integer and $|h|_\delta = \max_{|\mu|\le\delta}\sup_{w\in\mathcal{W}}|\partial^\mu h(w)|$.

Assumption 7: $a(h)$ is a scalar, $|a(h)| \le |h|_\delta$ for some $\delta \ge 0$, and there exists β_K such that as $K \longrightarrow \infty$, $a(p^{K\prime}\beta_K)$ is bounded away from zero while $E[\tau(w)\{p^K(w)'\beta_K\}^2] \longrightarrow 0$.

This assumption says that functional $a(h)$ is continuous in $|h|_\delta$, but *not* in mean square error. The lack of mean-square continuity will imply that the estimator $a(\hat{h})$ is not $\sqrt{n}$-consistent, and is also a useful regularity condition. Another restriction imposed is that $a(h)$ is a scalar, which is general enough to cover many cases of interest. When $a(h)$ is a vector, asymptotic normality with an estimated covariance matrix would follow from Assumption J iii) of Andrews (1991). That assumption of Andrews (1991) is difficult to verify. In contrast, Assumption 7 is a primitive condition, that is relatively easy to verify. This condition and Assumption 6 are mutually exclusive, and general enough to include many cases of interest. For example, as noted above, Assumption 6 includes the weighted average derivative. For another example, Assumption 7 obviously applies to the functional $a(h) = h(\bar{x},\bar{z}_1,\bar{u}) - \bar{\lambda}$ from equation (5.1), which is continuous with respect to the supremum norm and where it is straightforward to show that there is β_K such that $p^K(\bar{x},\bar{z}_1,\bar{u})'\beta_K - \bar{\lambda}$ is bounded away from zero but $E[\tau(w)\{p^K(w)'\beta_K\}^2] \longrightarrow 0$.

The next condition restricts the allowed rates of growth for K and L.

Assumption 8: $\sqrt{n}K^{-s/d} \longrightarrow 0$, $\sqrt{n}L^{-s_1/d_1} \longrightarrow 0$, for power series, $(K^9L + K^8L^2 + K^6L^3 + K^2L^6)/n \longrightarrow 0$, and for splines $(K^5L + K^4L^2 + K^3L^3 + KL^4)/n \longrightarrow 0$.

One can show that this condition requires that $s/d > 5/2$ and $s_1/d_1 > 2$, as was pointed out by a referee. It also limits the growth rate of K and L so that if each grows at the same rate they can grow no faster than $n^{1/6}$ for splines and $n^{1/10}$ for power series. These conditions are stronger than for single step series estimators discussed in Newey (1997), because of complications due to the multistage nature of the estimation problem.

The next condition is useful for the estimator $\hat{V}$ of the asymptotic variance to have the right properties.

Assumption 9: Either a) x is univariate, if a spline is used it is at least a quadratic one, and $\sqrt{n}K^{1-s} \longrightarrow 0$; b) x is multivariate, $p^K(w)$ is a power series, $h_0(w)$ is differentiable of all orders, there is a constant C with the absolute value of the jth derivative bounded above by $C(C)^j$, and $\sqrt{n}K^{-\epsilon} \longrightarrow 0$ for some $\epsilon > 0$.

This assumption helps guarantee that $\lambda_0(u)$ and its derivative can be approximated by

$p^K(w)$, which is important for consistency of the covariance matrix estimator. These conditions are not very general, but more general ones would require series approximation rates for derivatives, which are not available in the literature. This assumption is at least useful for showing that the asymptotic variance estimator will be consistent under some set of regularity conditions.

Next, for the nonnegative integer δ of Assumption 7, let $\zeta_\delta(K) = \max_{|\mu|\leq\delta}\sup_{\mathcal{W}}\|\partial^\mu p^K(w)\|$.

Theorem 5.1: If Assumptions 1-3, 5, either 6 or 7, and 8 and 9 are satisfied, then $\hat\theta = \theta_0 + O_p(\zeta_\delta(K)/\sqrt{n})$ *and*

$$\sqrt{n}V^{1/2}(\hat\theta - \theta_0) \xrightarrow{d} N(0,I), \quad \sqrt{n}\hat{V}^{1/2}(\hat\theta-\theta_0) \xrightarrow{d} N(0,I).$$

Furthermore, if Assumption 6 is satisfied,

$$\sqrt{n}(\hat\theta - \theta_0) \xrightarrow{d} N(0,\bar{V}), \quad \hat{V} \xrightarrow{p} \bar{V}.$$

In addition to the asymptotic normality needed for large sample confidence intervals this results gives convergence rate bounds and, when Assumption 6 holds, $\sqrt{n}$-consistency. Bounds on the convergence rate can also be derived from the results shown in Newey (1997) that for power series $\zeta_\delta(K) \leq CK^{1+2\delta}$ and for splines $\zeta_\delta(K) \leq CK^{(1/2)+\delta}$, where C is a generic positive constant. Thus, for example, when $a(\hat{h}) = \hat{h}(\bar{x},\bar{z}_1,\bar{u}) - \bar{\lambda}$, where $\delta = 0$, an upper bound on the convergence rate for power series would be $K/\sqrt{n}$ and for splines would be $\sqrt{K}/\sqrt{n}$.

When Assumption 9 b) is satisfied, so that $h_0(w)$ has derivatives of all orders, and $\Pi_0(z)$ also satisfies the same condition, the convergence rate can be made close to $1/\sqrt{n}$ by letting K grow slowly enough. In particular, $h_0(w)$ and $\Pi_0(z)$ are continuously differentiable of order s for *every* $s > 0$, so that if $K = Cn^\epsilon$ and $L = Cn^\epsilon$ for some positive, small ϵ, all of the conditions will be satisfied. Then the the convergence rate for $a(\hat{h}) = \hat{h}(\bar{x},\bar{z}_1,\bar{u}) - \bar{\lambda}$ for power series will then be $n^{(-1/2)+\epsilon}$, which will be close to $1/\sqrt{n}$ for ϵ small enough.

The centering of the limiting distribution at zero in this result is a reflection of "over fitting" that is present, meaning that the bias of these estimators shrinks faster than the variance. This over fitting is implied by Assumption 8. The order of the bias of the estimator is $K^{-s/d}$, so that Assumption 8 requires that the bias shrink faster than $1/\sqrt{n}$. Since $1/\sqrt{n}$ is generally the fastest the standard deviation of estimators can shrink (such as the sample mean), overfitting is present.

This feature of the asymptotic normality result that is shared by other theorems for series estimators, as in Andrews (1991) and Newey (1997). When combined with the condition that K go to infinity slower than n^{α} for $0 < \alpha < 1$, the bias shrinking faster than $1/\sqrt{n}$ means that the convergence rate for the estimator is bounded away from the optimal rate. This is different than asymptotic normality results for other types of nonparametric estimators, such as kernel regression. It would be good to relax this condition, but that is an extension that is beyond the scope of this paper.

6. Additive Semiparametric Models

In economic applications there is often a large number of covariates thereby making nonparametric estimation problematic. The well known curse of dimensionality can make the estimation of large dimensional nonparametric models difficult. One approach to this problem is to restrict the model in some way while retaining some nonparametric features. The single index model discussed earlier is one example of such a restricted model. Another example is a model that is additive in some components and a parametric in others. This type of model has received much attention in the literature as an approach to dimension reduction. Furthermore, it is particularly easy to estimate using a series estimator, by simply imposing restrictions on the approximating functions.

To describe this type of model, let $w_1 = (x,z_1)$ as before, and let w_{1j},

$(j = 0,1,\ldots,J)$, denote $J+1$ subvectors of w_1. A semiparametric additive model is

$$(6.1)\qquad g_0(w_1) = w_{10}'\gamma + \textstyle\sum_{j=1}^J g_j(w_{1j}).$$

This model can be estimated by specifying that the functions of x included in $p^K(w)$ consist of the elements of w_{10} and power series or splines in w_{1j}, $(j=1,\ldots,J)$. One could also impose similar restrictions on $\lambda(u)$ and the reduced form, by specifying that they depend on a linear combination plus some additive components. For example, one could specify a partially linear reduced form as $\Pi_0(z) = z_0'\pi_0 + \sum_{m=1}^M \Pi_m(z_m)$. This restriction could be imposed in estimation by specifying $r^L(z)$ to include the elements of z_0 and power series or splines in each z_m.

The coefficients γ of equation (6.1) are examples of functionals of $h(w) = g_0(w_1) + \lambda_0(u)$ that are mean-square continuous functionals of h, so that a series estimator will be $\sqrt{n}$-consistent under the conditions of Section 5. Let $q(w)$ be the residual from the mean-square projection of w_{10} on functions of the form $\sum_{j=1}^J g_j(w_{1j}) + \lambda_0(u)$, for the probability measure where $P(\mathcal{A}) = E[\tau(w)1(w\in\mathcal{A})]/E[\tau(w)]$. Assume that $E[\tau(w)q(w)q(w)']$ is nonsingular, an identification condition for γ. Then

$$(6.2)\qquad \gamma = E[\tau(w)\nu(w)h(w)],\quad \nu(w) = (E[\tau(w)q(w)q(w)'])^{-1}q(w).$$

Hence, $\sqrt{n}$-consistency and asymptotic normality of the series estimator for γ will follow from the results of Section 5. Robinson (1988) also gave some instrumental variable estimators for a semiparametric model that is linear in endogenous variables.

The regularity conditions of Section 5 can be weakened somewhat for this model. Basically, only the nonparametric part of the model need satisfy these conditions so that, for example, w_{10} need not be continuously distributed.

7. Empirical Example

To illustrate our approach we investigate the empirical relationship between the hourly wage rate and the number of annual hours worked. While early examinations treated the hourly wage rate as independent of the number of hours worked recent studies have incorporated an endogenous wage rate, (see, for example, Moffitt 1984, Biddle and Zarkin 1989, and Vella 1993), and have uncovered a non-linear relationship between the wage rate and annual hours. The theoretical underpinnings of this relationship can be assigned to an increasing importance of labor when it is involved in a greater number of hours (see Oi 1962). Alternatively Barzel (1973) argues that labor is relatively unproductive at low hours of work due to related start up costs. Moreover, at high hours of work Barzel argues that fatigue decreases labor productivity. These forces will generate hourly wage rates that initially increase but subsequently decrease as daily hours work increase. Finally, the taxation/labor supply literature argues that a non-linear relationship may exist due to the progressive nature of taxation rates. Thus the non-linear relationship between hours and wage rates is potentially the outcome of many influences and we capture it in the following model:

$$(7.1) \qquad y_i = z_{1i}'\beta + g_{20}(x_i) + \varepsilon_i, \quad x_i = z_i'\gamma + u_i,$$

where y_i is the log of the hourly wage rate of individual i, z_{1i} is a vector of individual characteristics, x_i is annual hours worked, z_i is a vector of exogenous variables that includes z_{1i}, β and γ are parameters, g_{20} is an unknown function, and ε_i and u_i are zero mean error terms such that $E[\varepsilon|u] \neq 0$. This is a semiparametric version of equation (1.1), like that discussed in Section 6, with a parametric reduced form. We use this specification because there are many individual characteristics in z_i, so that it would be difficult to apply fully nonparametric estimation due to the curse of dimensionality. We estimate this model using data on males from the 1989 wave of the Michigan Panel Survey Of Income Dynamics. To preserve comparability with previous empirical studies we follow similar data exclusions to those in Biddle and Zarkin (1989). Accordingly, we only include males aged between 22 and 55 who had worked between 1000 and 3500 hours in the previous year. This produced a sample of 1314 observations. Our measures of hours and wages are annual hours worked and the hourly wage rate respectively.

We first estimate the wage equation by linear OLS and report the relevant parameters in column 1 of Table 1. The hour effect is small and not significantly different from zero. Adjusting for the endogeneity of hours via linear two stage least squares (2SLS), however, indicates that the impact of hours is statistically significant although the effect is small. The results in column 2, on the basis of the t-statistic for the residuals, suggests hours are endogenous to wages.

To allow the g_{20} function to be non-linear we employed alternative specifications for g_{20} and various approximations for the manner in which we account for the endogeneity of hours. We also allow for some non-linearities in the reduced form by including non-linear terms for the experience and tenure variables. We first choose the number of non-linear terms in the first step and then, by employing the associated residuals from the chosen specification, we determine the number of approximating terms in the primary equation. We discriminate between alternative specifications on the basis of cross validation (CV) criterion. The CV criterion is well known to minimize asymptotic mean-square error when the first estimation is not present so we are hopeful that it will lead to estimates with good properties here. Below we consider its properties in a Monte Carlo study, and find that CV performs well.

Because the theory requires over fitting, where the bias is smaller than the variance asymptotically, it seems prudent to consider specifications with more terms than those that CV gives, particularly for inference. Accordingly, for both steps we identify the number of terms that minimizes the CV criterion and then add an additional term. For example, while the approximation which minimized the CV criterion for the first step was a fourth order polynomial in tenure and experience we generate the residuals from a model which includes a fifth order polynomial in these variables. Note that to enable comparisons, these were the residuals which were included in column 2 discussed above.

The CV values for a subset of the second step specifications examined are reported in Table 2. Although we also experimented with splines, in each step, we found that the specifications involving polynomial approximations appeared to dominate. The preferred specification is a fourth order polynomial in hours while accounting for the endogeneity through a third order polynomial in the residuals. The CV criterion for this specification is 182.643. The fourth column of Table 1 reports the estimates of the hours profile employing these approximations. The third column presents the hours coefficients while excluding the residuals. Due to the over fitting requirement we now take as a base case the specification with a fifth order polynomial in hours and a fourth order polynomial in the residual. We will also check the sensitivity of some results to this choice.

While Table 1 suggests the non-linearity and endogeneity are statistically important it is useful to examine their respective roles in determining the hours profile. For comparison purposes Figure 1 plots the change in the log of the hourly wage rate as annual hours increase, where the nonparametric specification is the over fitted case with $h = 5$ and $r = 4$. To facilitate comparisons each function has been centered at its mean over the observations, and the graph shows the deviation of the function from its mean. As annual hours affect the log of the hourly wage rate in an additive manner we simply plot the impact of hours on the log of the hourly wage rate. In Figure 1 we plot the quadratic 2SLS estimates like those of Biddle and Zarkin (1989) and the profile is very similar to theirs. In Figure 1 we also plot the predicted relationship between hours and wages for these data with and without the adjustment for endogeneity using our base specification. It indicates the relationship is highly non-linear. Furthermore, failing to adjust for the endogeneity leads to incorrect inferences regarding the overall impact of hours on wages and, more specifically, the value of the turning points.

Although we found that the polynomial approximations appeared to best fit the data we also identified the best fitting spline approximations. We employ cubic splines and the parameters we allow to be chosen by the data are the number of join points. We do this for the first and second steps and in this instance, we over fit the data by including an additional join point, in each step, above what is determined on the basis of the CV values. Note that the best fitting approximation involved 1 join points in the reduced form and a primary specification of 1 join point in the hours function and 1 join points for the residuals. In Figure 2 we plot the predicted relationship between hours and the wage rate from the over fitted spline approximation. This predicted relationship looks remarkably similar to that in Figure 1. Furthermore, the points noted above related to the failure to account for the endogeneity are similar.

Figures 1 and 2 indicate that at high numbers of hours the implied decrease in the wage rate is sufficient to decrease total earnings. This may be due to the small number of observations in this area of the profile and the associated large standard errors. Figures 3 and 4 explore this issue by plotting the 95 percent confidence intervals for our adjusted nonparametric estimates of the wage hours profile. It confirms that the estimated standard errors are large at the upper end of the hours range.

To quantify an average measure of the effect of hours on wages we consider the weighted average derivative of the estimated g_{20} function. We estimated the derivative over the range where the function is shown to be upward sloping in Figures 1 and 2. This region, 1300 to 2500 hours, represents 81 percent of the total sample. There were only 125 observations above 2600, so that not very many were excluded by ignoring the upper tail. The average derivative for the 1300-2500 part of the hours profile from the polynomial approximation, based on $h = 5$ and $r = 4$, is .000513 with a standard error of .000155. The corresponding estimate from our spline approximation of the weighted average derivative is .000505 with a standard error of .000147. These estimates are quite different than the linear 2SLS estimate reported in Table 1 which is consistent with the presence of nonlinearity. Furthermore, for the quadratic 2SLS in Figure 1 the average derivative is .000350 with a standard error of .000142, which is also quite different than the average derivative for the nonparametric estimators. This difference is present despite the relatively wide pointwise confidence bands in Figures 3 and 4. Thus, in our example the average wage change for most of the individuals in the sample is found to be substantially larger for our nonparametric approach than by linear or quadratic 2SLS.

Before proceeding we examined the robustness of these estimates of the average derivative to the specification of the reduced form and the alternative specifications of the reduced form and alternative approximations of the wage hours profile. First, we reduced the number of approximating terms in the reduced form. While we found that the CV criteria for each step increased with the exclusion of these terms in the reduced form we found that there was virtually no impact on the total profile and a very minor effect on our estimate of the average derivative of the profile for the region discussed above. For example, consider the estimates from the polynomial approximation. When we included only linear terms in the reduced form the estimated average derivative is .000515 with a standard error of .0000175. The corresponding values for the specification including quadratic terms is .000546 with a standard error of .000174. Finally, the addition of cubic terms resulted in an estimate of .000522 with a standard error of .000153. Changes in the specification of the model based on the spline approximations produced similarly small differences.

We also explored the sensitivity of the average derivative estimate to the number of terms in the series approximation. Both the average derivative and its standard error were relatively unaffected by increasing the number of terms in the approximations. For example, with an 8th order polynomial in hours and 7th order polynomial in the residual the average derivative estimate was .000519 with a standard error of .000156.

Finally we examine whether we are able to reject the additive structure implied by our model. We do this by including an interaction term capturing the product of hours and the residuals in our preferred specification. The t-statistic on this included variable is .231. The CV value for this specification is 182.924 while that for our preferred specification plus this interaction term and the term capturing the product of hours squared and the residual squared is 183.188. On the basis of these results we conclude there is no evidence against the additive structure of the model.

In addition to this empirical evidence we provide simulation evidence featuring the data examined above. The objective of this exercise is to explore the ability of our procedure to accurately estimate the unknown function in the empirical setting we examined above. We simulate the endogenous variable through the exogenous variables and parameter estimates from each of the polynomial and spline approximations reported above. We generate the hours variable by simulating the reduced form and incorporating a random component drawn from the reduced form empirical residual vector. From this reduced form we estimate the residuals and from the simulated hours vector we generate the higher order terms for hours. We then simulate wages by employing the simulated values of hours and residuals, along with the true exogenous variables, and the parameter estimates discussed above. The random component is drawn from the distribution of the empirical wage equation residuals. As we employ the parameter vector from the over fitted models in the empirical the polynomial model has a fifth order polynomial in the reduced form while the wage equation has a 5th order polynomial in hours and a fourth order polynomial in the residuals. The spline approximation employs a cubic spline with 2 join points in the reduced form while the wage equation has 2 join points for hours and 2 join points for the residuals. We examine the performance of our procedure for 3000 replications of the model.

In order to relax the parametric assumptions of the model we use the CV criterion in each step of the estimation. That is, for the model which employs the polynomial approximation we first choose the number of terms in the reduced form. Then on the basis of the residuals from the over fit reduced form we then choose the number of approximating terms in the wage equation. For the model generated by the spline approximation we do the same except we employ a cubic spline and choose the number of join points. For both approximations we trim the bottom and top two and a half percent observations, on the basis of the hours residuals, from the data for which we estimate the wage equation.

An important aspect of our procedure is the ability of the CV method to correctly identify the correct specification. To examine this we computed the CV criteria for the polynomial model for all 25 specifications of the wage equation combining up to a fifth order polynomial in hours and fifth order polynomial in residuals. As there was also 5 choices in the reduced form this generated 125 possibilities. For the spline model we examined up to 3 join points in the reduced form and then 3 each for hours and residuals in the wage equation. This produced 27 possible specifications.

From the Monte Carlo results we can evaluate the performance of an average derivative estimator like that we applied to the actual data. Once again we over fit the data such that we computed the estimator by choosing the number of terms that minimizes the CV criterion, plus one additional term, and computed standard errors for the same specification. First consider the results from the polynomial model. The mean of the average derivative estimates, across the Monte Carlo replications, was .000456, which represents a bias of 8.9 percent. The standard error across the replications was .000153, while the average of the estimated standard errors was .000152. Therefore the estimated standard errors accurately reflect the variability of the estimator in this experiment. For the spline approximation the average estimate was .000442 which represents a bias of 11.1 percent. The standard error across the replications was .000145, while the average of the estimated standard errors was .000144.

We also computed rejection frequencies for tests that the average derivative was equal to its true value. For the polynomial model the rejection rates at the nominal 5, 10 and 20 percent significance levels were 6.5, 11.8 and 22.5 percent respectively. The corresponding figures for the spline model were 7.4, 13.1 and 24.0 percent. Thus, asymptotic inference procedures turned out to be quite accurate in this experiment, lending some credence to our inference in the empirical example.

Appendix

We will first prove one of the results on identification.

Proof of Theorem 2.3: Consider differentiable functions $\delta(x,z_1)$ and $\gamma(u)$, and suppose that

$$0 = \delta(x,z_1) + \gamma(u) = \delta(\Pi(z)+u,z_1) + \gamma(u),$$

identically in z and u. Differentiating then gives

$$0 = \Pi_1(z)'\delta_x(x,z_1) + \delta_1(x,z_1), \quad 0 = \delta_x(x,z_1) + \gamma_u(u), \quad 0 = \Pi_2(z)'\delta_x(x,z_1).$$

If $\Pi_2(z)$ is full rank it follows from the last equation that that $\delta_x(x,z_1) = 0$. It then follows from the first two equations that $\delta_1(x,z_1) = 0$ and $\gamma_u(u) = 0$. Now, to show identification we will use Theorem 2.1. By differentiability of g and λ it suffices to consider $\delta(x,z_1)$ and $\gamma(u)$ that are differentiable. Also, if $\delta(x,z_1)+\gamma(u) = 0$ with probability one, then by continuity of the functions and the boundary having zero probability, this equality holds identically on the interior of the support of (u,z). Then, as above, all the partial derivatives of $\delta(x,z_1)$ and $\gamma(u)$ are zero with probability one, implying they are constant. Identification then follows by Theorem 2.1. QED.

To avoid repetition, it useful to prove consistency and asymptotic normality lemmas for a general two-step weighted least squares estimator. To state the Lemmas some notation is needed. Throughout the appendix C will denote a generic positive constant, that may be different in different uses. Let X be a vector of variables that includes x and z among its components and $w(X,\pi)$ a vector of functions of X and π, where π represents a possible value of $\Pi_0(z) = E[x|z]$. Then for $w = w(X,\Pi_0(z))$, it will be assumed that

(A.1) $E[y|X] = h_0(w), \quad E[x|z] = \Pi_0(z).$

Also, let $\psi_i = \psi(X_i)$ be a scalar, nonnegative function of X_i, $\hat{\Pi}(z)$ and $\hat{\tau}_i$ be as in the body of the paper, and

(A.2) $\hat{h}(w) = p^K(w)'\hat{\beta}, \quad \hat{\beta} = (\hat{P}'\hat{P})^{-1}\hat{P}'Y, \quad \hat{P} = [\hat{\tau}_1\psi_1\hat{p}_1,\ldots,\hat{\tau}_n\psi_n\hat{p}_n]', \quad \hat{p}_i = p^K(\hat{w}_i).$

Let $\mathcal{W} = \{w : \tau(w) = 1\}$, and for δ a nonnegative integer, let $|h|_\delta = \max_{|\mu|\le\delta}\sup_{w\in\mathcal{W}}|\partial^\mu h(w)|$.

Assumption A1: i) $\psi(X)$ is bounded and $w(X,\pi)$ is Lipschitz in π; ii) each component $w_j(X,\pi)$ either does not depend on π or $w_j(X,\Pi_0(z))$ is continuously distributed with bounded density; iii) For each K and L there are nonsingular matrices B and B_1 such that for $P^K(w) = Bp^K(w)$ and $R^L(z) = B_1r^L(z)$, $E[\tau(w)\psi(X)^2P^K(w)P^K(w)']$ and $E[R^L(z)R^L(z)']$ have smallest eigenvalues that are bounded away from zero, uniformly in K and L; iv) For each nonnegative integer δ there is $\zeta_\delta(K)$ with, $\max_{|\mu|\le\delta}\sup_{\mathcal{W}}\|\partial^\mu P^K(w)\| \le \zeta_\delta(K)$ and $\sup_{\mathcal{Z}}\|R^L(z)\| \le \xi(L)$; v) There exists δ, α, $\alpha_1 > 0$ and β_K and γ_L such that $|h_0-p^{K\prime}\beta_K|_\delta \le CK^{-\alpha}$ and $\sup_{\mathcal{Z}}\|\Pi_0(z)-\gamma_L r^L(z)\| \le CL^{-\alpha_1}$.

Lemma A1: If Assumptions 1 and A1 are satisfied for $d = 0$, $\zeta_0(K)^2K/n \longrightarrow 0$, *and* $[K^{1/2}\zeta_1(K) + \zeta_0(K)^2\xi(L)][(L/n)^{1/2} + L^{-\alpha_1}] \longrightarrow 0$, *then*

$$\int\tau(w)\psi(X)^2[\hat{h}(w)-h_0(w)]^2dF_0(X) = O_p(K/n + K^{-2\alpha} + L/n + L^{-2\alpha_1}).$$

If Assumptions 1 and A1 are satisfied for some nonnegative integer d, *then*

$$|\hat{h} - h_0|_d = O_p(\zeta_d(K)[(K/n)^{1/2} + K^{-\alpha} + (L/n)^{1/2} + L^{-\alpha_1}]).$$

Proof of Lemma A1: Note that the value of $\hat{h}$ is unchanged if a nonsingular constant linear transformation of $p^K(w)$ and/or $r^L(z)$ is used, so that it can be assumed that $p^K(w) = P^K(w)$ and $r^L(z) = R^L(z)$. Let $\tau_i = \tau(w_i)$, $p_i = p^K(w_i)$, $Q = E[\tau_i\psi_i^2p_ip_i']$. By

Assumption A1, the smallest eigenvalue of Q is bounded away from zero, so that the largest eigenvalue of Q^{-1} is bounded. Consider replacing $p^K(w)$ by $\tilde{p}^K(w) = Q^{-1/2}p^K(w)$. By the Cauchy-Schwartz inequality there is a constant C such that $\|\partial^{\mu}\tilde{p}^K(w)\| \le C\zeta_{|\mu|}(K)$, so that the hypotheses and the conclusion will be satisfied with this replacement. Since $E[\tau(w)\psi(X)^2\tilde{p}^K(w)\tilde{p}^K(w)'] = I$ by construction, it suffices to prove the results with $Q = 1$. By analogous reasoning, it suffices to prove the result for $Q_1 = E[r^L(z)r^L(z)'] = I$.

Next, let $\Delta_\pi = L^{1/2}/\sqrt{n} + L^{-\alpha_1}$, $\hat{\Pi}_i = \hat{\Pi}(z_i)$, and $\Pi_i = \Pi_0(z_i)$. Also, suppose for notational simplicity that that $\Pi(z)$ is a scalar function. Note that for γ_L from Assumption A1,

$$\sum_{i=1}^{n}\|\hat{\Pi}_i-\Pi_i\|^2/n \le C(\hat{\gamma}-\gamma_L)'\hat{Q}_1(\hat{\gamma}-\gamma_L) + C\sum_{i=1}^{n}\|\Pi_i-\gamma_L r^L(z_i)\|^2/n$$

$$\le C\int[\hat{\Pi}(z)-\gamma_L r^L(z)]^2dF(z) + C|(\hat{\gamma}-\gamma_L)'(\hat{Q}_1-1)(\hat{\gamma}-\gamma_L)| + O(K^{-2\alpha_1})$$

$$\le C\int[\hat{\Pi}(z)-\Pi_0(z)]^2dF(z) + C\int[\Pi(z)-\gamma_L r^L(z)]^2dF(z) + C\|\hat{\gamma}-\gamma_L\|^2o_p(1) + O(K^{-2\alpha_1})$$

$$\le O(\Delta_\pi^2) + C\|\hat{\gamma}-\gamma_L\|^2o_p(1),$$

where the last inequality follows by Theorem 1 of Newey (1997). Also, by eq. (A.2) of the Appendix in Newey (1997), it follows that $\|\hat{\gamma}-\gamma_L\|^2 = O_p(\Delta_\pi^2)$, so that by eq. (A.7a),

$$\text{(A.3)} \qquad \sum_{i=1}^{n}\|\hat{\Pi}_i-\Pi_i\|^2/n = O_p(\Delta_\pi^2).$$

Also, by Theorem 1 of Newey (1997),

$$\text{(A.3a)} \qquad \max_{i\le n}\|\hat{\Pi}_i-\Pi_i\| = O_p(\xi(L)\Delta_\pi).$$

Therefore, by Assumption A1 ii) and Lemma A3,

$$\text{(A.4)} \qquad \sum_{i=1}^{n}|\hat{\tau}_i-\tau_i|/n = O_p(\xi(L)\Delta_\pi).$$

Let $P = [\tau_1\psi_1p_1,\ldots,\tau_n\psi_np_n]'$ and $\tilde{Q} = P'P/n$. Note that $E[\|\tilde{Q}-I\|^2] \le$

$n^{-1}E[\tau_i\psi_i^4\sum_{j=1}^K p_{jK}(w_i)^2\sum_{k=1}^K p_{kK}(w_i)^2] \le Cn^{-1}\zeta_0(K)^2\sum_{k=1}^K E[\tau_i\psi_i^2 p_{kK}(w_i)^2] = Cn^{-1}\zeta_0(K)^2K \longrightarrow 0$, implying $\|\tilde{Q}-I\| \xrightarrow{p} 0$. Also, by $\mathcal{W}$ convex and a mean value expansion in w, and $w(X,\pi)$ Lipschitz in π, $\hat{\tau}_i\tau_i\|\hat{p}_i-p_i\| \le \zeta_1(K)\|\hat{w}_i-w_i\| \le C\zeta_1(K)\|\hat{\Pi}_i-\Pi_i\|$. Also, by the Markov inequality, $\sum_{i=1}^n\tau_i\psi_i^2\|p_i\|^2/n = O_p(K)$. Then by $\hat{\tau}_i^2 = \hat{\tau}_i$, $\tau_i^2 = \tau_i$, and ψ_i bounded,

$$\text{(A.5)}\quad \|\hat{Q}-\tilde{Q}\| \le \|\sum_{i=1}^n\hat{\tau}_i\tau_i\psi_i^2(\hat{p}_i\hat{p}_i' - p_ip_i')/n\| + C\|\sum_{i=1}^n(\hat{\tau}_i\tau_i-\hat{\tau}_i)\hat{p}_i\hat{p}_i'/n\|$$

$$+ C\|\sum_{i=1}^n(\hat{\tau}_i\tau_i-\tau_i)p_ip_i'/n\| \le C\sum_{i=1}^n\tau_i\hat{\tau}_i(\|\hat{p}_i-p_i\|^2 + 2\psi_i\|p_i\|\|\hat{p}_i-p_i\|)/n + C\zeta_0(K)^2\sum_{i=1}^n|\hat{\psi}_i-\psi_i|/n$$

$$\le C\zeta_1(K)^2\sum_{i=1}^n\|\hat{\Pi}_i-\Pi_i\|^2/n + C(\sum_{i=1}^n\tau_i\psi_i^2\|p_i\|^2/n)^{1/2}(\sum_{i=1}^n\tau_i\hat{\tau}_i\|\hat{p}_i-p_i\|^2/n)^{1/2}$$

$$+ O_p(\zeta_0(K)^2\xi(L)\Delta_\pi) = O_p(\zeta_1(K)^2\Delta_\pi^2 + K^{1/2}\zeta_1(K)\Delta_\pi + \zeta_0(K)^2\xi(L)\Delta_\pi) = o_p(1).$$

Let $\eta_i = \psi_i[y_i-h_0(w_i)]$, $\eta = (\eta_1,\dots,\eta_n)$, and $\tilde{X} = (X_1,\dots,X_n)$. Then by independence of the observations $E[\psi_iy_i|\tilde{X}] = \psi_iE[y_i|\tilde{X}] = \psi_ih_0(w_i)$, so $E[\eta_i|\tilde{X}] = 0$. Furthermore, $E[\eta_i^2|\tilde{X}]$ is bounded by ψ_i and $\mathrm{Var}(y_i|\tilde{X}) = \mathrm{Var}(y|X_i)$ bounded, and by independence of the observations, $E[\eta_i\eta_j|\tilde{X}] = E[\eta_i\eta_j|X_i,X_j] = E[\eta_iE[\eta_j|\eta_i,X_i,X_j]|X_i,X_j] = E[\eta_iE[\eta_j|X_j]|X_i,X_j] = 0$ for $i \ne j$. Then by $\hat{P}$ depending only on $\tilde{X}$,

$$\text{(A.6)}\quad E[\|(\hat{P}-P)'\eta/n\|^2|\tilde{X}] \le Cn^{-2}\sum_{i=1}^n\psi_i^2\|\hat{\tau}_i\hat{p}_i-\tau_ip_i\|^2 \le Cn^{-1}\sum_{i=1}^n|\hat{\tau}_i-\tau_i|(\|\hat{p}_i\|^2+\|p_i\|^2)/n$$

$$+ Cn^{-1}\sum_{i=1}^n\hat{\tau}_i\tau_i\|\hat{p}_i-p_i\|^2/n = O_p(n^{-1}[\zeta_0(K)^2\xi(L)\Delta_\pi + \zeta_1(K)^2\Delta_\pi^2]) = o_p(n^{-1}),$$

where the second to last equality follows similarly to eq. (A.5). Then, since a standard result is that $Y_n = O_p(\Delta_n)$ if $E[|Y_n||\tilde{X}] = O_p(\Delta_n)$, it follows that $\|(\hat{P}-P)'\eta/n\|^2 = o_p(1/n)$. Also, $E[\|P'\eta/n\|^2] = E[E[\|P'\eta\|^2/n|\tilde{X}]] = E[\sum_{i=1}^n\tau_i^2\psi_i^4p_i'p_i\mathrm{Var}(y_i|X_i)/n^2] \le CE[\tau_i\psi_i^2p_i'p_i]/n = C\mathrm{tr}(Q)/n = C\mathrm{tr}(I)/n = CK/n$. Therefore, by the triangle inequality,

$$\text{(A.7)}\quad \|\hat{P}'\eta/n\|^2 \le C[\|(\hat{P}-P)'\eta/n\|^2 + \|P'\eta/n\|^2] = o_p(1) + O_p(K/n) = O_p(K/n).$$

Then by eqs. (A.5) and (A.7) and the smallest eigenvalue of $\hat{Q}$ bounded away from zero with probability approaching one, for $\hat{M} = \hat{P}(\hat{P}'\hat{P})^{-1}\hat{P}'$, $\eta'\hat{M}\eta/n = (\eta'\hat{P}/n)\hat{Q}^{-1}(\hat{P}'\eta/n) \le$

$O_p(1)\|\hat{P}'\eta/n\|^2 = O_p(K/n)$. Let $\hat{h}_i = \hat{\tau}_i\psi_i\hat{h}(\hat{w}_i)$, $\tilde{h}_i = \hat{\tau}_i\psi_i h_0(\hat{w}_i)$, $h_i = \tau_i\psi_i h_0(w_i)$, and $\hat{h}$, $\tilde{h}$, and h be the corresponding $n \times 1$ vectors (i.e. $\hat{h} = (\hat{h}_1,\dots,\hat{h}_n)'$). Let β be such that $\sup_{\mathcal{W}}|h_0(w)-p^K(w)'\beta| = O(K^{-\alpha})$. Then by M and $M-I$ idempotent, $(M-I)\tilde{h} = (M-I)(\tilde{h}-\hat{P}\beta)$, $h_0(w)$ Lipschitz, and $\sum_{i=1}^n\|\hat{\Pi}_i-\Pi_i\|^2/n = O_p(\Delta_\pi^2)$, for $\Delta_h = \sqrt{K}/\sqrt{n} + K^{-\alpha} + \Delta_\pi$,

$$\|\hat{h}-\tilde{h}\|^2/n \le \|\hat{M}\eta + \hat{M}(h-\tilde{h}) + (M-I)\tilde{h}\|^2/n \le C(\eta'M\eta/n + \textstyle\sum_{i=1}^n\hat{\tau}_i\psi_i^2[h_0(w_i)-h_0(\hat{w}_i)]^2/n$$

$$+ \textstyle\sum_{i=1}^n\hat{\tau}_i\psi_i^2[h_0(\hat{w}_i)-\hat{p}_i'\beta]^2/n \le O_p(K/n) + C\sum_{i=1}^n\|\hat{\Pi}_i-\Pi_i\|^2/n + O(K^{-2\alpha}) = O_p(\Delta_h^2).$$

It then follows by the smallest eigenvalue of $\hat{Q}$ bounded away from zero with probability approaching one that, for $\bar{h} = \hat{P}\beta$ and $\tilde{y} = (\hat{\tau}_1\psi_1 y_1,\dots,\hat{\tau}_n\psi_n y_n)'$,

$$\|\hat{\beta}-\beta\|^2 \le O_p(1)(\hat{\beta}-\beta)'\hat{Q}(\hat{\beta}-\beta) = O_p(1)(\tilde{y}-\bar{h})'M(\tilde{y}-\bar{h})/n \le O_p(1)[\eta'M\eta/n + (\tilde{h}-h)'M(\tilde{h}-h)$$

$$+ (\tilde{h}-\bar{h})'M(\tilde{h}-\bar{h})]/n \le O_p(K/n) + O_p(1)\textstyle\sum_{i=1}^n\hat{\tau}_i\psi_i^2[h_0(\hat{w}_i)-h_0(w_i)]^2/n$$

$$+ O_p(1)\textstyle\sum_{i=1}^n\hat{\tau}_i\psi_i^2[h_0(\hat{w}_i)-\hat{p}_i'\beta]^2/n = O_p(\Delta_h^2).$$

Then by the triangle inequality,

$$\{\int\tau(w)\psi(X)^2[\hat{h}(w)-h_0(w)]^2dF_0(X)\}^{1/2} \le \{\int\tau(w)\psi(X)^2[p^K(w)'(\hat{\beta}-\beta)]^2dF_0(X)\}^{1/2}$$

$$+ \{\int\tau(w)\psi(X)^2\{p^K(w)'\beta-h_0(w)]^2dF(w)\}^{1/2} = \|\hat{\beta}-\beta\| + O(K^{-\alpha}) \le O_p(\Delta_h),$$

giving the first conclusion. Also, by the triangle inequality,

$$|\hat{h}(w)-h_0(w)|_\delta \le |p^K(w)'\beta-h_0(w)|_\delta + |p^K(w)'(\hat{\beta}-\beta)|_\delta$$

$$\le O(K^{-\alpha}) + \zeta_\delta(K)\|\hat{\beta}-\beta\| = O_p(\zeta_\delta(K)\Delta_h). \quad \text{QED.}$$

To state Lemma A2, some additional notation is needed. Let $\hat{\Sigma}_1$, Σ_1, $\hat{Q}_1$, and Q_1 be as given in the body of the paper. Also, let $p_i = p^K(w_i)$, $r_i = p^L(z_i)$,

$$\hat{\Sigma} = \sum_{i=1}^{n}\hat{\tau}_i\psi_i^2\hat{p}_i\hat{p}_i'[y_i-\hat{h}(\hat{w}_i)]^2/n, \quad \Sigma = E[\tau_i\psi_i^2 p_i p_i' \mathrm{Var}(y_i|X_i)],$$

$$\hat{Q} = \hat{P}'\hat{P}/n, \quad Q = E[\tau_i\psi_i^2 p_i p_i'],$$

$$\hat{H} = \sum_{i=1}^{n}\hat{\tau}_i\psi_i\hat{p}_i\{[\partial\hat{h}(\hat{w}_i)/\partial w]'\partial w(X_i,\hat{\Pi}_i)/\partial\pi\otimes r_i'\}/n,$$

$$H = E[\tau_i\psi_i p_i\{[\partial h_0(w_i)/\partial w]'\partial w(X_i,\Pi_i)/\partial\pi\otimes r_i'\}].$$

$$\hat{V} = A\hat{Q}^{-1}(\hat{\Sigma} + \hat{H}\hat{Q}_1^{-1}\hat{\Sigma}_1\hat{Q}_1^{-1}\hat{H}')\hat{Q}^{-1}A', \quad V = AQ^{-1}(\Sigma + HQ_1^{-1}\Sigma_1Q_1^{-1}H')Q^{-1}A'.$$

For notational convenience, K and L subscripts for V are suppressed.

Assumption A2: i) $\|a(h)\| \le C|h|_\delta$; ii) $h_0(w)$ and $w(X,\pi)$ are twice continuously differentiable in w and π respectively, and the first and second derivatives are bounded; iii) either a) there are $\nu(w)$, β_K such that $a(h_0) = E[\tau(w)\psi(X)^2\nu(w)h_0(w)]$, $a(p_{kK}) = E[\tau(w)\psi(X)^2\nu(w)p_{kK}(w)]$, and $E[\tau(w)\psi(X)^2\|\nu(w)-\beta_K p^K(w)\|^2] \longrightarrow 0$, or; b) $a(h)$ is a scalar and there exists $(\tilde{\beta}_K)$ such that for $h_K(w) = p^K(w)'\tilde{\beta}_K$, $E[h_K(w)^2] \longrightarrow 0$, and $a(h_K)$ is bounded away from zero.

Under iii) b), let $d(X) = [\partial h_0(w)/\partial w]'\partial w(X,\Pi_0(z))/\partial\pi$, recall the definition of $\mathcal{L}$ as the set of limit points of $r^L(z)'\gamma_L$, and let $\rho(z)$ be the matrix of projections of elements of $\tau(w)\psi(X)\nu(w)d(X)$ on $\mathcal{L}$, and

$$\bar{V} = E[\tau(w)\psi(X)^2\nu(w)\nu(w)'\mathrm{Var}(y|X)] + E[\rho(z)\mathrm{Var}(x|z)\rho(z)'].$$

Lemma A2: If Assumptions A1 and A2 are satisfied, $Var(y|X)$ *is bounded away from zero,* $\xi(L)$ *and* $\zeta_d(K)$ *are bounded away from zero as* L *and* K *grow, and each of the following go to zero as* $n \longrightarrow \infty$: $\sqrt{n}K^{-\alpha}$, $\sqrt{n}L^{-\alpha_1}$, $\zeta_0(K)^2K^2/n$, $(K^2L+L^3)\zeta_1(K)^2/n$, $KL\zeta_0(K)^4\xi(L)^2/n$, $L^2\zeta_0(K)^2\xi(L)^4/n$, *and* $\zeta_0(K)^2\zeta_1(K)^2(LK+L^2)/n$, *then* $\hat{\theta} = \theta_0 + O_p(\zeta_d(K)/\sqrt{n})$ *and*

$$\sqrt{n}V^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,I), \quad \sqrt{n}\hat{V}^{1/2}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,I).$$

Furthermore, if Assumption A2, iii) a) is satisfied,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,\bar{V}), \quad \hat{V} \xrightarrow{p} \bar{V}.$$

Proof of Lemma A2: First, let

$$\Delta_\pi = L^{1/2}/\sqrt{n} + L^{-\alpha_1}, \quad \Delta_h = K^{1/2}/\sqrt{n} + K^{-\alpha} + \Delta_\pi, \quad \Delta_{Q1} = \xi(L)L^{1/2}/\sqrt{n},$$

$$\Delta_Q = [K^{1/2}\zeta_1(K) + \zeta_0(K)^2\xi(L)]\Delta_\pi + \zeta_0(K)K^{1/2}/\sqrt{n},$$

$$\Delta_H = [L^{1/2}\zeta_1(K) + \zeta_0(K)\xi(L)^2]\Delta_\pi + K^{1/2}\xi(L)/\sqrt{n}.$$

Note that by $\sqrt{n}K^{-\alpha} \longrightarrow 0$ and $\sqrt{n}K^{-\alpha_1} \longrightarrow 0$, $\Delta_\pi = (L^{1/2}/\sqrt{n})(1 + L^{-1/2}\sqrt{n}L^{-\alpha_1}) = O(L^{1/2}/\sqrt{n})$ and $\Delta_h = O(K^{1/2}/\sqrt{n} + L^{1/2}/\sqrt{n})$. Therefore, it follows from the convergence rate conditions and $\zeta_\delta(K)$ and $\xi(L)$ bounded away from zero that $\xi(L)^2L^2/n \le CL^2\zeta_0(K)^2\xi(L)^4/n \longrightarrow 0$, and

$$\text{(A.8)} \qquad K^{1/2}\Delta_Q = O([K\zeta_1(K) + K^{1/2}\zeta_0(K)^2\xi(L)]L^{1/2}/\sqrt{n} + \zeta_0(K)K/\sqrt{n})$$

$$= O(\{K^2L\zeta_1(K)^2 + KL\zeta_0(K)^4\xi(L)^2 + K^2\zeta_0(K)^2\}^{1/2}/\sqrt{n}) \longrightarrow 0.$$

$$L^{1/2}\Delta_H = O([L\zeta_1(K) + L^{1/2}\zeta_0(K)\xi(L)^2]L^{1/2}/\sqrt{n} + L^{1/2}K^{1/2}\xi(L)/\sqrt{n})$$

$$= O(\{L^3\zeta_1(K)^2 + L^2\zeta_0(K)^2\xi(L)^4 + LK\xi(L)^2\}^{1/2}/\sqrt{n}) \longrightarrow 0.$$

$$\sqrt{n}\zeta_0(K)\Delta_\pi^2 = O(\zeta_0(K)L/\sqrt{n}) \longrightarrow 0,$$

$$\zeta_0(K)L^{1/2}\zeta_1(K)\Delta_h = O(\{\zeta_0(K)^2L\zeta_1(K)^2(K+L)\}^{1/2}/\sqrt{n}) \longrightarrow 0.$$

It also follows from these results that $L^{1/2}\Delta_{Q1} \longrightarrow 0$, $\zeta_0(K)\Delta_h \longrightarrow 0$ and $\zeta_1(K)\Delta_h \longrightarrow 0$.

For simplicity the remainder of the proof will be given for the scalar $\Pi_0(z)$ case. Also, as in the proof of Lemma A1 it suffices to show the result with $p^K(w) = P^K(w)$, $r^L(z) = R^L(z)$, and Q and Q_1 equal to identity matrices. In this case, $V = A[\Sigma + H\Sigma_1H']A'$. Let F be a symmetric square root of V^{-1}. By $\sigma^2(X) = Var(y|X)$ bounded away from zero and $Q = I$, $\Sigma - CI$ is positive semidefinite. Therefore,

$$\text{(A.9)} \qquad \|FA\| = \{tr[FAA'F']\}^{1/2} \le \{tr[CFA\Sigma A'F']\}^{1/2} \le tr\{CFVF'\}^{1/2} = C.$$

Suppose that Assumption A2,- iii), a) is satisfied. Then $A = E[\tau(w)\psi(X)^2\nu(w)p^K(w)']$. Let $\nu_K(w) = Ap^K(w)$. By $Q = I$, $E[\tau(w)\psi(X)^2\|\nu(w)-\nu_K(w)\|^2] \le E[\tau(w)\psi(X)^2\|\nu(w)-\tilde{\beta}_Kp^K(w)\|^2] \longrightarrow 0$. Also let $d(X) = [\partial h_0(w)/\partial w]'\partial w(X,\Pi_0(z))/\partial\pi$, $b_{KL}(z) = E[\tau(w)\psi(X)d(X)\nu_K(w)r^L(z)']r^L(z)$, and $b_L(z) = E[\tau(w)\psi(X)d(X)\nu(w)r^L(z)']r^L(z)$. Since the mean square error (MSE) of a least squares projection is no greater than the MSE of the random variable being projected, $E[\|b_{KL}(z)-b_L(z)\|^2] \le E[\tau(w)\psi(X)^2d(X)^2\|\nu_K(w)-\nu(w)\|^2] \le CE[\tau(w)\psi(X)^2\|\nu_K(w)-\nu(w)\|^2] \longrightarrow 0$ as $K \longrightarrow \infty$. Furthermore, by Assumption A2, ii), b), $E[\|b_L(z)-\rho(z)\|^2] \longrightarrow 0$ as $L \longrightarrow \infty$. Then by boundedness of $\sigma^2(X) = Var(y|X)$ and $Var(x|z)$, and by the fact that MSE convergence implies convergence of second moments, it follows that

$$\text{(A.10)} \qquad V = E[\tau(w)\psi(X)^2\nu_K(w)\nu_K(w)'\sigma^2(X)] + E[b_{KL}(z)Var(x|z)b_{KL}(z)'] \longrightarrow \bar{V}.$$

This shows that F is bounded. Suppose that Assumption A2, iii) b) is satisfied. Then by the Cauchy-Schwartz inequality, $|a(h_K)| = |A\tilde{\beta}_K| \le \|A\|\|\tilde{\beta}_K\| = \|A\|(E[h_K(x)^2])^{1/2}$, so that $\|A\| \longrightarrow \infty$. Also, $V \ge A\Sigma A' \ge C\|A\|^2$, so F is also bounded under Assumption 5.

Next, by the proof of Lemma A1,

(A.11) $\quad \|\hat{Q}-I\| = O_p(\Delta_Q) = o_p(1), \quad \|\hat{Q}_1-I\| = O_p(\Delta_{Q1}) = o_p(1).$

Further, for $\bar{H} = \sum_{i=1}^{n}\hat{\tau}_i\psi_i\hat{p}_i d(X_i)r_i'/n$, similarly to Lemma A1,

(A.12) $\quad \|\bar{H}-H\| = O_p(\Delta_H) = o_p(1),$

Now, $\|\hat{Q}-I\| \xrightarrow{p} 0$ implies that the smallest eigenvalue of $\hat{Q}$ is bounded away from zero with probability approaching one, implying that the largest eigenvalue of $\hat{Q}^{-1}$ is $O_p(1)$, implying $\|B\hat{Q}^{-1}\| \le \|B\|O_p(1)$ for any matrix B. Therefore,

(A.13) $\quad \|FA(\hat{Q}^{-1}-I)\| \le \|FA(I-\hat{Q})\hat{Q}^{-1}\| \le \|FA\|\|\hat{Q}-I\|O_p(1) \xrightarrow{p} 0.$

Also, $\|FA\hat{Q}^{-1/2}\|^2 = \mathrm{tr}(FA\hat{Q}^{-1}A'F') \le CO_p(1)\|FA\|^2 = O_p(1).$

Next, let $\tilde{\beta}$ be such that $|p^K(\cdot)'\tilde{\beta} - h_0(\cdot)|_\delta = O_p(K^{-\alpha})$. Then

(A.14) $\quad \|\sqrt{n}Fa(p^{K\prime}\tilde{\beta} - h_0)\| \le C\sqrt{n}\|F\||p^K(\cdot)'\tilde{\beta} - h_0(\cdot)|_\delta = O_p(\sqrt{n}K^{-\alpha}) = o_p(1).$

Also, for $\tilde{h}_i = \hat{\tau}_i\psi_i h_0(\hat{w}_i)$ and $\tilde{h} = (\tilde{h}_1,\ldots,\tilde{h}_n)'$,

(A.15) $\quad \|FA\hat{Q}^{-1}\hat{P}'(\tilde{h}-\hat{P}\tilde{\beta})/\sqrt{n}\| \le C\sqrt{n}\|FA\hat{Q}^{-1}\hat{P}'/\sqrt{n}\|\sup_{\mathcal{W}}|p^K(w)'\tilde{\beta} - h_0(w)|$

$$\le C\sqrt{n}[\mathrm{tr}(FA\hat{Q}^{-1}A'F')]^{1/2}O(K^{-\alpha}) = O_p(\sqrt{n}K^{-\alpha}) = o_p(1).$$

Then by $A\tilde{\beta} = a(p^{K\prime}\tilde{\beta})$, $a(\hat{h}) = A\hat{\beta}$, eqs. (A.14) and (A.15), and the triangle inequality, for $h_i = \hat{\tau}_i\psi_i h_0(w_i)$ and $h = (h_1,\ldots,h_n)'$,

(A.16) $\quad \sqrt{n}F[a(\hat{h})-a(h_0)] = FA\hat{Q}^{-1}\hat{P}'\eta/\sqrt{n} + FA\hat{Q}^{-1}\hat{P}'(h-\tilde{h})/\sqrt{n} + o_p(1).$

Let $\Pi = (\Pi_1,\ldots,\Pi_n)'$, $u_i = x_i-\Pi_i$, $U = (u_1,\ldots,u_n)'$, γ be such that $\sup_{\mathcal{Z}}|\Pi_0(z)-r^L(z)'\gamma| = O(L^{-\alpha_1})$, and $d_i = d(X_i)$. By a second order mean-value expansion of each $h(\hat{w}_i)$ around w_i, and by eq. (B.0),

$$(A.17)\qquad FA\hat{Q}^{-1}\hat{P}'(h-\tilde{h})/\sqrt{n} = FA\hat{Q}^{-1}\textstyle\sum_{i=1}^{n}\hat{\tau}_i\psi_i\hat{p}_i d_i[\hat{\Pi}_i-\Pi_i]/\sqrt{n} + \hat{\rho} \;= FA\hat{Q}^{-1}\bar{H}\hat{Q}_1^{-1}R'U/\sqrt{n}$$

$$+ FA\hat{Q}^{-1}\bar{H}\hat{Q}_1^{-1}R'(\Pi-R'\gamma)/\sqrt{n} + FA\hat{Q}^{-1}\textstyle\sum_{i=1}^{n}\hat{\tau}_i\psi_i\hat{p}_i d_i[r_i'\gamma-\Pi_i]/\sqrt{n} + \hat{\rho}$$

$$\|\hat{\rho}\| \le C\sqrt{n}\|FA\hat{Q}^{-1/2}\|\zeta_0(K)\textstyle\sum_{i=1}^{n}\|\hat{\Pi}_i-\Pi_i\|^2/n = O_p(\sqrt{n}\zeta_0(K)\Delta_\pi^2) = o_p(1),$$

Also, by d_i bounded and $n\bar{H}\hat{Q}_1^{-1}\bar{H}'$ equal to the matrix sum of squares from the multivariate regression of $\hat{\tau}_i\psi_i\hat{p}_i d_i$ on r_i, $\bar{H}\hat{Q}_1^{-1}\bar{H}' \le \sum_{i=1}^{n}\hat{\tau}_i\psi_i^2\hat{p}_i\hat{p}_i' d_i^2/n \le C\hat{Q}$. Therefore,

$$(A.18)\qquad \|FA\hat{Q}^{-1}\bar{H}\hat{Q}_1^{-1}R'(\Pi-R'\gamma)/\sqrt{n}\| \le \|FA\hat{Q}^{-1}\bar{H}\hat{Q}_1^{-1}R'/\sqrt{n}\|\sqrt{n}\cdot\sup_{\mathcal{Z}}|\Pi_0(z)-r^L(z)'\gamma|$$

$$\le [\mathrm{trace}(FA\hat{Q}^{-1}\bar{H}\hat{Q}_1^{-1}\hat{Q}_1\hat{Q}_1^{-1}\bar{H}'\hat{Q}^{-1}A'F')]^{1/2}O(\sqrt{n}L^{-\alpha_1}) \le C\|FA\hat{Q}^{-1/2}\|O(\sqrt{n}L^{-\alpha_1})$$

$$= O_p(\sqrt{n}L^{-\alpha_1}) = o_p(1).$$

Similarly,

$$(A.19)\qquad \|FA\hat{Q}^{-1}\textstyle\sum_{i=1}^{n}\hat{\tau}_i\psi_i\hat{p}_i d_i[r_i'\gamma-\Pi_i]/\sqrt{n}\ \| \le C\|FA\hat{Q}^{-1/2}\|O(\sqrt{n}L^{-\alpha_1}) = o_p(1).$$

Next, note that $E[\|R'U/\sqrt{n}\|^2] = tr(\Sigma_1) \le Ctr(I_L) \le L$ by $E[u^2|z]$ bounded, so by the Markov inequality,

$$(A.20)\qquad \|R'U/\sqrt{n}\| = O_p(L^{1/2}).$$

Also, note that $\|FA\hat{Q}^{-1}\bar{H}\hat{Q}_1^{-1}\| \le O_p(1)\|FA\hat{Q}^{-1/2}\| = O_p(1)$. Therefore,

$$(A.21)\qquad \|FA\hat{Q}^{-1}\bar{H}(\hat{Q}_1^{-1}-I)R'U/\sqrt{n}\| \le \|FA\hat{Q}^{-1}\bar{H}\hat{Q}_1^{-1}\|\|\hat{Q}_1-I\|\|R'U/\sqrt{n}\| = O_p(L^{1/2}\Delta_{Q1}) = o_p(1).$$

By similar reasoning, and by eq. (A.8),

$$(A.22)\qquad \|FA\hat{Q}^{-1}(\bar{H}-H)R'U/\sqrt{n}\| \le \|FA\hat{Q}^{-1}\|\|\bar{H}-H\|\|R'U/\sqrt{n}\| = O_p(\Delta_H L^{1/2}) = o_p(1).$$

Noting also that HH' is the population matrix mean-square of the regression of

$\tau_i\psi_ip_id_i$ on r_i, so that $HH' \le CI$, it follows that $E[\|HR'U/\sqrt{n}\|^2] = tr(H\Sigma_1H') \le CK$. Therefore, $\|HR'U/\sqrt{n}\| = O_p(K^{1/2})$, and

(A.23) $$\|FA(\hat{Q}^{-1}-I)HR'U/\sqrt{n}\| \le \|FA\hat{Q}^{-1}\|\|(I-\hat{Q})\|\|HR'U/\sqrt{n}\| = O_p(\Delta_QK^{1/2}) = o_p(1).$$

Combining eqs. (A.16)-(A.23) and the triangle inequality gives

(A.24) $$FA\hat{Q}^{-1}\hat{P}'(h-\tilde{h})/\sqrt{n} = FAHR'U/\sqrt{n} + o_p(1).$$

Next, it follows as in the proof of Lemma A1 that $\|\hat{Q}^{-1/2}(\hat{P}-P)'\eta/\sqrt{n}\| = O_p(\zeta_0(K)^2\xi(L)\Delta_\pi + \zeta_1(K)^2\Delta_\pi^2) = o_p(1)$, implying

(A.25) $$\|FA\hat{Q}^{-1}(\hat{P}-P)'\eta/\sqrt{n}\| \le \|FA\hat{Q}^{-1/2}\|\|\hat{Q}^{-1/2}(\hat{P}-P)'\eta/\sqrt{n}\| = o_p(1).$$

Also, by $E[\eta|\tilde{X}] = 0$,

(A.26) $$E[\|FA(\hat{Q}^{-1}-I)P'\eta/\sqrt{n}\|^2|\tilde{X}] = tr(FA(I-\hat{Q})\hat{Q}^{-1}\Sigma\hat{Q}^{-1}(I-\hat{Q})A'F')$$

$$\le \|FA(I-\hat{Q})\hat{Q}^{-1}\|^2 \le \|FA\|^2\|I-\hat{Q}\|^2O_p(1),$$

so that $\|FA(\hat{Q}^{-1}-I)P'\eta/\sqrt{n}\| \xrightarrow{p} 0$. Then combining eqs. (A.25)-(A.26) with eq. (A.16) and the triangle inequality gives

(A.27) $$\sqrt{n}F[a(\hat{h})-a(h_0)] = FA(P'\eta/\sqrt{n} + HR'U/\sqrt{n}) + o_p(1).$$

Next, for any vector ϕ with $\|\phi\| = 1$ let $\phi'FA[\tau_i\psi_ip_i\eta_i + Hr_iu_i]/\sqrt{n} = Z_{in}$. Note that Z_{in} are i.i.d. across i for a given n, $E[Z_{in}] = 0$, and $Var(Z_{in}) = 1/n$. Furthermore, for any $\epsilon > 0$, $\|FA\| \le C$ and $\|FAH\| \le C\|FA\| \le C$ by $CI-HH'$ positive semidefinite, so that

$$nE[1(|Z_{in}| > \epsilon)Z_{in}^2] = n\epsilon^2E[1(|Z_{in}| > \epsilon)(Z_{in}/\epsilon)^2] \le n\epsilon^{-2}E[|Z_{in}|^4]$$

$$\le Cn\epsilon^{-2}\|\phi\|^4\{E[\|\tau_ip_i\|^4E[\eta_i^4|X_i]] + E[\|r_i\|^4E[u_i^4|z_i]]\}/n^2$$

$$\le Cn^{-1}\{\zeta_0(K)^2E[\|\tau_i\psi_ip_i\|^2] + \xi(L)^2E[\|r_i\|^2]\} = O([\zeta_0(K)^2K+ \xi(L)^2L]/n) = o(1).$$

Therefore, $\sqrt{n}F(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,I)$ follows by the Lindbergh-Feller theorem and equation (A.27).

As previously noted, $CI - H\Sigma H'$ is positive semi-definite, so that $V \le C\|A\|^2$. Suppose $a(h)$ is a scalar. It follows from the above proof that $A \ne 0$ for all n large enough. Note that for any β, the Cauchy-Schwartz inequality implies $|p^{K\prime}\beta|_\delta \le \zeta_\delta(K)\|\beta\|$, so that $\|A\|^2 = |a(Ap^K)| \le |Ap^K|_\delta \le \zeta_\delta(K)\|A\|$. Dividing by $\|A\|$ then gives $\|A\|^2 \le \zeta_\delta(K)^2$. Therefore, $\hat{\theta}-\theta_0 = O_p(V^{1/2}/\sqrt{n}) = O_p(\zeta_\delta(K)/\sqrt{n})$. This result for the scalar $a(h)$ covers the case of Assumption A2, iii), b). In the other case it follows from $V \longrightarrow \bar{V}$ that $\hat{\theta}-\theta_0 = O_p(1/\sqrt{n}) = O_p(\zeta_\delta(K)/\sqrt{n})$.

Next, by Lemma A1, $\max_{i\le n}|\hat{h}_i-\tilde{h}_i| = O_p(\zeta_0(K)\Delta_h) = o_p(1)$. Also, it follows by Theorem 1 of Newey (1997) that $\max_{i\le n}|\hat{\Pi}_i-\Pi_i| = O_p(\xi(L)\Delta_\pi) = o_p(1)$, so that by $h_0(w)$ and $w(X,\pi)$ Lipschitz in w and π respectively, $\max_{i\le}|\tilde{h}_i-h_i| = o_p(1)$. Hence, by the triangle inequality, $\max_{i\le n}|\hat{h}_i-h_i| = o_p(1)$. Note that by $\hat{\tau}_i = \hat{\tau}_i^2$, $\hat{\tau}_i(\hat{\eta}_i^2-\eta_i^2) = \hat{\tau}_i[-2\eta_i(\hat{h}_i-h_i)+(\hat{h}_i-h_i)^2] = \hat{\nu}_i$. Also, for $\hat{D} = FA\hat{Q}^{-1}\hat{P}'\mathrm{diag}\{1+|\eta_1|,\ldots,1+|\eta_n|\}\hat{P}\hat{Q}^{-1}A'F'/n$, $E[\hat{D}|\tilde{X}] \le CFA\hat{Q}^{-1}A'F = O_p(1)$. Let $\tilde{\Sigma} = \sum_{i=1}^n\hat{\tau}_i\hat{p}_i\hat{p}_i'\eta_i^2/n$. Then

$$\|FA\hat{Q}^{-1}(\hat{\Sigma}-\tilde{\Sigma})\hat{Q}^{-1}A'F'\| = \|FA\hat{Q}^{-1}\hat{P}'\mathrm{diag}\{\hat{\nu}_1,\ldots,\hat{\nu}_n\}\hat{P}\hat{Q}^{-1}A'F'/n\|$$

$$\le C\mathrm{tr}(\hat{D})\max_{i\le n}|\hat{h}_i-h_i| = O_p(1)o_p(1) = o_p(1).$$

Also, by $E[\eta_i^2|\tilde{X}]$ uniformly bounded, $E[\sum_{i=1}^n\eta_i^2|\hat{\tau}_i-\tau_i|/n|\tilde{X}] \le C\sum_{i=1}^n|\hat{\tau}_i-\tau_i|/n$ and $E[\sum_{i=1}^n\eta_i^2\|\hat{w}_i-w_i\|^2/n|\tilde{X}] \le C\sum_{i=1}^n\|\hat{w}_i-w_i\|^2/n$, so that $\sum_{i=1}^n\eta_i^2|\hat{\tau}_i-\tau_i|/n = O_p(\xi(L)\Delta_\pi)$ and $\sum_{i=1}^n\eta_i^2\|\hat{w}_i-w_i\|^2/n = O_p(\Delta_\pi^2)$. Then similarly to the proof of $\|\hat{Q}-I\| = O_p(\Delta_Q)$,

$$\|\tilde{\Sigma}-\Sigma\| \le \|\textstyle\sum_{i=1}^n(\hat{\tau}_i\hat{p}_i\hat{p}_i'-\tau_ip_ip_i')\eta_i^2/n\| + \|\textstyle\sum_{i=1}^n\tau_ip_ip_i'\eta_i^2/n - \Sigma\|$$

$$= O_p(\Delta_Q) + O_p(\{E[\tau_i\|p_i\|^4E[\eta_i^4|X_i]]/n\}^{1/2}) = O_p(\Delta_Q + \zeta_0(K)^2K/n) = o_p(1).$$

Therefore, $\|FA\hat{Q}^{-1}(\tilde{\Sigma}-\Sigma)\hat{Q}^{-1}A'F'\| \le \|FA\hat{Q}^{-1}\|^2\|\tilde{\Sigma}-\Sigma\| \xrightarrow{p} 0$. Furthermore, $\|B\Sigma\| \le$

$C\|B\|$ for any matrix B, by $CI - \Sigma$ positive semi-definite, so that $\|FA(\hat{Q}^{-1}\Sigma\hat{Q}^{-1}-\Sigma)A'F'\| \le \|FA(\hat{Q}^{-1}-I)\Sigma\hat{Q}^{-1}A'F'\| + \|FA\Sigma(\hat{Q}^{-1}-I)A'F'\| \le \|FA\hat{Q}^{-1}\|^2\|(I-\hat{Q})\Sigma\| + \|FA\Sigma\|\|I-\hat{Q}\|\|FA\hat{Q}^{-1}\| \le O_p(1)\|1-\hat{Q}\| + \|FA\|o_p(1)O_p(1) = o_p(1)$. Then by the triangle inequality, it suffices to show that $FA(\hat{Q}^{-1}\hat{H}\hat{Q}_1^{-1}\hat{\Sigma}_1\hat{Q}_1^{-1}\hat{H}'\hat{Q}^{-1} - H\Sigma_1H')A'F' \xrightarrow{p} 0$.

To show this last result, note first that it follows similarly to the argument for $\|\hat{\Sigma}-\Sigma\| \xrightarrow{p} 0$ that $\|\hat{\Sigma}_1-\Sigma_1\| \xrightarrow{p} 0$. Let $\hat{d}_i = \{[\partial\hat{h}(\hat{w}_i)/\partial w]'\partial w(X_i,\hat{\Pi}_i)/\partial\pi$ and $d_i = d(X_i)$. Then by the conclusion of Lemma A1, $\partial w(X,\pi)/\partial\pi$ bounded,

$$\textstyle\sum_{i=1}^{n}\hat{\tau}_i|\hat{d}_i-d_i|^2/n \le C(\sup_{\mathcal{W}}|\hat{h}-h_0|_1^2 + \sum_{i=1}^{n}\|\hat{\Pi}_i-\Pi_i\|^2/n) = O_p(\zeta_1(K)^2\Delta_h^2).$$

Therefore,

$$\|\hat{H}-\bar{H}\| \le C(\textstyle\sum_{i=1}^{n}\hat{\tau}_i\|\hat{p}_i\|^2\|r_i\|^2/n)^{1/2}(\sum_{i=1}^{n}\hat{\tau}_i|\hat{d}_i-d_i|^2/n)^{1/2} = O_p(\zeta_0(K)L^{1/2}\zeta_1(K)\Delta_h) = o_p(1).$$

Hence, by eq. (A.12) and the triangle inequality, $\|\hat{H}-H\| \xrightarrow{p} 0$. The conclusion then follows similarly to previous arguments. For example, by logic like that above,

$$\|FA\hat{Q}^{-1}\hat{H}\hat{Q}_1^{-1}(\hat{\Sigma}_1-\Sigma_1)\hat{Q}_1^{-1}\hat{H}'\hat{Q}^{-1}A'F'\| \le \|FA\hat{Q}^{-1}\hat{H}\hat{Q}_1^{-1}\|^2\|\hat{\Sigma}_1-\Sigma_1\|$$

$$\le tr(FA\hat{Q}^{-1}\hat{H}\hat{Q}_1^{-1}\hat{H}'\hat{Q}^{-1}A'F')o_p(1) = o_p(1).$$

It then follows by similar arguments and the triangle inequality that $F\hat{V}F' \xrightarrow{p} I$. In the case of Assumption A2, iii, b), where $a(h)$ is a scalar, it follows by taking a square root that that $\hat{V}^{-1/2}/V^{-1/2} \xrightarrow{p} 1$, so that $\sqrt{n}\hat{V}^{-1/2}(\hat{\theta}-\theta_0) = (\hat{V}^{-1/2}/V^{-1/2})\sqrt{n}V^{-1/2}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,I)$. In the other case the last conclusion follows similarly from $F \longrightarrow \bar{V}^{-1/2}$. QED.

Lemma A3: If v_i is i.i.d. and continuously distributed with bounded density and $\max_{i\le n}|\hat{v}_i-v_i| = O_p(\delta_n)$, $\delta_n \longrightarrow 0$, then $\sum_{i=1}^{n}|1(a\le\hat{v}_i\le b)-1(a\le v_i\le b)|/n = O_p(\delta_n)$.

Proof: By the density of v_i bounded, for any $\Delta_n > 0$, $\Delta_n \longrightarrow 0$, by the Markov inequality,

$$\sum_{i=1}^{n}[1(|v_i-a|\le\Delta_n)+1(|v_i-b|\le\Delta_n)]/n = O_p(\mathrm{Prob}(|v_i-a|\le\Delta_n)+\mathrm{Prob}(|v_i-b|\le\Delta_n)) = O_p(\Delta_n).$$

We also use the well known result that $Y_n = O_p(1)$ if and only if $\varepsilon_n Y_n \xrightarrow{p} 0$ for all positive sequences with $\varepsilon_n \longrightarrow 0$ slowly enough. Consider any positive sequence ε_n such that ε_n goes to zero slower than δ_n^2, i.e. $(\varepsilon_n)^{-1/2}\delta_n \longrightarrow 0$. Then $\nu_n = (\varepsilon_n)^{1/2} \longrightarrow 0$, so $\nu_n \max_{i\le n}|\hat{v}_i-v_i|/\delta_n \xrightarrow{p} 0$. It follows that $\max_{i\le n}|\hat{v}_i-v_i| \le \delta_n/\nu_n$ with probability approaching one (w.p.a.1) as $n \longrightarrow \infty$. Then w.p.a.1,

$$\varepsilon_n\sum_{i=1}^{n}|1(a\le\hat{v}_i\le b)-1(a\le v_i\le b)|/\delta_n n = \varepsilon_n\sum_{i=1}^{n}|1(a\le v_i+[\hat{v}_i-v_i]\le b)-1(a\le v_i\le b)|/\delta_n n$$

$$\le \varepsilon_n\sum_{i=1}^{n}[1(|v_i-a|\le\delta_n/\nu_n)+1(|v_i-b|\le\delta_n/\nu_n)]/\delta_n n = \varepsilon_n\delta_n^{-1}O_p(\delta_n/\nu_n) = O_p(\nu_n) = o_p(1). \quad \text{QED.}$$

Proof of Lemma 4.1: Because series estimators are invariant to nonsingular linear transformations of the approximating functions, a location and scale shift allows us to assume that $\mathcal{W} = [0,1]^{2d}$ and $\mathcal{Z} = [0,1]^{d_1}$. Also, in the polynomial case, the component powers can be replaced by polynomials of the same order that are orthonormal with respect to the uniform distribution on $[0,1]$, and in the spline case by a B-spline basis. In both cases the resulting vector of functions is a nonsingular linear combination of the original vector, because of the assumption that the order of the multi-index is increasing. Also, assume that the same operation is carried out on the first step approximating functions $r^L(z)$. For any nonnegative integer j,

$$\zeta_j(K) = \max_{|\mu|\le j}\sup_{w\in\mathcal{W}}\|\partial^{\mu}p^K(w)\|, \quad \xi_j(L) = \max_{|\mu|\le j}\sup_{z\in\mathcal{Z}}\|\partial^{\mu}p^L(z)\|.$$

Then it follows from Newey (1997) that in the polynomial case,

$$\zeta_j(K) \le CK^{1+2j}, \quad \xi(L) \le CL,$$

and in the spline case that

$$\zeta_j(K) \le K^{.5+j}, \quad \xi(L) \le L^{.5}.$$

It follows from these inequalities and Assumption 4 that

$$[\zeta_0(K)\zeta_1(K) + \zeta_0(K)^2\xi_0(L)]\Delta_\pi \longrightarrow 0, \quad \Delta_\pi = (L/n)^{1/2} + L^{-s_1/d_1},$$

so the conclusion follows by Lemma A1. QED.

Proof of Theorem 4.2: Without changing the notation let $F_0(w)$ denote the conditional distribution of w given $\tau(w) = 1$, let $\tilde{\lambda}(u) = \hat{\lambda}(u) - \int\hat{\lambda}(u)dF_0(u)$, and $\tilde{\lambda}_0(u) = \lambda_0(u) - \int\lambda_0(u)dF_0(u)$. For $w_1 = (x,z_1)$, note also that $\tilde{g}(w_1) = \hat{g}(w_1)-\int\hat{g}(w_1)dF_0(w_1)$ and $\tilde{g}_0(w_1) = g_0(w_1)-\int g_0(w_1)dF_0(w_1)$. Also by the density of w being bounded and bounded away from zero, so is the conditional density on $\mathcal{W}$, and the corresponding marginals. Therefore, for $\hat{\Delta} = \int\{\hat{h}(w)-h_0(w)\}dF_0(w)$

$$\int[\hat{h}(w)-h_0(w)]^2dF_0(w) \ge C\int[\hat{g}(w_1)+\hat{\lambda}(u)-g_0(w_1)-\lambda_0(u)]^2dF_0(w_1)dF_0(u)$$

$$= C\int[\tilde{g}(w_1)-\tilde{g}_0(w_1) + \tilde{\lambda}(u)-\tilde{\lambda}_0(u) + \hat{\Delta}]^2dF_0(w_1)dF_0(u)$$

$$= C\int[\tilde{g}(w_1)-\tilde{g}_0(w_1)]^2dF_0(w_1) + C\int[\tilde{\lambda}(u)-\tilde{\lambda}_0(u)]^2dF_0(u) + C\hat{\Delta}^2$$

$$\ge C\max\{\int[\tilde{g}(w_1)-\tilde{g}_0(w_1)]^2dF_0(w_1), \int[\tilde{\lambda}(u)-\tilde{\lambda}_0(u)]^2dF_0(u), \hat{\Delta}^2\}$$

where the product terms drop out by the construction of $\tilde{g}(u)$, etc., e.g. $\int$ $\int[\tilde{g}(w_1)-\tilde{g}_0(w_1)][\tilde{\lambda}(u)-\tilde{\lambda}_0(u)]dF_0(w_1)dF_0(u) = \int[\tilde{g}(w_1)-\tilde{g}_0(w_1)]dF_0(w_1)\cdot\int[\tilde{\lambda}(u)-\tilde{\lambda}_0(u)]dF_0(u) = 0$. QED.

Proof of Theorem 4.3: Follows from Lemma 4.1 by fixing the value of u at any point in $\mathcal{W}$. QED.

Proof of Theorem 5.1: For power series, using the inequalities in the proof of Theorem 4.1, it follows that

$$\zeta_0(K)^2K^2 \le CK^4, \quad (K^2L+L^3)\zeta_1(K)^2 \le C(K^2L+L^3)K^6, \quad KL\zeta_0(K)^4\xi(L)^2 \le CK^5L^3,$$

$$L^2\zeta_0(K)^2\xi(L)^4 \le CK^2L^6, \quad \zeta_0(K)^2\zeta_1(K)^2(LK+L^2) \le CK^8(KL + L^2),$$

and for splines that

$$\zeta_0(K)^2K^2 \le CK^3, \quad (K^2L+L^3)\zeta_1(K)^2 \le C(K^2L+L^3)K^3, \quad KL\zeta_0(K)^4\xi(L)^2 \le CK^3L^2,$$

$$L^2\zeta_0(K)^2\xi(L)^4 \le CKL^4, \quad \zeta_0(K)^2\zeta_1(K)^2(LK+L^2) \le CK^4(KL + L^2).$$

It then follows from the rate conditions of Theorem 4.2 that the rate conditions of Lemma A2 are satisfied. The conclusion then follows from Lemma A2, similarly to the proof of Theorem 4.1. QED.

References

Agarwal, G.G. and Studden, W.J. (1980): "Asymptotic Integrated Mean Square Error Using Least Squares and Bias Minimizing Spline," *Annals of Statistics* 8, 1307-1325.

Ahn, H. (1994) "Nonparametric Estimation of Conditional Choice Probabilities in a Binary Choice Model Under Uncertainty," working paper, Virginia Polytechnic Institute.

Andrews, D.W.K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Models," *Econometrica* 59, 307-345.

Andrews, D.W.K. and Whang, Y.J. (1990): "Additive Interactive Regression Models: Circumvention of the Curse of Dimensionality," *Econometric Theory* 6, 466-479.

Barzel, Y. (1973) "The Determination of Daily Hours and Wages," *Quarterly Journal of Economics* 87, 220-238.

Biddle, J. and Zarkin, G. (1989) "Choice among Wage-Hours Packages: An Empirical Investigation of Male Labor Supply," *Journal of Labor Economics* 7, 415-37.

Breiman, L. and Friedman, J.H. (1985) "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association.* 80 580-598.

Brown, B.W. and Walker, M.B. (1989): "The Random Utility Hypothesis and Inference in Demand Systems," *Econometrica* 47, 814-829.

Cox, D.D. 1988, "Approximation of Least Squares Regression on Nested Subspaces," *Annals of Statistics* 16, 713-732.

Hastie, T. and R. Tibshirani (1991): *Generalized Additive Models*, Chapman and Hall, London.

Linton, O. and Nielsen, J.B. (1995): "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika* 82, 93-100.

Linton, O., E. Mammen, and J. Nielsen (1997): "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions," preprint.

Moffitt, R. (1984) "The Estimation of a Joint Wage-Hours Labor Supply Model" *Journal of Labor Economics* 2, 550-66.

Newey, W.K. (1984) "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters* 14, 201-206.

Newey, W.K. (1994) "Kernel Estimation of Partial Means and a General 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 (1989) "Nonparametric Instrumental Variables Estimation," working paper, MIT Department of Economics.

Oi, W. (1962) "Labor as a Quasi Fixed Factor," *Journal of Political Economy* 70, 538-555.

Powell, M.J.D. (1981). *Approximation Theory and Methods.* Cambridge, England: Cambridge University Press.

Robinson, P. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica,* 56, 931-954.

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

Stone, C.J. (1982) "Optimal global rates of convergence for nonparametric regression," *Annals of Statistics* 10, 1040-1053.

Stone, C.J. (1985) "Additive regression and other nonparametric models," *Annals of Statistics.* 13 689-705.

Tjostheim, D. and Auestad, B. (1994): "Nonparametric Identification of Nonlinear Time Series: Projections," *Journal of the American Statistical Association* 89, 1398-1409.

Vella, F. (1991) "A Simple Two-Step Estimator for Approximating Unknown Functional Forms in Models with Endogenous Explanatory Variables," Australian National University, Department of Economics, working paper.

Vella, F. (1993) "Nonwage Benefits in a Simultaneous Model of Wages and Hours: Labor Supply Functions of Young Females," *Journal of Labor Economics* 11, 704-723

Table 1:

Estimate

	OLS	2sls	NP	NPIV
h	.000017	.000387	-.0094	-.01097
	(.5597)	(2.7447)	(2.5513)	(2.9170)
h^2	----	----	7.0151e-6	7.4577e-6
			(2.6363)	(2.7986)
h^3	----	----	-2.1707e-9	-2.0250e-9
			(2.6304)	(2.4640)
h^4	----	----	2.3692e-13	1.8929e-13
			(2.5537)	(2.0311)
r	----	-.000384	----	-.0005
		(2.6732)		(2.7643)
r^2	----	----	----	-6.6464e-8
				(.5211)
r^3				3.4978e-10
				(2.7516)

Notes:i) h denotes hours worked and r is the reduced form residual.

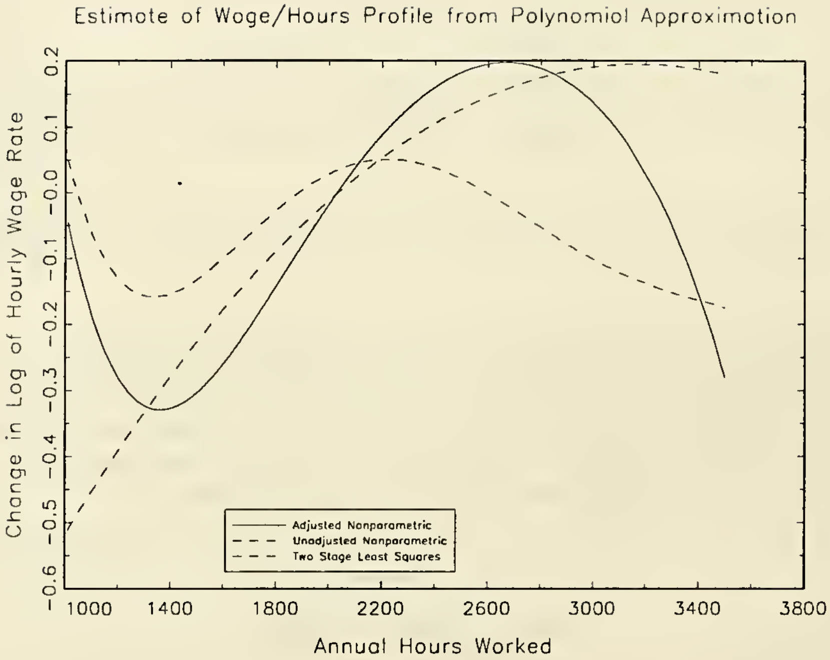
ii) The regressors in the wage equation included education dummies, union status, tenure, full-time work experience, black dummy and regional variables.

iii) The variables included in Z which are excluded from X are marital status, health status, presence of young children, rural dummy and non-labor income.

iv) Absolute value of t-statistics are reported in parentheses.

Table 2:

Specification	Cross Validation Value
h=0; r=0	185.567
h=1; r=1	183.158
h=2; r=2	183.738
h=3; r=3	182.899
h=4; r=4	182.882
h=5; r=5	183.128



Estimote of Woge/Hours Profile from Polynomiol Approximotion
Change in Log of Hourly Wage Rate
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
1000
1400
1800
2200
2600
3000
3400
3800
Annual Hours Worked
Adjusted Nonporametric
Unodjusted Nonporometric
Two Stoge Leost Squores

Figure 1:

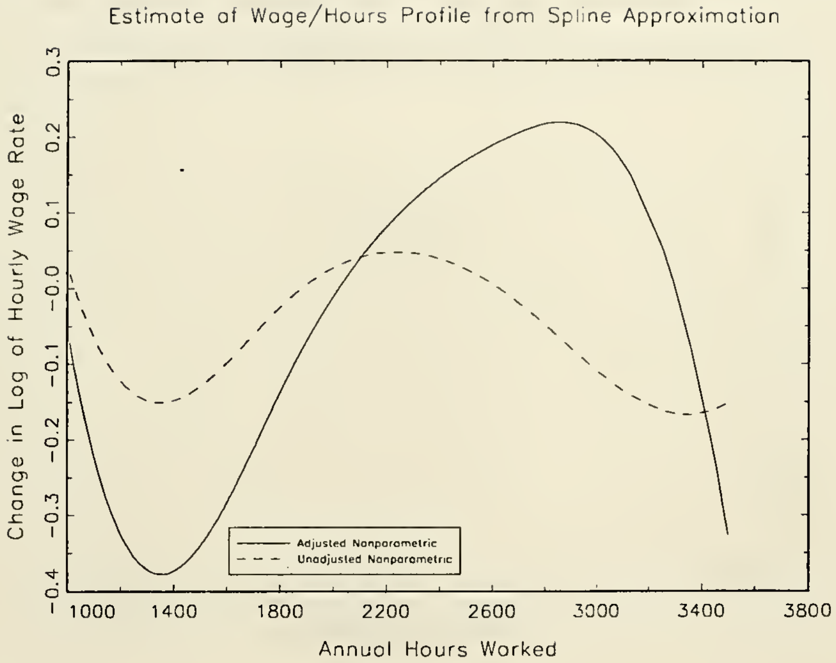

Estimate of Wage/Hours Profile from Spline Approximation
Change in Log of Hourly Wage Rate
0.3
0.2
0.1
−0.0
−0.1
−0.2
−0.3
−0.4
1000
1400
1800
2200
2600
3000
3400
3800
Annual Hours Worked
Adjusted Nonparametric
Unadjusted Nonparametric

Figure 2:

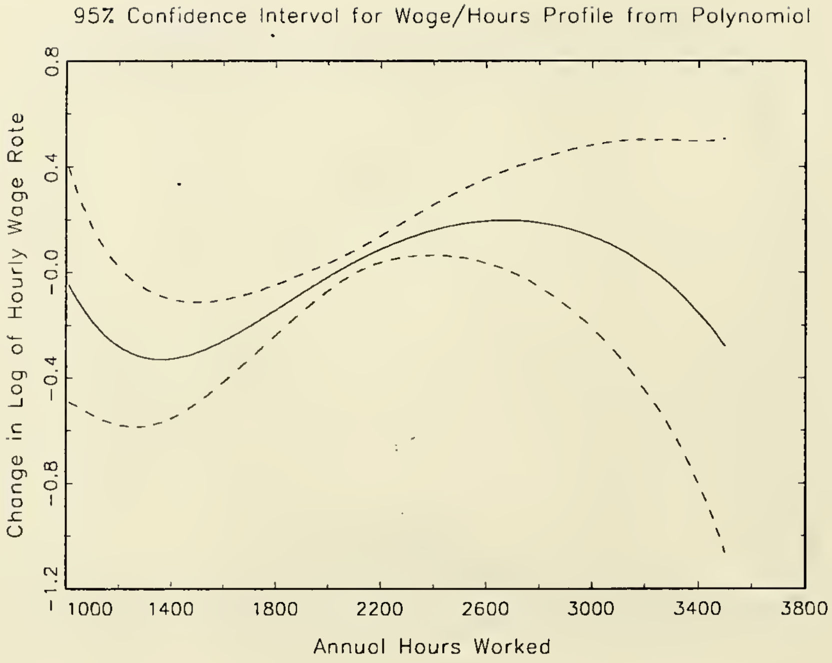

95% Confidence Intervol for Woge/Hours Profile from Polynomiol
Change in Log of Hourly Wage Rote
0.8
0.4
-0.0
-0.4
-0.8
-1.2
1000
1400
1800
2200
2600
3000
3400
3800
Annuol Hours Worked

Figure 3:

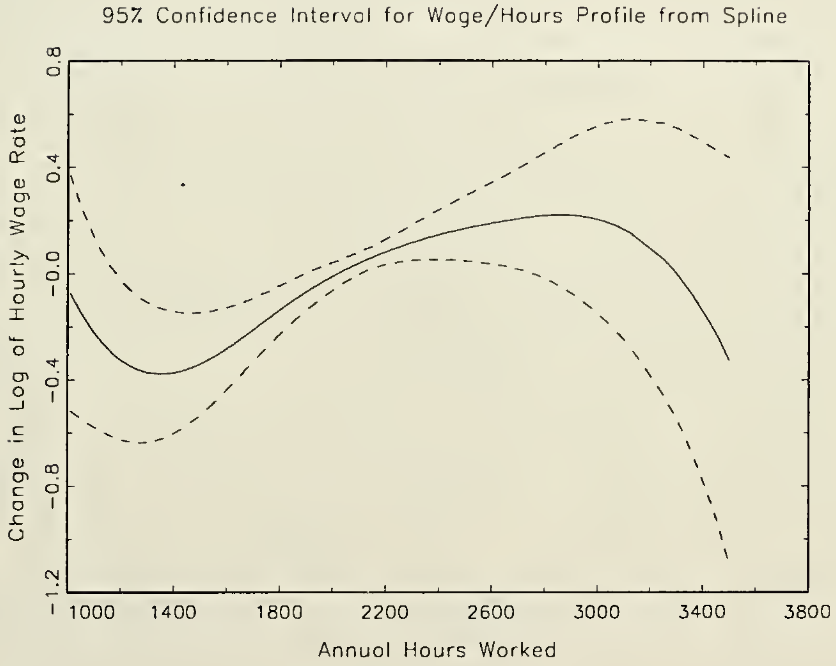

95% Confidence Interval for Wage/Hours Profile from Spline
Change in Log of Hourly Wage Rate
0.8
0.4
-0.0
-0.4
-0.8
-1.2
1000
1400
1800
2200
2600
3000
3400
3800
Annual Hours Worked

Figure 4: