Nonparametric Instrumental Variables Estimation of a Quantile Regression Model

Joel Horowitz
Sokbae Lee

NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION OF A QUANTILE REGRESSION MODEL

by

Joel L. Horowitz
Department of Economics
Northwestern University
Evanston, IL 60208
U.S.A.

and

Sokbae Lee
Department of Economics
University College London
London WC1E 6BT
United Kingdom

May 2006

Abstract

We consider nonparametric estimation of a regression function that is identified by requiring a specified quantile of the regression "error" conditional on an instrumental variable to be zero. The resulting estimating equation is a nonlinear integral equation of the first kind, which generates an ill-posed-inverse problem. The integral operator and distribution of the instrumental variable are unknown and must be estimated nonparametrically. We show that the estimator is mean-square consistent, derive its rate of convergence in probability, and give conditions under which this rate is optimal in a minimax sense. The results of Monte Carlo experiments show that the estimator behaves well in finite samples.

JEL Codes: C13, C31

Key words: Statistical inverse, endogenous variable, instrumental variable, optimal rate, nonlinear integral equation, nonparametric regression

NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION OF A QUANTILE REGRESSION MODEL

1. Introduction

Quantile regression models are increasingly important in applied econometrics. This paper is concerned with nonparametric estimation of a quantile regression model that has a possibly endogenous explanatory variable and is identified through an instrumental variable. Specifically, we estimate the function g in the model

(1.1) $\quad Y = g(X) + U$

(1.2) $\quad \boldsymbol{P}(U \leq 0 \mid W = w) = q$,

where Y is the dependent variable, X is an explanatory variable, W is an instrument for X, U is an unobserved random variable, and q is a known constant satisfying $0 < q < 1$. We do not assume that $\boldsymbol{P}(U \leq 0 \mid X = x)$ is independent of x. Therefore, the explanatory variable X may be endogenous in the quantile regression model (1.1)-(1.2). The function g is assumed to satisfy regularity conditions but is otherwise unknown. In particular, it does not belong to a known, finite-dimensional parametric family. The data are an iid random sample, $\{Y_i, X_i, W_i : i = 1.,,,.n\}$, of (Y, X, W). We present an estimator of g, derive its L_2 rate of convergence in probability, and provide conditions under which this rate is the fastest possible in a minimax sense.

Estimators of linear quantile regression models with endogenous right-hand side variables are described by Amemiya (1982), Powell (1983), Chen and Portnoy (1996), Chernozhukov and Hansen (2005a), and Ma and Koenker (2006). Chernozhukov and Hansen (2004) and Januszewski (2002) use such models in economic applications. Research on nonparametric estimation of quantile regression models is relatively recent. Chesher (2003) considers a triangular-array structure that is not necessarily additively separable and investigates nonparametric identification of derivatives of the unknown functions in that structure. Chernozhukov and Hansen (2005b) give conditions under which g in (1.1)-(1.2) is identified. Chernozhukov, Imbens, and Newey (2004) give conditions for consistency of a series estimator of g in a quantile instrumental-variables model that is not necessarily additively separable. The rate of convergence of their estimator is unknown.

There has also been research on nonparametric estimation of g in the model

(1.3) $\quad Y = g(X) + U$; $\quad \boldsymbol{E}(U \mid W = w) = 0$.

1

Here, as in (1.1)-(1.2), X is a possibly endogenous explanatory variable and W is an instrument for X, but the quantile restriction (1.2) is replaced by the condition $E(U \mid W = w) = 0$. Blundell and Powell (2003); Darolles, Florens, and Renault (2002); Florens (2003); Newey and Powell (2003); Newey, Powell, and Vella (1999); Carrasco, Florens, and Renault (2005); Horowitz (2005); and Hall and Horowitz (2005) discuss estimators of g in (1.3). Our estimator of g in (1.1)-(1.2) is related to Hall's and Horowitz's (2005) estimator of g in (1.3), but for reasons that will now be explained, estimating g in (1.1)-(1.2) presents problems that are different from those of estimating g in (1.3).

In both (1.1)-(1.2) and (1.3), the relation that identifies g is an operator equation,

$$(1.4) \quad \mathcal{T}g = \theta,$$

say, where $\mathcal{T}$ is a non-singular integral operator and θ is a function. $\mathcal{T}$ and θ are unknown but can be estimated consistently without difficulty. However, $\mathcal{T}^{-1}$ is discontinuous in both (1.1)-(1.2) and (1.3), so g cannot be estimated consistently by replacing $\mathcal{T}$ and θ in (1.4) with consistent estimators. This "ill-posed inverse problem" is familiar in the literature on integral equations. See, for example, Groetsch (1984); Engl, Hanke, and Neubauer (1996), and Kress (1999). It is dealt with by regularizing (that is, modifying) $\mathcal{T}$ to make the resulting inverse operator continuous. As in Darolles, Florens, and Renault (2002) and Hall and Horowitz (2005), we use Tikhonov (1963a, 1963b) regularization. This consists of choosing the estimator $\hat{g}$ to solve

$$(1.5) \quad \underset{\varphi \in \mathcal{G}}{\text{minimize}} : \left\| \hat{\mathcal{T}}\varphi - \hat{\theta} \right\|^2 + a_n \left\| \varphi \right\|^2,$$

where $\mathcal{G}$ is a parameter set (a set of functions in this case), $\hat{\mathcal{T}}$ and $\hat{\theta}$, respectively, are consistent estimators of $\mathcal{T}$ and θ; $\{a_n\}$ is a sequence of non-negative constants that converges to 0 as $n \to \infty$; and

$$\|v\|^2 = \int v(x)^2 dx$$

for any square integrable function v. In model (1.3), $\mathcal{T}$ and $\hat{\mathcal{T}}$ are linear operators, and the first-order condition for (1.5) is a linear integral equation (a Fredholm equation of the first kind). In (1.1)-(1.2), however, $\mathcal{T}$ and $\hat{\mathcal{T}}$ are nonlinear operators, and the first-order condition for (1.5) is a nonlinear integral equation.

The nonlinearity of $\mathcal{T}$ in (1.1)-(1.2) complicates the task of deriving the rate of convergence of $\hat{g}$ to g. In contrast to the situation in many other nonlinear estimation

problems, using a Taylor series expansion to make a linear approximation to the first-order condition is unattractive because, as a consequence of the ill-posed inverse problem, the approximation error dominates other sources of estimation error and controls the rate of convergence of $\hat{g}$ unless very strong assumptions are made about the probability density function of (Y, X, W). To avoid making these assumptions, we use a modified version of a method developed by Engl, Hanke, and Neubauer (1996, Theorem 10.7) to derive the rate of convergence of $\hat{g}$. This method works directly from the objective function in (1.5), rather than from the first-order condition. It requires us to assume that the norm of the second Fréchet derivative of $\mathcal{T}$ is not too large (see assumption 6 in Section 3.2). It is an open question whether the same rate of convergence of $\hat{g}$ can be attained without making this assumption or one that is similar to it.

The remainder of the paper is organized as follows. Section 2 presents the estimator for the special case in which X and W are scalars. Section 3 gives conditions under which the estimator is consistent, obtains its rate of convergence, and shows that the rate of convergence is the fastest possible (in a minimax sense) under our assumptions. Section 4 extends the results of sections 2 and 3 to a multivariate model in which X is a vector that may have some exogeneous components. Section 5 presents the results of a Monte Carlo investigation of the estimator's finite-sample performance. Concluding comments are given in Section 6. The proofs of theorems are in the appendix, which is Section 7.

2. The Estimator

This section describes our estimator of g in (1.1)-(1.2) when X and W are scalars. Let $F_{Y|XW}$ denote the distribution function of Y conditional on (X, W). We assume that the conditional distribution of Y has a probability density function, $f_{Y|XW}$, with respect to Lebesgue measure. We also assume that (X, W) has a probability density function with respect to Lebesgue measure, f_{XW}. Let f_W denote the marginal density of W. Define $F_{YXW} = F_{Y|XW} f_{XW}$ and $f_{YXW} = f_{Y|XW} f_{XW}$. Assume without loss of generality that the support of (X, W) is contained in $[0,1]^2$.

It follows from (1.1)-(1.2) that

(2.1) $\quad \int_0^1 F_{YXW}[g(x), x, w] dx = q f_W(w)$

3

for almost every w. We assume that (2.1) uniquely identifies g up to a set of x values whose Lebesgue measure is 0. Chernozhukov and Hansen (2005b, Theorem 4) give sufficient conditions for this assumption to hold.

Now define the operator $\mathcal{T}$ on $L_2[0,1]$ by

$$(2.2) \quad (\mathcal{T}\varphi)(w) = \int_0^1 F_{YXW}[\varphi(x), x, w]dx,$$

where φ is any function in $L_2[0,1]$. Then (2.1) is equivalent to the operator equation

$$\mathcal{T}g = qf_W.$$

Identifiability of g is equivalent to assuming that $\mathcal{T}$ is invertible. Thus,

$$(2.3) \quad g = q\mathcal{T}^{-1}f_W.$$

However, $\mathcal{T}^{-1}$ is discontinuous because the Fréchet derivative of $\mathcal{T}$ is a completely continuous operator and, therefore, has an unbounded inverse if f_{YXW} is "well behaved". Consequently, g cannot be estimated consistently by replacing $\mathcal{T}$ and f_W with consistent estimators in (2.3). As was explained in Section 1, we use Tikhonov regularization to deal with this problem.

We now describe the version of problem (1.5) that we solve to obtain the regularized estimator of g. We assume that f_{YXW} has $r > 0$ continuous derivatives with respect to any combination of its arguments.[1] Let K denote a continuously differentiable kernel function whose support is $[-1,1]$, is symmetrical about 0, and that satisfies

$$(2.4) \quad \int_{-1}^{1} v^j K(v)dv = \begin{cases} 1 \text{ if } j = 0 \\ \\ 0 \text{ if } 1 \le j \le \max(1, r-1). \end{cases}$$

Let $h > 0$ denote a bandwidth parameter, and define $K_h(v) = K(v/h)$ for any $v \in [-1,1]$. We estimate f_W, f_{YXW}, and F_{YXW}, respectively, by

$$\hat{f}_W(w) = \frac{1}{nh}\sum_{i=1}^{n} K_h(w - W_i)$$

$$\hat{f}_{YXW}(y, x, w) = \frac{1}{nh^3}\sum_{i=1}^{n} K_h(y - Y_i)K_h(x - X_i)K_h(w - W_i),$$

and

$$\hat{F}_{YXW}(y, x, w) = \int_{-\infty}^{y} \hat{f}_{YXW}(v, x, w)dv.$$

Define the operator $\hat{\mathcal{T}}$ by

$$(\hat{\mathcal{T}}\varphi)(w) = \int_0^1 \hat{F}_{YXW}[\varphi(x), x, w] dx.$$

for any $\varphi \in L_2[0,1]$. Let $\|\varphi\|$ denote the L_2 norm of φ. Define $\mathcal{G} = \{\varphi \in L_2[0,1] : \|\varphi\|^2 \le M\}$ for some constant $M < \infty$. Our estimator of g is any solution to the problem

$$(2.5) \qquad \hat{g} = \arg\min_{\varphi \in \mathcal{G}} S_n(\varphi) \equiv \left\| \hat{\mathcal{T}}\varphi - q\hat{f}_W \right\|^2 + a_n \|\varphi\|^2.$$

Under our assumptions, a function $\hat{g}$ that minimizes S_n always exists, though it may not be unique (Bissantz, Hohage, and Munk. 2004).

3. Theoretical Properties

This section gives conditions under which the estimator $\hat{g}$ of Section 2 is consistent, derives the rate at which $\|\hat{g} - g\|^2$ converges in probability to 0, and gives conditions under which this is the fastest possible rate, in a minimax sense.

3.1 Consistency

This section gives conditions under which $E\|\hat{g} - g\|^2 \to 0$ as $n \to \infty$. Define $\delta_n = h^{2r} + (nh)^{-1}$. Make the following assumptions.

Assumption 1: (a) The function g is identified. That is, (2.1) specifies $g(x)$ uniquely up to a set of x values whose Lebesgue measure is 0. (b) $\|g\|^2 \le M$ for some constant $M < \infty$.

Assumption 2: f_{YXW} has $r > 0$ continuous derivatives with respect to any combination of its arguments. These derivatives and f_{YXW} are bounded in absolute value by M.

Assumption 3: As $n \to \infty$, $a_n \to 0$, $\delta_n \to 0$, and $\delta_n / a_n \to 0$.

Assumption 4: The kernel function K is supported on $[-1,1]$, continuously differentiable, symmetrical about 0, and satisfies (2.4).

In assumption 2, r need not be an integer. If r is not an integer, then the assumption means that

$$\left| D^{[r]}[f_{YXW}(y_1, x_1, w_1) - f_{YXW}(y_2, x_2, w_2)] \right| \le M \left\| (y_1, x_1, w_1) - (y_2, x_2, w_2) \right\|^{r-[r]},$$

where $[r]$ is the integer part of r, $D^{[r]} f_{YXW}$ denotes any order $[r]$ derivative of f_{YXW}, and $\|z_1 - z_2\|$ is the Euclidean distance between the points z_1 and z_2.

Assumptions 1(b) and 2 are standard in nonparametric estimation. Assumptions 3 and 4 are satisfied by a wide range of choices of h, a_n and K. The choice of h in applications is discussed in Section 3.2.

Theorem 1: Let assumptions 1-4 hold. Then

(3.1) $\lim_{n \to \infty} E \|\hat{g} - g\|^2 = 0$.

Result (3.1) implies that $\|\hat{g} - g\|^2$ converges in probability to 0 as $n \to \infty$. The next section obtains the rate of convergence.

3.2 Rate of Convergence

In model (1.3), where $\mathcal{T}$ is a linear operator, the source of the ill-posed inverse problem is that sequence of singular values of $\mathcal{T}$ (or, equivalently, eigenvalues of $\mathcal{T}^* \mathcal{T}$, where $\mathcal{T}^*$ is the adjoint of $\mathcal{T}$) converges to 0. Consequently, the rate at which $\|\hat{g} - g\|^2$ converges to 0 depends on the rate of convergence of the singular values (or eigenvalues). See Hall and Horowitz (2005). In (1.1)-(1.2), where $\mathcal{T}$ is nonlinear, the source of the ill-posed inverse problem is convergence to 0 of the singular values of the Fréchet derivative of $\mathcal{T}$ at g. Denote this derivative by T_g. The rate of convergence of $\|\hat{g} - g\|^2$ in (1.1)-(1.2) depends on the rate of convergence of the singular values of T_g or, equivalently, of the eigenvalues of $T_g^* T_g$, where T_g^* is the adjoint of T_g. Accordingly, the regularity conditions for our rate of convergence result are framed in terms of the spectral representation of $T_g^* T_g$.

It is easy to show that the Frechet derivative of $\mathcal{T}$ at g is the operator T_g defined by

$$(T_g \varphi)(w) = \int_0^1 f_{YXW}[g(x), x, w] \varphi(x) dx .$$

The adjoint operator is defined by

$$(T_g^* \varphi)(w) = \int_0^1 f_{YXW}[g(w), w, x] \varphi(x) dx .$$

We assume that $T_g^* T_g$ is non-singular, so its eigenvalues are strictly positive. Let $\{\lambda_j, \phi_j : j = 1, 2, ...\}$ denote the eigenvalues and orthonormal eigenvectors of $T_g^* T_g$ ordered so that $\lambda_1 \geq \lambda_2 \geq ... > 0$. Under our assumptions, $T_g^* T_g$ is a completely continuous operator, so $\{\phi_j\}$ forms a basis for $L_2[0,1]$. Therefore, we may write

(3.2) $g(x) = \sum_{j=1}^{\infty} b_j \phi_j(x)$,

where the Fourier coefficients b_j are given by

$$b_j = \int_0^1 g(x) \phi_j(x) dx.$$

We make the following additional assumptions.

<u>Assumption 5</u>: (a) There are constants $\alpha > 1$, $\beta > 1$, and $C_0 < \infty$ such that $\beta - 1/2 \le \alpha < 2\beta - 1$, $j^{-\alpha} \le C_0 \lambda_j$, and $|b_j| \le C_0 j^{-\beta}$ for all $j \ge 1$. (b) Moreover, $r \ge \max[(2\beta + \alpha - 1)/2, (3\beta + \alpha - 1/2)/(\alpha + 1)]$.

<u>Assumption 6</u>: There is a finite constant $L > 0$ such that

(3.3) $\left\| \mathcal{T}(g_1) - \mathcal{T}(g_2) - T_{g_2}(g_1 - g_2) \right\| \le 0.5 L \left\| g_1 - g_2 \right\|^2$

for any $g_1, g_2 \in L_2[0,1]$ and

(3.4) $\sum_{j=1}^{\infty} \dfrac{b_j^2}{\lambda_j} < 1/L$.

<u>Assumption 7</u>: The tuning parameters h and a_n satisfy $h = C_h n^{-1/(2r+1)}$ and $a_n = C_a n^{-\alpha/(2\beta+\alpha)}$, where C_h and C_a are positive, finite constants.

In assumption 5(a), α characterizes the severity of the ill-posed inverse problem. As α increases, the problem becomes more severe and, consequently, the fastest possible rate of convergence of any estimator of g decreases. The parameter β characterizes the complexity of g in the sense that if β is large, then g can be well approximated by the finite-dimensional parametric model that is obtained by truncating the series on the right-hand side of (3.2) at $j = J$ for some small integer J. A finite-dimensional g can be estimated with a $n^{-1/2}$ rate of convergence, so the rate of convergence of $\left\| \hat{g} - g \right\|^2$ increases as β increases.

Assumption (5a) places tight restrictions on the rates of convergence of λ_j and b_j. This is unavoidable for obtaining optimal rates of convergence with Tikhonov regularization. The restrictions are not needed for consistent estimation, as Theorem 1 shows. In linear inverse problems, the restrictions can sometimes be relaxed by using other forms of regularization. See, for example, Carrasco, Florens, and Renault (2005) and Bissantz, Hohage, Munk, and Ruymgaart (2006). However, there has been little research on the application of non-Tikhonov methods to

nonlinear inverse problems, especially when $\mathcal{T}$ in (1.4) is unknown and must be estimated. Much additional research will be needed to determine the extent to which such methods are useful in nonlinear problems with an unknown $\mathcal{T}$.

Together, assumptions 2 and 5(a) imply that $T_g^* T_g$ is a completely continuous linear operator, so its eigenvectors form a basis for $L_2[0,1]$. Assumption 2 implies that inequality (3.3) holds for some $L < \infty$, so (3.3) amounts to the definition of L. Inequality (3.4) restricts the norm of the second Fréchet derivative of $\mathcal{T}$. A sufficient condition for (3.4) is

$$\sup_{y,x,w} \left| \partial f_{YXW}(y,x,w) / \partial y \right| < \left(\sum_{j=1}^{\infty} \frac{b_j^2}{\lambda_j} \right)^{-1} .$$

Restrictions similar to (3.4) are well-known in the theory of nonlinear integral equations. It is an open question whether an estimator of g that has our rate of convergence can be achieved without making an assumption similar to (3.4). Assumptions 2 and 5(b) imply that f_W, and $\mathcal{T}\varphi$ for any $\varphi \in L_2[0,1]$ can be estimated with a rate of convergence in probability of $O_p[n^{-(\beta-1/2+\alpha/2)/(2\beta+\alpha)}]$ and that f_{YXW} can be estimated with a rate of convergence of $O_p[n^{-(\beta-1/2)/(2\beta+\alpha)}]$. These rates are needed to obtain our rate of convergence of $\hat{g}$. Under assumption 7, h has the optimal rate of convergence for nonparametric estimation of an r-times differentiable density function of a scalar argument. Accordingly, h can be chosen in applications by, say, cross-validation applied to $\hat{f}_W$.

Let $\mathcal{H} = \mathcal{H}(M, C_0, \alpha, \beta, L)$ be the set of distributions of (Y, X, W) that satisfy assumptions 1, 2, 5, and 6 with fixed values of M, C_0, α, β, and L. Our rate-of-convergence result is given by the following theorem.

<u>Theorem 2</u>: Let assumptions 1-2 and 4-7 hold. Then

$$\lim_{D \to \infty} \limsup_{n \to \infty} \sup_{H \in \mathcal{H}} P_H \left[\| \hat{g} - g \|^2 > D n^{-(2\beta-1)/(2\beta+\alpha)} \right] = 0 .$$

As expected, the rate of convergence of $\| \hat{g} - g \|^2$ decreases as α increases (the ill-posed inverse problem becomes more severe) and increases as β increases (g increasingly resembles a finite-dimensional parametric model).

The next theorem shows that the convergence rate in Theorem 2 is optimal in a minimax sense under our assumptions. Let $\{\tilde{g}_n\}$ denote any sequence of estimators of g satisfying $\|\tilde{g}_n\|^2 \le M$ for some $M < \infty$.

Theorem 3: Let assumptions 1-2 and 4-7 hold with $\alpha \geq 2$. Then for every finite $D > 0$

$$\liminf_{n \to \infty} \sup_{H \in \mathcal{H}} P_H \left[\left\| \tilde{g}_n - g \right\|^2 > Dn^{-(2\beta-1)/(2\beta+\alpha)} \right] > 0.$$

Our proof of Theorem 3 requires $\alpha \geq 2$. It is an open question whether the rate of convergence $n^{-(2\beta-1)/(2\beta+\alpha)}$ is optimal when $1 < \alpha < 2$.

4. Multivariate Model

This section extends the results of Sections 2 and 3 to a multivariate model in which X is a vector that may have some exogenous components. We rewrite model (1.1)-(1.2) as

(4.1) $Y = g(X,Z) + U$

(4.2) $P(U \leq 0 \mid W = w, Z = z) = q$,

where Y is the dependent variable, $X \in \mathbb{R}^\ell$ ($\ell \geq 1$) is a vector of possibly endogenous explanatory variables, $Z \in \mathbb{R}^m$ ($m \geq 0$) is a vector of exogenous explanatory variables, and $W \in \mathbb{R}^\ell$ is a vector of instruments for X, U is an unobserved random variable, and q is a known constant satisfying $0 < q < 1$. If $m = 0$, then Z is not in the model. The problem is to estimate g nonparametrically from data consisting of the independent random sample $\{Y_i, X_i, W_i, Z_i : i = 1, ..., n\}$.

The estimator is obtained by applying the method of Section 2 after conditioning on Z. To do this, let $K_{\ell,h}(v) = \prod_{j=1}^{\ell} K_h(v^j / h)$, where v^j is the j'th component of the ℓ-vector v. Define $K_{m,h}(v)$ for an m-vector v similarly. Let f_{WZ} and f_{YXWZ}, respectively, denote the probability density functions of (W, Z) and (Y, X, W, Z). Define

$$F_{YXWZ}(y, x, w, z) = \int_{-\infty}^{y} f_{YXWZ}(v, x, w, z) dv.$$

Define the following kernel estimators of f_{WZ}, f_{YXWZ}, and F_{YXWZ}:

$$\hat{f}_{WZ}(w, z) = \frac{1}{nh^{\ell+m}} \sum_{i=1}^{n} K_{\ell,h}(w - W_i) K_{m,h}(z - Z_i),$$

$$\hat{f}_{YXWZ}(w, z) = \frac{1}{nh^{2\ell+m+1}} \sum_{i=1}^{n} K_h(y - Y_i) K_{\ell,h}(x - X_i) K_{\ell,h}(w - W_i) K_{m,h}(z - Z_i),$$

and

$$\hat{F}_{YXWZ}(y, x, w, z) = \int_{-\infty}^{y} \hat{f}_{YXWZ}(v, x, w, z) dv.$$

9

For each $z \in [0,1]^m$, define the operators $\mathcal{T}_z$ and $\hat{\mathcal{T}}_z$ on $L_2[0,1]^\ell$ by

$$(\mathcal{T}_z \varphi_z)(w) = \int_{[0,1]^\ell} F_{YXWZ}[\varphi_z(x), x, w, z] dx$$

and

$$(\hat{\mathcal{T}}_z \varphi_z)(w) = \int_{[0,1]^\ell} \hat{F}_{YXWZ}[\varphi_z(x), x, w, z] dx,$$

where φ_z is any function on $L_2[0,1]^\ell$.

The function g satisfies

(4.3) $(\mathcal{T}_z g)(w, z) = q f_{WZ}(w, z)$.

The function $g(\cdot, z)$ is identified if (4.3) has a unique solution (up to a set of x values of Lebesgue measure 0) for the specified z. Define $\mathcal{G} = \{\varphi \in L_2[0,1]^\ell : \|\varphi\|^2 \leq M\}$ for some constant $M < \infty$. For each $z \in [0,1]^m$, the estimator of $g(\cdot, z)$ is any solution to the problem

(4.4) $\hat{g}(\cdot, z) = \arg \min_{\varphi_z \in \mathcal{G}} \left\{ \int_{[0,1]^\ell} [(\hat{\mathcal{T}}_z \varphi_z)(w) - q \hat{f}_{WZ}(w, z)]^2 dw + a_n \int_{[0,1]^\ell} \varphi_z(w)^2 dw \right\}$.

4.1 Consistency

This section gives conditions under which $E\|\hat{g} - g\|^2 \to 0$ as $n \to \infty$ in model (4.1)-(4.2). Define $\delta_n = h^{2r} + (nh^{\ell+m})^{-1}$. Make the following assumptions, which are extensions of the assumptions of Section 3.1.

<u>Assumption 1′</u>:   (a) The function g is identified.   (b) $\int_{[0,1]^\ell} g(x, z)^2 dx \leq M$ for each $z \in [0,1]^m$ and some constant $M < \infty$.

<u>Assumption 2′</u>: f_{YXWZ} has $r > 0$ continuous derivatives with respect to any combination of its arguments. These derivatives and f_{YXWZ} are bounded in absolute value by M.

<u>Assumption 3′</u>:   As $n \to \infty$, $a_n \to 0$, $\delta_n \to 0$, and $\delta_n / a_n \to 0$.

<u>Theorem 4</u>:   Let assumptions 1′-3′ and 4 hold. Then for each $z \in [0,1]^m$,

$$\lim_{n \to \infty} E \int_{[0,1]^\ell} [\hat{g}(x, z) - g(x, z)]^2 dx = 0.$$

4.2 Rate of Convergence

As when X and W are scalars, the rate of convergence of $\hat{g}$ in the multivariate model depends on the rate of convergence of the singular values of the Fréchet derivative of $\mathcal{T}_z$.

10

Accordingly, let T_{gz} denote the Fréchet derivative of $\mathcal{T}_z$ at g, and let T_{gz}^* denote the adjoint of T_{gz}. T_{gz} and T_{gz}^*, respectively, are the operators defined by

$$(T_{gz}\varphi_z)(w) = \int_{[0,1]^\ell} f_{YXWZ}[g(x,z),x,w,z]\varphi_z(x)dx$$

and

$$(T_{gz}^*\varphi_z)(w) = \int_{[0,1]^\ell} f_{YXWZ}[g(w,z),w,x,z]\varphi_z(x)dx.$$

Assume that for each $z \in [0,1]^\ell$, $T_{gz}^* T_{gz}$ is non-singular. Let $\{(\lambda_{zj}, \phi_{zj}): j = 1,2,...\}$ denote the eigenvalues and eigenvectors of $T_{gz}^* T_{gz}$ ordered so that $\lambda_{z1} \geq \lambda_{z2} \geq ... > 0$. The eigenvectors $\{\phi_{zj}\}$ form a complete, orthonormal basis for $L_2[0,1]^\ell$. Thus, for each $z \in L_2[0,1]^\ell$ we can write

$$g(x,z) = \sum_{j=1}^{\infty} b_{zj}\phi_{zj}(x),$$

where

$$b_{zj} = \int_{[0,1]^\ell} g(x,z)\phi_z(x)dx.$$

Now make the following additional assumptions, which extend those of Section 3.2.

<u>Assumption 5′</u>: (a) There are constants $\alpha > 1$, $\beta > 1$, and $C_0 < \infty$ such that $\beta - 1/2 \leq \alpha < 2\beta - 1$, $j^{-\alpha} \leq C_0 \lambda_{zj}$, and $|b_{zj}| \leq C_0 j^{-\beta}$ uniformly in $z \in [0,1]^m$ for all $j \geq 1$. (b) Moreover, $r \geq \max\{[\ell(2\beta + \alpha - 1) - m]/2, r^*\}$, where r^* is the largest root of the equation
$$4[(\alpha + 1)/(2\beta + \alpha)]r^2 - 2[(\ell + m)(2\beta - 1)/(2\beta + \alpha) + \ell + 1 - m]r - (\ell + 1)m = 0.$$
<u>Assumption 6′</u>: There is a finite constant $L > 0$ such that

$$\left\| \mathcal{T}_z(g_1) - \mathcal{T}_z(g_2) - T_{g_2 z}(g_1 - g_2) \right\| \leq 0.5L\|g_1 - g_2\|^2$$

for any $g_1, g_2 \in L_2[0,1]$ uniformly in $z \in [0,1]^m$ and

$$\sum_{j=1}^{\infty} \frac{b_{zj}^2}{\lambda_{zj}} < 1/L.$$

uniformly in $z \in [0,1]^m$.

<u>Assumption 7′</u>: The tuning parameters h and a_n satisfy $h = C_h n^{-1/(2r+\ell+m)}$ and $a_n = C_a n^{-\tau\alpha/(2\beta+\alpha)}$, where $\tau = 2r/(2r+m)$ and C_h and C_a are positive, finite constants.

11

Let $\mathcal{H}_M = \mathcal{H}_M(M, C_0, \alpha, \beta, L, r, \ell, m)$ be the set of distributions of (Y, X, Z, W) that satisfy assumptions 1′, 2′, 5′, and 6′ with fixed values of M, C_0, α, β, L, r, ℓ, and m. The multivariate extension of Theorems 2 and 3 is

Theorem 5: Let assumptions 1′-2′, 4, and 5′-7′ hold. Then for each $z \in [0,1]^m$,

$$(4.5) \quad \lim_{D \to \infty} \limsup_{n \to \infty} \sup_{H \in \mathcal{H}_M} P_H \left\{ \int_{[0,1]^\ell} [\hat{g}(x,z) - g(x,z)]^2 dx > Dn^{-\tau(2\beta-1)/(2\beta+\alpha)} \right\} = 0.$$

If, in addition,

$$(4.6) \quad [\ell(2\beta + \alpha - 1) - m]/2 \geq r^*,$$

then for each $z \in [0,1]^m$

$$(4.7) \quad \liminf_{n \to \infty} \sup_{H \in \mathcal{H}_M} P_H \left\{ \int_{[0,1]^\ell} [\hat{g}(x,z) - g(x,z)]^2 dx > Dn^{-\tau(2\beta-1)/(2\beta+\alpha)} \right\} > 0.$$

If $m = 0$, then (4.6) simplifies to $\alpha \geq 1 + 1/\ell$. The rate of convergence in Theorem 5 is the same as that in Theorem 3 if $\ell = 1$ and $m = 0$. The theorem shows that increasing m decreases the rate of convergence of $\hat{g}$ for any fixed r. This is the familiar curse of dimensionality of nonparametric estimation. In addition, assumption 5′ implies that as ℓ increases, r must also increase to maintain the rate of convergence $n^{-\tau(2\beta-1)/(2\beta+\alpha)}$. This is a form of the curse of dimensionality that is associated with the endogenous explanatory variable X.

5. Monte Carlo Experiments

This section reports the results of Monte Carlo simulations that illustrate the finite-sample performance of the estimator that is obtained by solving (2.5). Samples of (Y, X, W) were generated from the model

$$Y = g(X) + U,$$

where

$$g(x) = 2^{1/2} \sum_{j=1}^{\infty} (-1)^{j+1} j^{-6} \sin(j\pi x)$$

and (U, X, W) is sampled from the distribution whose density is

$$f_{UXW}(u,x,w) = C\sum_{j=1}^{\infty} a_j \sin\left[\frac{j\pi}{2}(u+1)\right]\sin(j\pi x)\sin(j\pi w)$$

$$+0.15\sum_{j=1}^{\infty} d_j \sin[j\pi(u+1)]\sin(2j\pi x)\sin(j\pi w)$$

for $(U,X,W)\in[-1,1]\times[0,1]^2$. In this density,

$$a_j = \begin{cases} j^{-3} & \text{if } j=1,3,5,... \\ 0 & \text{otherwise,} \end{cases}$$

$d_j = j^{-10}$, and

$$C = \left[\left(\frac{16}{\pi^3}\right)\sum_{j=1}^{\infty}\frac{a_j}{j^3}\right]^{-1}.$$

The density is cumbersome algebraically, but it is convenient for computations and has a conventional shape. This is illustrated in Figures 1-3, which show graphs of the density of W, the density of X conditional on W, and the density of U conditional on (X,W). For computational purposes, the infinite series were truncated at $j=100$.

The kernel function used for density estimation is $K(x) = (15/16)(1-x^2)^2$ for $|x|\leq1$. The estimates of g were computed by using the Levenberg-Marquardt method (Engl, Hanke and Neubauer 1996, p. 285). The starting function was obtained by carrying out a cubic quantile regression of Y on X without controlling for endogeneity.

Each experiment consisted of estimating g at the 99 points $x=0.01,0.02,...,0.99$. The experiments were carried out in GAUSS using GAUSS pseudo-random number generators. There were 500 Monte Carlo replications in each experiment. Experiments were carried out using sample sizes $n=200$ and $n=800$.

The results are summarized in Table 1 and illustrated graphically in Figures 4-5, respectively, for $n=200$ and $n=800$. For each sample size, there are two values of the bandwidth parameter, h (0.5 and 0.8), and two values of the regularization parameter, a_n (0.5 and 1). For estimation of f_{YXW} the bandwidth $2h$ is used in the Y direction and h is used in the X and W directions, because the standard deviation of Y is about double those of X and W. The results in Table 1 show that the empirical integrated variance and integrated mean-square error (IMSE) decrease as the sample size increases. The bias does not decrease if a_n remains fixed. This is not surprising because a_n is the main source of estimation bias. Figures 4-5 show

13

$g(x)$ (dashed line), the Monte Carlo approximation to $E[\hat{g}(x)]$ (solid line), and the estimates, $\hat{g}$, whose IMSEs are the 25th, 50th, and 75th percentiles of the IMSEs of the 500 Monte Carlo replications. The figures show, not surprisingly, that $\hat{g}$ is biased but that its shape is similar to that of g.

6. Conclusions

This paper has presented a nonparametric instrumental variables estimator of a quantile regression model, derived the estimator's rate of mean-square convergence in probability, and given conditions under which this rate is the fastest possible in a minimax sense. The estimator's finite-sample performance has been illustrated by a small set of Monte Carlo experiments.

Several topics remain for future research. The problem of deriving the asymptotic distribution of $\hat{g}$ appears to be quite difficult. In contrast to the situation in many other nonlinear estimation problems, asymptotic normality cannot be obtained by using a Taylor series approximation to linearize the first-order condition for (2.5). This is because, as was explained in Section 1, the ill-posed inverse problem causes the error of the linear approximation to dominate other sources of estimation error unless very strong assumptions are made about the distribution of (Y, X, W). This problem does not arise with the mean-regression estimator of Hall and Horowitz (2005) and Horowitz (2005), because the first-order condition in the mean regression is a linear equation for $\hat{g}$. Other topics for future research include determining whether the rate of convergence of Theorem 2 is optimal when $1 < \alpha < 2$ (or when $r < r^*$ in Theorem 5) and finding a method to choose the regularization parameter a_n in applications.

7. Mathematical Appendix: Proofs of Theorems

This appendix provides proofs of Theorems 1-3. Theorems 4-5 can be proved by following the same steps after conditioning on Z.

6.1 Proof of Theorem 1

The proof is a modification of the proof of Theorem 2 of Bissantz, Hohage, and Munk (2004). By (2.5),

$$(6.1) \quad \left\| \hat{T}(\hat{g}) - q\hat{f}_W \right\|^2 + a_n \left\| \hat{g} \right\|^2 \leq \left\| \hat{T}(g) - q\hat{f}_W \right\|^2 + a_n \left\| g \right\|^2 .$$

In addition, $E \left\| \hat{T}(g) - T(g) \right\|^2 = O(\delta_n)$ and $E \left\| \hat{f}_W - f_W \right\|^2 = O(\delta_n)$. Therefore, by assumption 3,

$$E\left\|\hat{\mathcal{T}}(g) - q\hat{f}_W\right\|^2 \le 2E\left\|\hat{\mathcal{T}}(g) - \mathcal{T}(g)\right\|^2 + 2qE\left\|\hat{f}_W - f_W\right\|^2$$

$$= O(\delta_n).$$

Combining this result with (6.1) gives

$$E\left\|\hat{\mathcal{T}}(\hat{g}) - q\hat{f}_W\right\|^2 + a_n E\|\hat{g}\|^2 \le C\delta_n + a_n\|g\|^2$$

for some constant $C < \infty$ and all sufficiently large n. Therefore, by assumption 3,

$$\limsup_{n\to\infty} E\|\hat{g}\|^2 \le \|g\|^2.$$

Note, in addition, that $E\left\|\hat{\mathcal{T}}(\hat{g}) - q\hat{f}_W\right\|^2 \to 0$ as $n \to \infty$. Moreover, assumptions 1(b) and 2

imply that $\mathcal{T}$ is weakly closed. Consistency now follows from arguments identical to those used

to prove Theorem 2 of Bissantz, Hohage, and Munk (2004, p. 1777). Q.E.D.

6.2 Proof of Theorem 2

Assumptions 1-2 and 4-7 hold throughout this section. Let $\langle\cdot,\cdot\rangle$ denote the inner product

in $L_2[0,1]$. Define $\omega_g = (T_g T_g^*)^{-1} T_g g$ and

(6.2) $\qquad \tilde{g} = g - a_n (T_g^* T_g + a_n I)^{-1} T_g^* \omega_g$,

where I is the identity operator. Observe that by (3.4),

(6.3) $\qquad L\|\omega_g\| < 1$.

Let $\hat{r}$ and $\tilde{r}$ be Taylor series remainder terms with the properties that

(6.4) $\qquad \mathcal{T}(\hat{g}) = q f_W + T_g(\hat{g} - g) + \hat{r}$

and

(6.5) $\qquad \mathcal{T}(\tilde{g}) = q f_W + T_g(\tilde{g} - g) + \tilde{r}$,

where $q f_W = \mathcal{T}(g)$. By (3.3), $\|\hat{r}\| \le (L/2)\|\hat{g} - g\|^2$, and $\|\tilde{r}\| \le (L/2)\|\tilde{g} - g\|^2$.

<u>Lemma 6.1</u>: For any $g \in \mathcal{G}$,

$$\left(1 - L\|\omega_g\|\right)\|\hat{g} - g\|^2 \le \|\tilde{g} - g\|^2 + a_n^{-1}\left\|\hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}) + \tilde{r} + qf_W - q\hat{f}_W\right\|^2$$

$$+ \left\langle 2T_g(\tilde{g} - g) + a_n\omega_g, \omega_g \right\rangle + a_n^{-1}\left\|T_g(\tilde{g} - g)\right\|^2$$

$$+ 2\left\langle qf_W - q\hat{f}_W, \omega_g \right\rangle + 2a_n^{-1}\left\langle \tilde{r} + qf_W - q\hat{f}_W, T_g(\tilde{g} - g) \right\rangle$$

$$+ 2\left\langle \hat{\mathcal{T}}(\hat{g}) - \mathcal{T}(\hat{g}), \omega_g \right\rangle + 2a_n^{-1}\left\langle \hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}), T_g(\tilde{g} - g) \right\rangle.$$

Proof: By (6.5),

(6.6) $\quad \left\|\hat{\mathcal{T}}(\tilde{g}) - q\hat{f}_W\right\|^2 = \left\|\hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}) + \tilde{r} + qf_W - q\hat{f}_W\right\|^2 + \left\|T_g(\tilde{g} - g)\right\|^2$

$$+ 2\left\langle \tilde{r} + qf_W - q\hat{f}_W, T_g(\tilde{g} - g) \right\rangle + 2\left\langle \hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}), T_g(\tilde{g} - g) \right\rangle.$$

Also,

(6.7) $\quad \left\|\hat{\mathcal{T}}(\hat{g}) - q\hat{f}_W\right\|^2 = \left\|\hat{\mathcal{T}}(\hat{g}) - q\hat{f}_W + a_n\omega_g\right\|^2 + a_n^2\left\|\omega_g\right\|^2$

$$- 2a_n\left\langle \hat{\mathcal{T}}(\hat{g}) - \mathcal{T}(\hat{g}), \omega_g \right\rangle - 2a_n\left\langle \hat{\mathcal{T}}(\hat{g}) - qf_W, \omega_g \right\rangle$$

$$- 2a_n\left\langle qf_W - q\hat{f}_W, \omega_g \right\rangle - 2a_n\left\|\omega_g\right\|^2.$$

Moreover,

(6.8) $\quad \left\langle \tilde{g} - g, g \right\rangle = \left\langle \tilde{g} - g, T_g^*\omega_g \right\rangle$

$$= \left\langle T_g(\tilde{g} - g), \omega_g \right\rangle.$$

By (6.4),

(6.9) $\quad \left\langle \hat{g} - g, g \right\rangle = \left\langle \hat{g} - g, T_g^*\omega_g \right\rangle$

$$= \left\langle T_g(\hat{g} - g), \omega_g \right\rangle$$

$$= \left\langle \mathcal{T}(\hat{g}) - qf_W, \omega_g \right\rangle - \left\langle \hat{r}, \omega_g \right\rangle.$$

By (2.5),

(6.10) $\quad \left\|\hat{\mathcal{T}}(\hat{g}) - q\hat{f}_W\right\|^2 + a_n^2\|\hat{g}\|^2 \le \left\|\hat{\mathcal{T}}(\tilde{g}) - q\hat{f}_W\right\|^2 + a_n^2\|\tilde{g}\|^2.$

Rearranging and expanding terms in (6.10) gives

$$(6.11) \quad \|\hat{g} - g\|^2 \le a_n^{-1}\left[\left\|\hat{\mathcal{T}}(\tilde{g}) - q\hat{f}_W\right\|^2 - \left\|\hat{\mathcal{T}}(\hat{g}) - q\hat{f}_W\right\|^2\right] + \|\tilde{g} - g\|^2$$

$$+ 2\langle \tilde{g} - g, g\rangle - 2\langle \hat{g} - g, g\rangle.$$

Combining (6.11) with (6.6)-(6.9) gives

$$(6.12) \quad \|\hat{g} - g\|^2 \le 2\langle \hat{r}, \omega_g\rangle - a_n^{-1}\left\|\hat{\mathcal{T}}(\hat{g}) - q\hat{f}_W + a_n\omega_g\right\|^2 + \|\tilde{g} - g\|^2$$

$$+ a_n^{-1}\left\|\hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}) + \tilde{r} + qf_W - q\hat{f}_W\right\|^2 + \langle 2T_g(\tilde{g} - g) + a_n\omega_g, \omega_g\rangle$$

$$+ a_n^{-1}\left\|T_g(\tilde{g} - g)\right\|^2 + 2\langle qf_W - q\hat{f}_W, \omega_g\rangle + 2a_n^{-1}\langle \tilde{r} + qf_W - q\hat{f}_W, T_g(\tilde{g} - g)\rangle$$

$$+ 2\langle \hat{\mathcal{T}}(\hat{g}) - \mathcal{T}(\hat{g}), \omega_g\rangle + 2a_n^{-1}\langle \hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}), T_g(\tilde{g} - g)\rangle.$$

The lemma follows by noting that the second term on the right-hand side of (6.12) is non-positive and that $2\langle \hat{r}, \omega_g\rangle \le L\|\omega_g\|\|\hat{g} - g\|^2$. Q.E.D.

Lemma 6.2: For any $g \in \mathcal{G}$,

$$(6.13) \quad a_n^{-1}\left\|\hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}) + \tilde{r} + qf_W - q\hat{f}_W\right\|^2$$

$$\le 4a_n^{-1}\left(\left\|\hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g})\right\|^2 + \frac{L^2}{4}\|\tilde{g} - g\|^4 + \left\|qf_W - q\hat{f}_W\right\|^2\right),$$

$$(6.14) \quad \langle 2T_g(\tilde{g} - g) + a_n\omega_g, \omega_g\rangle + a_n^{-1}\left\|T_g(\tilde{g} - g)\right\|^2 = a_n^3\left\|(T_gT_g^* + a_nI)^{-1}\omega_g\right\|^2,$$

$$(6.15) \quad \left|2\langle qf_W - q\hat{f}_W, \omega_g\rangle + 2a_n^{-1}\langle \tilde{r} + qf_W - q\hat{f}_W, T_g(\tilde{g} - g)\rangle\right|$$

$$\le L\|\omega_g\|\|\tilde{g} - g\|^2 + 2a_n\left\|qf_W - q\hat{f}_W\right\|\left\|(T_gT_g^* + a_nI)^{-1}\omega_g\right\|$$

$$+ La_n\|\tilde{g} - g\|^2\left\|(T_gT_g^* + a_nI)^{-1}\omega_g\right\|,$$

and

17

(6.16) $\left| 2\left\langle \hat{\mathcal{T}}(\hat{g}) - \mathcal{T}(\hat{g}), \omega_g \right\rangle + 2a_n^{-1}\left\langle \hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}), T_g(\tilde{g} - g) \right\rangle \right|$

$$\leq 2a_n \left\| \hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}) \right\| \left\| (T_g T_g^* + a_n I)^{-1} \omega_g \right\| + 2\left\| \omega_g \right\| \left\| \left[\hat{\mathcal{T}}(\hat{g}) - \mathcal{T}(\hat{g}) \right] - \left[\hat{\mathcal{T}}(g) - \mathcal{T}(g) \right] \right\|$$

$$+ 2\left\| \omega_g \right\| \left\| \left[\hat{\mathcal{T}}(\tilde{g}) - \mathcal{T}(\tilde{g}) \right] - \left[\hat{\mathcal{T}}(g) - \mathcal{T}(g) \right] \right\|.$$

Proof: Inequality (6.13) follows from (6.5) and the relation

$$\| A + B + C \|^2 \leq 4\left(\| A \|^2 + \| C \|^2 + \| C \|^2 \right)$$

for any functions A, B, and C.

To show (6.14), note that

(6.17) $a_n(T_g T_g^* + a_n I)^{-1} = I - T_g(T_g^* T_g + a_n I)^{-1} T_g^*.$

By (6.2)

(6.18) $T(\tilde{g} - g) = -a_n T_g(T_g^* T_g + a_n I)^{-1} T_g^* \omega_g.$

It follows from (6.17) and (6.18) that

(6.19) $a_n \omega_g + T(\tilde{g} - g) = a_n^2 (T_g T_g^* + a_n I)^{-1} \omega_g.$

Taking the squares of the norms of both sides of (6.19) and expanding the term on the left-hand side yields

(6.20) $a_n^2 \left\| \omega_g \right\|^2 + \left\| T_g(\tilde{g} - g) \right\|^2 + 2\left\langle a_n \omega_g, T_g(\tilde{g} - g) \right\rangle = a_n^4 \left\| (T_g T_g^* + a_n I)^{-1} \omega_g \right\|^2.$

Then (6.14) follows by dividing both sides of (6.20) by a_n.

We now turn to (6.15). First note that

(6.21) $\left\langle \tilde{r} + qf_W - q\hat{f}_W, T_g(\tilde{g} - g) \right\rangle = \left\langle \tilde{r} + qf_W - q\hat{f}_W, T_g(\tilde{g} - g) + a_n \omega_g \right\rangle$

$$-\left\langle \tilde{r} + qf_W - q\hat{f}_W, a_n \omega_g \right\rangle.$$

It follows from (6.19) and (6.21) that

(6.22) $2\left\langle qf_W - q\hat{f}_W, \omega_g \right\rangle + 2a_n^{-1}\left\langle \tilde{r} + qf_W - q\hat{f}_W, T_g(\tilde{g} - g) + a_n \omega_g \right\rangle$

$$= 2a_n \left\langle \tilde{r} + qf_W - q\hat{f}_W, (T_g T_g^* + a_n I)^{-1} \omega_g \right\rangle - 2\left\langle \tilde{r}, \omega_g \right\rangle.$$

Then (6.15) follows by applying the Cauchy-Schwarz and triangle inequalities to (6.22).

Now we prove (6.16). Observe that by (6.19) and algebra like that yielding (6.21),

18

$$2\left\langle \hat{T}(\hat{g}) - T(\hat{g}), \omega_g \right\rangle + 2a_n^{-1}\left\langle \hat{T}(\tilde{g}) - T(\tilde{g}), T_g(\tilde{g} - g) \right\rangle$$

$$= 2a_n\left\langle \hat{T}(\tilde{g}) - T(\tilde{g}), (T_g T_g^* + a_n I)^{-1}\omega_g \right\rangle + 2\left\langle \left[\hat{T}(\hat{g}) - T(\hat{g})\right] - \left[\hat{T}(\tilde{g}) - T(\tilde{g})\right], \omega_g \right\rangle,$$

which yields (6.16) by the Cauchy-Schwarz and triangle inequalities. Q.E.D.

<u>Lemma 6.3</u>: The following relations hold uniformly over $H \in \mathcal{H}$.

(a) $\left\| \tilde{g} - g \right\|^2 = O[n^{-(2\beta-1)/(2\beta+\alpha)}]$,

(b) $a_n^{-1}\left\| \tilde{g} - g \right\|^4 = O[n^{-(2\beta-1)/(2\beta+\alpha)}]$,

(c) $a_n^3\left\| (T_g T_g^* + a_n I)^{-1}\omega_g \right\|^2 = O[n^{-(2\beta-1)/(2\beta+\alpha)}]$,

(d) $a_n\left\| \tilde{g} - g \right\|^2 \left\| (T_g T_g^* + a_n I)^{-1}\omega_g \right\| = O[n^{-(2\beta-1)/(2\beta+\alpha)}]$,

(e) $a_n\left\| q f_W - q\hat{f}_W \right\| \left\| (T_g T_g^* + a_n I)^{-1}\omega_g \right\| = O_p[n^{-(2\beta-1)/(2\beta+\alpha)}]$,

(f) $a_n\left\| \hat{T}(\tilde{g}) - T(\tilde{g}) \right\| \left\| (T_g T_g^* + a_n I)^{-1}\omega_g \right\| = O_p[n^{-(2\beta-1)/(2\beta+\alpha)}]$,

(g)   There are random variables $\Delta_n = O_p[n^{-(\beta-1/2)/(2\beta+\alpha)}]$ and $\Gamma_n = o_p(1)$ such that

$$\left\| \left[\hat{T}(\hat{g}) - T(\hat{g})\right] - \left[\hat{T}(g) - T(g)\right] \right\| \le \Delta_n \left\| \hat{g} - g \right\| + \Gamma_n \left\| \hat{g} - g \right\|^2,$$

(h) $\left\| \left[\hat{T}(\tilde{g}) - T(\tilde{g})\right] - \left[\hat{T}(g) - T(g)\right] \right\| = O_p[n^{-(2\beta-1)/(2\beta+\alpha)}]$.

<u>Proof</u>: To prove (a), note that by (6.2) and $T_g^*\omega_g = g$,

$$\tilde{g}(x) - g(x) = -a_n \sum_{j=1}^{\infty} \frac{1}{\lambda_j + a_n}\phi_j(x)\left\langle \phi_j, T_g^*\omega_g \right\rangle$$

$$= -a_n \sum_{j=1}^{\infty} \frac{b_j}{\lambda_j + a_n}\phi_j(x).$$

Therefore,

$$\left\| \tilde{g} - g \right\|^2 = a_n^2 \sum_{j=1}^{\infty} \frac{b_j^2}{(\lambda_j + a_n)^2}$$

$$= O[n^{-(2\beta-1)/(2\beta+\alpha)}],$$

where the second line follows from arguments identical to those used to prove equation (6.4) of Hall and Horowitz (2005). This proves (a). It follows from (a) that

$$a_n^{-1} \|\tilde{g} - g\|^4 = O[n^{-(2\beta-1)/(2\beta+\alpha)}]$$

whenever $\alpha < 2\beta - 1$, thereby proving (b).

We now turn to (c). Define $\psi_j = T_g \phi_j / \|T_g \phi_j\|$. Use $\omega_g = (T_g T_g^*)^{-1} T_g g$ and the singular

value decomposition $T_g^* \psi_j = \lambda_j^{1/2} \phi_j$ to obtain

$$(T_g T_g^* + a_n I)^{-1} \omega_g = \sum_{j=1}^{\infty} \frac{1}{\lambda_j (\lambda_j + a_n)} \psi_j \langle \psi_j, T_g g \rangle$$

$$= \sum_{j=1}^{\infty} \frac{1}{\lambda_j (\lambda_j + a_n)} \psi_j \langle T_g^* \psi_j, g \rangle$$

$$= \sum_{j=1}^{\infty} \frac{1}{\lambda_j (\lambda_j + a_n)} \psi_j \langle \lambda_j^{1/2} \phi_j, g \rangle$$

$$= \sum_{j=1}^{\infty} \frac{b_j}{\lambda_j^{1/2} (\lambda_j + a_n)} \psi_j.$$

Therefore,

(6.23) $\quad \left\| (T_g T_g^* + a_n I) \omega_g \right\|^2 = \sum_{j=1}^{\infty} \frac{b_j^2}{\lambda_j (\lambda_j + a_n)^2}$

$$= O[a_n^{(2\beta-3\alpha-1)/\alpha}]$$

by arguments like those used to prove equation (6.4) of Hall and Horowitz (2005). Therefore, (c)

follows from $a_n = C_a n^{-\alpha/(2\beta+\alpha)}$.

To prove (d), note that by (6.23)

(6.24) $\quad a_n \left\| (T_g T_g^* + a_n I)^{-1} \omega_g \right\| = O[a_n^{(2\beta-\alpha-1)/(2\alpha)}]$

$$= O[n^{-(2\beta-\alpha-1)/(4\beta+2\alpha)}].$$

Therefore, (d) follows from (6.24) and (a), because $\alpha < 2\beta - 1$. Now by assumptions 2 and 5(b),

$$\left\| \hat{f}_W - f_W \right\|^2 = O_p[n^{-(2\beta-1+\alpha)/(2\beta+\alpha)}]$$

and

$$\left\| \hat{T} - T \right\|^2 = O_p[n^{-(2\beta-1+\alpha)/(2\beta+\alpha)}]$$

uniformly over $\mathcal{H}$. Therefore, (e) and (f) follow from (6.24).

We now turn to (g). By the mean value theorem,

$$\left\{ \left[\hat{\mathcal{T}}(\hat{g}) - \mathcal{T}(\hat{g}) \right] - \left[\hat{\mathcal{T}}(g) - \mathcal{T}(g) \right] \right\}(w)$$

$$= \int_0^1 \{ \hat{f}_{YXW}[\bar{g}(x), x, w] - f_{YXW}[\bar{g}(x), x, w] \} [\hat{g}(x) - g(x)] dx,$$

where $\bar{g}$ is between $\hat{g}$ and g. Then by the Cauchy-Schwarz inequality

$$\left\| \left[\hat{\mathcal{T}}(\hat{g}) - \mathcal{T}(\hat{g}) \right] - \left[\hat{\mathcal{T}}(g) - \mathcal{T}(g) \right] \right\|^2$$

$$= \int_0^1 \left(\int_0^1 \{ \hat{f}_{YXW}[\bar{g}(x), x, w] - f_{YXW}[\bar{g}(x), x, w] \} [\hat{g}(x) - g(x)] dx \right)^2 dw$$

$$\leq \int_0^1 \left(\int_0^1 \{ \hat{f}_{YXW}[\bar{g}(x), x, w] - f_{YXW}[\bar{g}(x), x, w] \}^2 dx \int_0^1 [\hat{g}(x) - g(x)]^2 dx \right) dw$$

$$= \int_0^1 \int_0^1 \{ \hat{f}_{YXW}[\bar{g}(x), x, w] - f_{YXW}[\bar{g}(x), x, w] \}^2 dx dw \| \hat{g} - g \|^2.$$

But

$$\int_0^1 \int_0^1 \{ \hat{f}_{YXW}[\bar{g}(x), x, w] - f_{YXW}[\bar{g}(x), x, w] \}^2 dx dw = O_p[n^{-(2\beta-1)/(2\beta+\alpha)}] + \| \hat{g} - g \|^2 o_p(1)$$

by assumptions 2 and 5(b), thereby yielding (g). Finally, (h) can be proved by combining (a) with arguments similar to those used to prove (g). The lemma is now proved because the foregoing arguments hold uniformly over $H \in \mathcal{H}$. Q.E.D.

Proof of Theorem 2: The theorem follows by combining the results of lemmas 6.1-6.3 with $L \| \omega_g \| < 1$. Q.E.D.

6.3 Proof of Theorem 3

It suffices to find a sequence of finite-dimensional models $\{ g_n \} \in \mathcal{H}$ for which

$$\liminf_{n \to \infty} P_H \left[\| \tilde{g}_n - g_n \|^2 > Dn^{-(2\beta-1)/(2\beta+\alpha)} \right] > 0.$$

To this end, let m denote the integer part of $n^{-1/(2\beta+\alpha)}$ and f_{XW} denote the density of (X, W). Since $\alpha \geq 2$, $(2\beta + \alpha - 1)/2 \geq (3\beta + \alpha - 1/2)/(\alpha + 1)$. Let $r = (2\beta + \alpha - 1)/2$. Assume that $f_{XW}(x, w) \leq C$ for all $(x, w) \in [0,1]^2$ and some constant $C < \infty$. Let

$$Y = g_n(X) + U,$$

where U is independent of (X, W), and $P(U \leq 0) = q$. Let F_U and f_U, respectively, denote the distribution function and density of U. Assume that $f_U(0) > 0$ and that F_U is twice continuously differentiable everywhere with $|F_U''(u)| < M$ for all u and some $M < \infty$. Define the operator Q on $L_2[0,1]$ by

$$(\Pi g)(x) = \int_0^1 \pi(x, z) g(z) dz ,$$

for any $g \in L_2[0,1]$, where

$$\pi(x, z) = f_U(0)^2 \int_0^1 f_{XW}(x, w) f_{XW}(z, w) dw .$$

Let $\{\lambda_j, \phi_j : j = 1, 2, ...\}$ denote the orthonormal eigenvalues and eigenvectors of Π ordered so that $\lambda_1 \geq \lambda_2 \geq ... > 0$. Assume that $j^\alpha \lambda_j$ is bounded away from 0 and 1 for all j. Set

$$g_n(x) = \theta \sum_{j=m}^{\infty} j^{-\beta} \phi_j(x)$$

for some finite, constant $\theta > 0$. Then for any $h \in L_2[0,1]$,

$$(Th)(w) = \int_0^1 F_U[h(x) - g_n(x)] f_{XW}(x, w) dx ,$$

and the Fréchet derivative of T at g_n is

$$(T_{g_n} h)(w) = f_U(0) \int_0^1 f_{XW}(x, w)[h(x) - g_n(x)] dx .$$

Assumption 6 is satisfied with $L = MC$ whenever $\theta > 0$ is sufficiently small.

Now let $\hat{\theta}$ be an estimator of θ. Then

$$\hat{g}(x) \equiv \hat{\theta} \sum_{j=m}^{\infty} j^{-\beta} \phi_j(x) .$$

is an estimator of $g_n(x)$. Moreover,

(6.25) $\left\| \hat{g} - g_n \right\|^2 = (\hat{\theta} - \theta)^2 R_n ,$

where $R_n = \sum_{j=m}^{\infty} j^{-2\beta}$. Note that $n^{(2\beta-1)/(2\beta+\alpha)} R_n$ is bounded away from 0 and 1 as $n \to \infty$. In addition, f_W is estimated by

$$\hat{f}_w(w) = q^{-1}(T\hat{g})(w) .$$

Define $\psi_j = T_g \phi_j / \left\| T_g \phi_j \right\|$. Then a Taylor series approximation and singular value expansion give

$$\hat{f}_W(w) - f_W(w) = q^{-1}(\hat{\theta} - \theta) \sum_{j=m}^{\infty} j^{-\beta}(T_g \phi_j)(w) + (\hat{\theta} - \theta)^2 O[n^{-(2\beta-1)/(2\beta+\alpha)}]$$

.

$$= q^{-1}(\hat{\theta} - \theta) \sum_{j=m}^{\infty} j^{-\beta} \lambda_j^{1/2} \psi_j(w) + (\hat{\theta} - \theta)^2 O[n^{-(2\beta-1)/(2\beta+\alpha)}].$$

Now,

$$n^{(2\beta+\alpha-1)/(2\beta+\alpha)} \left\| \sum_{j=m}^{\infty} j^{-\beta} \lambda_j^{1/2} \psi_j \right\|^2$$

is bounded away from 0 and ∞ as $n \to \infty$. Therefore, there is a finite constant $C_\theta > 0$ such that

(6.26) $\quad (\hat{\theta} - \theta)^2 \geq C_\theta n^{(2\beta+\alpha-1)/(2\beta+\alpha)} \left\| \hat{f}_W - f_W \right\|^2.$

Combining (6.25) and (6.26) shows that there is a finite constant $C_g > 0$ such that

(6.27) $\quad n^{(2\beta-1)/(2\beta+\alpha)} \left\| \hat{g} - g_n \right\|^2 \geq C_g n^{(2\beta+\alpha-1)/(2\beta+\alpha)} \left\| \hat{f}_W - f_W \right\|^2.$

The theorem now follows from (6.27) and the observation that with $r = (2\beta + \alpha - 1)/2$,

$O_p[n^{(2\beta+\alpha-1)/(2\beta+\alpha)}]$ is the fastest possible minimax rate of convergence of $\left\| \hat{f}_W - f_W \right\|^2$. Q.E.D.

[1] The results of this paper hold even if f_{YXW} or its derivatives are discontinuous at one or more boundaries of the support of (Y, X, W) provided that the kernel function K is replaced by a boundary kernel.

REFERENCES

Amemiya, T. (1982). Two stage least absolute deviations estimators. *Econometrica*, 50, 689-711.

Bissantz, N., T. Hohage, and A. Munk (2004). Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems*, 20, 1773-1789.

Bissantz, N., T. Hohage, A. Munk, and F. Ruymgaart (2006). Convergence rates of general regularization methods for statistical invetrse problems and applications, working paper, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen, Germany.

Blundell, R. and J.L. Powell (2003). Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications*, M. Dewatripont, L.-P. Hansen, and S.J. Turnovsky, eds., 2, 312-357, Cambridge: Cambridge University Press.

Carrasco, M., J.-P. Florens, and E. Renault (2005). Linear inverse problems in structural econometrics: estimation based on spectral decomposition and regularization. In *Handbook of Econometrics*, Vol. 6, E.E. Leamer and J.J. Heckman, eds, Amsterdam: North-Holland, forthcoming.

Chen, L. and S. Portnoy (1996). Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equation models. *Communications in Statistics, Theory and Methods*, 25, 1005-1032.

Chernozhukov, V. and C. Hansen (2005a). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, forthcoming.

Chernozhukov, V. and C. Hansen (2005b). An IV model of quantile treatment effects. *Econometrica*, 73, 245-261.

Chernozhukov, V. and C. Hansen (2004). The effects of 401(k) participation on the wealth distribution: an instrumental quantile regression analysis. *Review of Economics and Statistics*, 86, 735-751.

Chernozhukov, V., G.W. Imbens, and W.K. Newey (2004). Instrumental variable identification and estimation of nonseparable models via quantile conditions. Working paper, Department of Economics, Massachusetts Institute of Technology, Cambridge, MA.

Chesher, A. (2003). Identification in nonseparable models. *Econometrica*, 71, 1405-1441.

Darolles, S., J.-P. Florens, and E. Renault (2002). Nonparametric instrumental regression. Working paper, GREMAQ, University of Social Science, Toulouse, France.

Engl, H.W., M. Hanke, and A. Neubauer (1996). *Regularization of Inverse Problems*. Dordrecht: Kluwer Academic Publishers.

Florens, J.-P. (2003). Inverse problems and structural econometrics: the example of instrumental variables. In *Advances in Economics and Econometrics: Theory and Applications*, Dewatripont, L.-P. Hansen, and S.J. Turnovsky, eds., 2, 284-311, Cambridge: Cambridge University Press.

Groetsch, C. (1984). *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. London: Pitman.

Hall, P. and J.L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics*, 33, 2904-2929.

Horowitz, J.L. (2005). Asymptotic normality of a nonparametric instrumental variables estimator. Working paper, Department of Economics, Northwestern University, Evanston, IL.

Januszewski, S.I. (2002). The effect of air traffic delays on airline prices. Working paper, Department of Economics, University of California at San Diego, La Jolla, CA.

Kress, R. (1999). *Linear Integral Equations*, 2nd edition. New York: Springer.

Ma, L. and R. Koenker (2006). Quantile regression methods for recursive structural equation models. *Journal of Econometrics*, forthcoming.

Newey, W.K. and J.L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71, 1565-1578.

Newey, W.K., J.L. Powell, and F. Vella (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67, 565-603.

Powell, J.L. (1983). The asymptotic normality of two-stage least absolute deviations estimators, *Econometrica*. 50, 1569-1575.

Tikhonov, A.N. (1963a). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4, 1035-1038.

Tikhonov, A.N. (1963b). Regularization of incorrectly posed problems. *Soviet Mathematics Doklady*, 4, 1624-1627.

Table 1. *Results of Monte Carlo Experiments*

n	a_n	h	Bias^2	Variance	MISE
200	0.5	0.5	0.0226	0.0199	0.0425
	0.5	0.8	0.0795	0.0220	0.1015
	1.0	0.5	0.0146	0.0195	0.0341
	1.0	0.8	0.0247	0.0203	0.0450
800	0.5	0.5	0.0233	0.0048	0.0282
	0.5	0.8	0.0816	0.0054	0.0870
	1.0	0.5	0.0151	0.0047	0.0199
	1.0	0.8	0.0258	0.0050	0.0308

Note: Bias^2, Variance, and MISE, respectively, are Monte Carlo approximations to $\int_0^1 (E[\hat{g}(x)] - g(x))^2 dx$, $\int_0^1 (\hat{g}(x) - E[\hat{g}(x)])^2 dx$, and $\int_0^1 (\hat{g}(x) - g(x))^2 dx$.

Figure 1: The Density of W

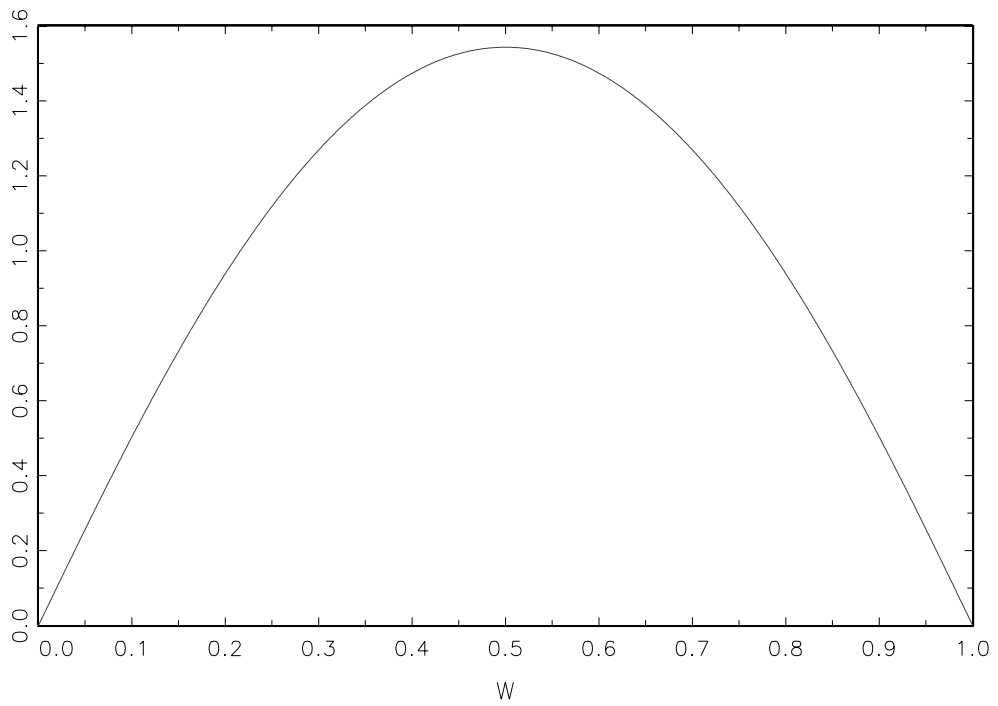The Probablity Density Function of W

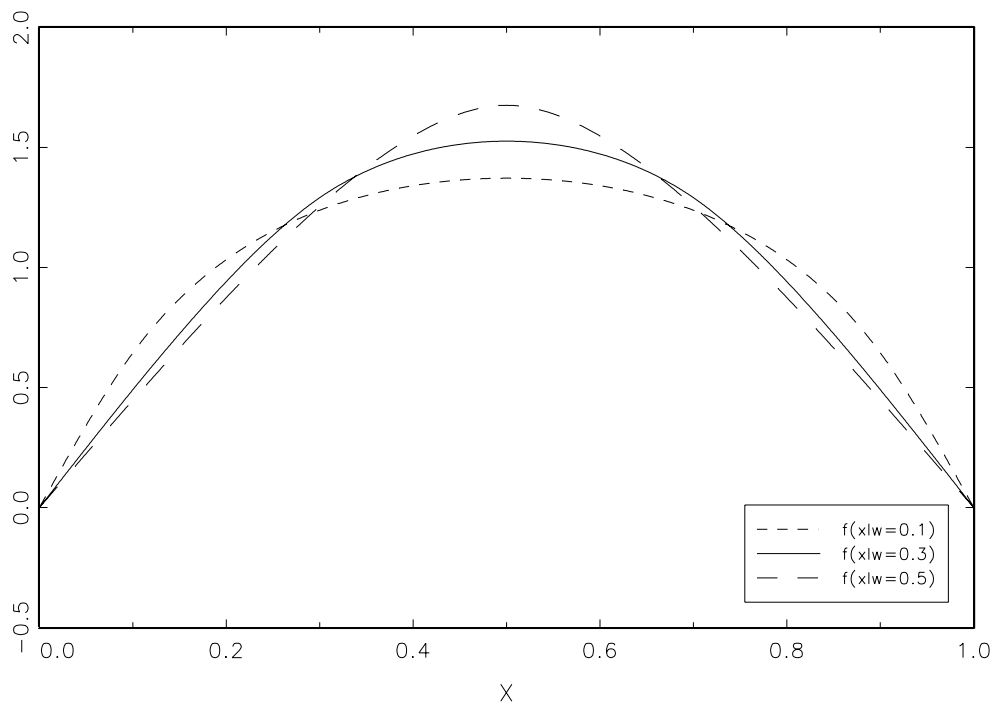Figure 2: The Density of X Conditional on W

The Density of X Conditional on W

Figure 3: The Density of U Conditional on X and W
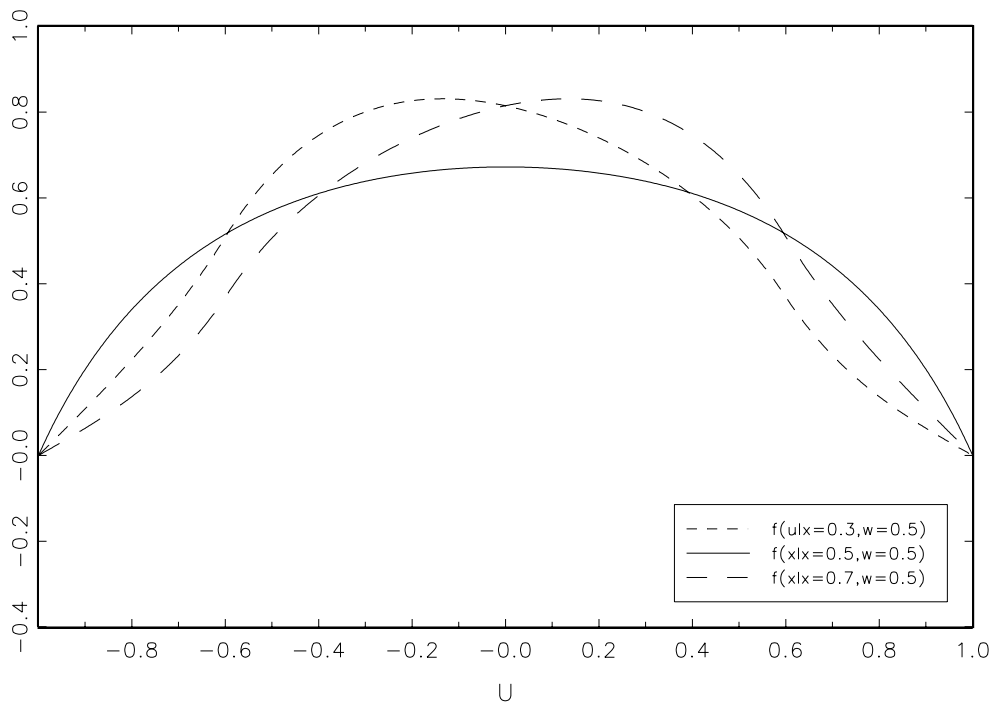

The Density of U Conditional on (X,W)

Figure 4: Monte Carlo Results for $n = 200$

a = 0.50 and h = 0.50

a = 0.50 and h = 0.80

a = 1.00 and h = 0.50

a = 1.00 and h = 0.80

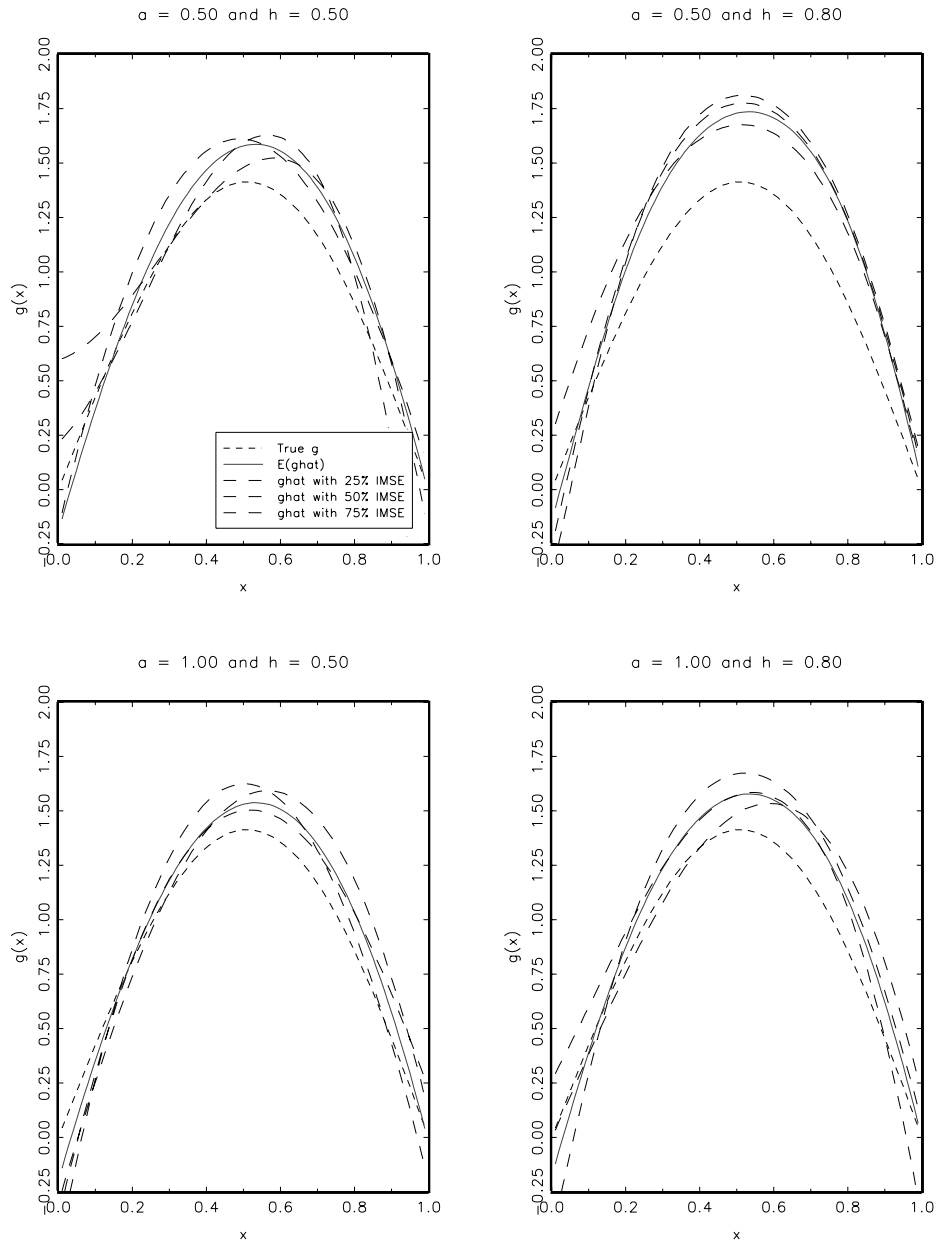True g
E(ghat)
ghat with 25% IMSE
ghat with 50% IMSE
ghat with 75% IMSE

Figure 5: Monte Carlo Results for $n = 800$