Instrumental Nonparametric Estimation Under Conditional Moment Restrictions: A Nonlinear Tikhonov Approach

Jerome M. Krief *

November 12, 2011

Abstract

This paper treats the estimation of an unknown function θ_0 which is identified via the conditional moment restriction $E[\rho(Z,\theta_0)|W]=0$ where W is a continuously distributed instrument and ρ is a known mapping. Using the principle of Tikhonov regularization, a new estimator minimizing some penalized objective is shown to be MSE consistent under a relatively strong Hilbert norm. This Tikhonov estimator achieves a rate of convergence in probability equal to $n^{-1/(4+\gamma)}$, where γ is a positive constant. Furthermore, γ vanishes under adequate smoothness conditions. This methodology does not require the unknown function to be continuous. Some Monte Carlo experiments are conducted highlighting the good finite sample properties of this estimator.

Key words: Tikhonov regularization, Instrumental nonparametric estimation, Nonlinear integral equations.

*University of Miami, Department of Economics, Box 248126, Coral Gables, FL 33124. I would like to thank Marine Carrasco, Peter Hall, Thorsten Hohage, Jeff Racine and Eric Renault for comments and suggestions.

1 Introduction

This paper considers the general model

$$E[\rho(Z, \theta_0)|W] = 0 \tag{1}$$

almost surely (a.s.), where $Z \in \mathbb{R}^{|Z|}$ for some natural number $|Z| \geq 1$, $W \in \mathbb{R}$ is a continuous instrument excluded from Z, ρ is a known mapping from $\mathcal{Z} \times \mathcal{H}$ to the real line, and θ_0 is a parameter of interest belonging to $\mathcal{H} \subset L^2(\mathcal{X}, Leb)$, where $L^2(\mathcal{X}, Leb)$ denotes the space of functions from $\mathcal{X} \subset \mathbb{R}$ to the real line which are square integrable with respect to Lebesgue measure.

Such conditional moment restrictions frequently arise in econometrics due to data limitations which make the generalized residual $\rho(Z, \theta_0)$ correlated with some component of Z. In that case, the availability of some external variable W uncorrelated with $\rho(Z, \theta_0)$ may offer statistical identification of the unknown parameter θ_0. Important examples satisfying (1) include the endogenous regression model with a known link (Ai and Chen 2003, Horowitz and Mammen 2004), the endogenous nonparametric regression model (Hall and Horowitz 2005, Darolles, Fan, Florens, and Renault 2011), and the endogenous nonparametric quantile regression model (Horowitz and Lee 2007) which encompasses the instrumental variable regression for nonseparable models of Chernozhukov, Imbens, and Newey (2007) and the quantile treatment effect model of Chernozhukov and Hansen (2005).

If θ_0 admits a certain number of well-behaved derivatives then one can estimate θ_0 consistently using the estimator proposed in Newey and Powell (2003) or the Sieves Minimum Distance estimator described in Ai and Chen (2003). However, if θ_0 is not continuous then these estimators are inconsistent. This paper presents a methodology which does not require θ_0 to be continuous.

To understand informally the motivation behind the methodology developed in this paper, write $E_{Z|W}$ for the expectation operator with respect to the distribution of Z conditional on W. An intuitive approach is to solve the empirical counterpart of (1), say $\hat{E}_{Z|W}[\rho(Z, \theta)|W] = 0$, where $\hat{E}_{Z|W}$ is some consistent estimator of $E_{Z|W}$ constructed from data. It turns out that this approach does not ensure consistency because the solution, viewed as a mapping from the space of distributions into the parameter space, is not continuous in data (Engl, Kunisch, and Neubauer 1989). Unlike the case where θ_0 belongs to a finite-dimensional parameter space (see Hansen 1982,

Judge et al. 1980), this lack of continuity or "ill-posedness" is neither data-related nor an identification problem. It is simply inherent in the fact that the parameter space does not have a finite dimension so that small estimation errors can be amplified *ad* infinity.

One way to mitigate the ill-posedness phenomenon is to assume that θ_0 is part of a compact parameter space (with respect to some norm) because this makes the solution continuous by the Arzela theorem (see Gallant and Nychka 1987, Engl and Kügler 2005). This is the approach chosen in Newey and Powell (2003) and Ai and Chen (2003).

An alternative strategy is to regularize the solution, that is, to modify it to a continuous mapping. Various regularization schemes are available but this paper applies the *Tikhonov* regularization method developed in Bissantz, Hohage, and Munk (2004). This only requires that θ_0 belong to a Hilbert space, and therefore can accommodate a more general class of functions, notably those discontinuous on a Lebesgue null set. This Tikhonov scheme achieves regularization by penalizing some objective. The proof to show consistency works directly from the objective. The is convenient in the context of our problem because this regularization technique does not require linearizing the first-order conditions, a theoretical challenge when the unknown parameter solves a nonlinear integral equation due to the ill-posedness previously mentioned.

Under mild requirements, the restriction in (1) implies that θ_0 solves $T\theta = 0$ for some nonlinear operator T known up to some densities. The method of Tikhonov regularization is used, replacing T by a nonparametric estimator which permits recovering a consistent estimator of θ_0. If certain density functions have enough derivatives and if some generalized Fourier coefficients of θ_0 collapse fast enough then $||\hat{\theta} - \theta_0|| = O_p(n^{-\frac{1}{4+\gamma}})$ where $||.||$ is a Hilbert norm and γ is a positive constant. Furthermore, γ becomes negligible if the densities in question are sufficiently differentiable. Hence, under some smoothness conditions, one can get close to the rate of $n^{-1/4}$ without assuming that the parameter space is compact. This rate of convergence is slower than that derived in Ai and Chen (2003). This is not surprising because the parameter space is not restricted to be compact in this paper. This rate of convergence corresponds to the fastest one achievable with the Tikhonov regularization scheme for what shall be called the linear case, characterized by $\mathcal{H} \equiv L^2(\mathcal{X}, Leb)$ and $\rho(Z, \theta) \equiv Z_1 - A(\theta)(Z_2)$ for some given linear operator A such as in the endogenous nonparametric regression model (see Carrasco 2007). This result, arising under weaker smoothness–source conditions, stems from the fact that $\mathcal{H}$ is assumed closed in this paper. In effect, this makes the optimization problem solved in this article constrained, unlike the linear case. It is common in the linear ill-posed

literature to have the constrained estimator achieve a rate of convergence no slower than the unconstrained one for a given set of source conditions (Carrasco and Florens 2010).

The estimator presented in this paper shares similarities with those in the literature for the nonparametric endogenous quantile regression (Horowitz and Lee 2007, Chernozhukov, Gagliardini, and Scaillet 2009). For the linear case, Tikhonov methods are available (Hall and Horowitz 2005, Darolles, Florens, and Renault 2006, 2007 Section 4). When ρ is nonlinear in θ_0 and $\mathcal{H}$ is not compact, no methodology is available as far as I know.

The rest of this paper is organized as follows. Section 2 provides a summary of the estimation procedure. Section 3 presents the assumptions and results. Finally, Section 4 exhibits a Monte Carlo experiment in the context of the endogenous regression model with a known link function. All the proofs are located in the Appendix.

At this point, it is convenient to introduce some notations used throughout the subsequent sections. Given $f : \mathbb{R} \longrightarrow \mathbb{R}$, define $f^{(j)}(t)$ to be its j^{th} derivative at t whenever this latter exists. For any f and g belonging to $L^2(\mathcal{X}, Leb)$, write $< f, g >_{\mathcal{X}} \equiv \int_{\mathcal{X}} f(x)g(x)dx$ and $||f||_{\mathcal{X}} \equiv \sqrt{< f, f >_{\mathcal{X}}}$. For a measure space $(\mathcal{X}, B, \ell)$ where ℓ indicates the Lebesgue measure on $\mathcal{X}$, a property will be called "Leb a.e.x" if the property is true except on a Lebesgue null set, i.e., it is true for all $x \in \mathcal{X} \setminus N$ with $\ell(N) = 0$. $\mathcal{L}(E, F)$, where E and F are two complete Hilbert spaces, denotes the space of linear bounded operators from E to F. When $\mathbf{T}$ belongs to $\mathcal{L}(E, F)$, use $||\mathbf{T}|| \equiv sup_{\{x \in E : ||x||_E = 1\}} ||\mathbf{T}x||_F$ where $||.||_E$, respectively $||.||_F$, is the norm induced by the inner product on E, respectively on F. Also, for a sequence φ_n of elements of E and $\varphi \in E$, the symbol $\varphi_n \rightsquigarrow \varphi$ denotes weak convergence in E, i.e., $\lim_n < \varphi_n, e >=< \varphi, e >$ for any $e \in E$. For a natural number $s \geq 2$, define $\mathcal{K}_s$=$\{f : \mathbb{R} \longrightarrow \mathbb{R}$, f continuous, symmetrical, supported on $[-1, 1]$, $\int f(t)dt = 1$, $\int t^u f(t)dt = 0$ for $u = 1, ..., s-1$ and $\int t^s f(t)dt \neq 0\}$. For a given natural number $p \geq 1$, define $\mathcal{K}_{p,s}$ =$\{\kappa : \mathbb{R}^p \to \mathbb{R}$, $\kappa(t_1, ..., t_p) \equiv \prod_{j=1}^p k(t_j)$, $k \in \mathcal{K}_s$ $\}$. Given a strictly positive sequence $\{a_n\}_{n \geq 1}$ and a sequence $\{c_n\}_{n \geq 1}$, the symbol $c_n/a_n \asymp 1$ means that c_n/a_n is bounded away from 0 and infinity.

2 Summary of the Estimation Procedure

In this paper $\mathcal{H}$ must be closed and convex. Thus, we may assume without loss of generality that $\mathcal{H} \equiv \{\theta \in L^2(\mathcal{X}, Leb) : ||\theta||_{\mathcal{X}} \leq M\}$ where $M < \infty$ is a given real number. As explained in Section 3, other choices for $\mathcal{H}$ are

possible.

Write $\rho_\theta(Z) \equiv \rho(Z, \theta)$ and assume that there exists a partition of $Z' = (S', V')$ where $S \in \mathbb{R}^{|S|}$ is discrete with bounded support $\mathcal{S}$ and $V \in \mathbb{R}^{|V|}$ is continuous with $|S| + |V| \geq 1$. Recall that $|V| = 0$ is possible if $\rho_\theta(z)$ is a functional where only the path of θ determines the value for a given z. Identification requires that (V', W) admit a density with respect to Lebesgue measure denoted f_{VW}, and we may assume without loss of generality that its support is contained in $[0,1]^{|V|+1}$. Finally, denote the probability density function (PDF) of $S|V = v, W = w$ by $p(.|v, w)$. If $|S| = 0$, we adopt the convention that $p(.|v, w) \equiv 1$, and if $|V| = 0$, use $f_{VW} \equiv f_W$ and $p(.|v, w) \equiv p(.|w)$, where $p(.|w)$ indicates the PDF of $S|W = w$. These and (1) imply that θ_0 satisfies the following integral equation

$$(T\theta)(w) \equiv \int_{\mathcal{S}\times[0,1]^{|V|}} \rho_\theta(s, v) f(s, v, w) d\chi(s) dv = 0, \tag{2}$$

up to a Lebesgue null set of $[0, 1]$, where

$$f(s, v, w) \equiv p(s|v, w) f_{VW}(v, w),$$

with χ indicating counting measure. Little is known as to how to ensure that θ_0 is the unique solution. Some results are provided in Chen, Chernozhukov, Lee, and Newey (2011) to obtain local identification. Section 3 provides generic conditions under which θ_0 is identified up to a Lebesgue null set. That is, if there exists some other element in $\mathcal{H}$ satisfying (1) then it must differ from θ_0 only on a subset of $\mathcal{X}$ having zero Lebesgue measure. The possibility of such a subset does not matter for consistency purposes because of the global metric chosen to show consistency.

To expose the exact nature of the estimation challenge, suppose T admits a Fréchet derivative at θ_0, say T_1, that is, there exists some linear bounded operator $T_1 : L^2(\mathcal{X}, Leb) \to L^2([0, 1], Leb)$ satisfying

$$T(\theta) = T_1(\theta - \theta_0) + R(\theta) \text{ for all } \theta \in \mathcal{H} \text{ with } ||R(\theta)||_{[0,1]} = o(||\theta - \theta_0||_{\mathcal{X}}).$$

For the models discussed in Section 1, the operator is Fréchet differentiable under mild smoothness conditions. Under the assumptions of this article, this latter operator has the form

$$(T_1\varphi)(w) = \int_{\mathcal{X}} \varphi(x) F(f_{ZW}, \theta_0)(x, w) dx,$$

for some functional F. Although F is known for the models discussed in Section 1, this is not required. Introduce

$T_1^* : L^2([0,1], Leb) \to L^2(\mathcal{X}, Leb)$ the adjoint of T_1 given by

$$(T_1^*\varphi)(x) = \int_{[0,1]} \varphi(w) F(f_{ZW}, \theta_0)(x, w) dw.$$

Under some integrability conditions given in Section 3, $T_1^* T_1$ becomes a compact operator, and therefore admits a spectrum (Kress 1999 Theorem 15.16), say $\sigma \equiv \{\lambda_j, j \geq 1\}$ with $\lambda_1 \geq \lambda_2 \geq ...$, repeating an eigenvalue by its order of multiplicity if needed. Technically, the ill-posedness phenomenon described in Section 1 arises because the sequence of eigenvalues converges to 0, which makes $(T_1^* T_1)^{-1}$ discontinuous. In that case, forming the sample analogue of (2) does not yield consistency since the estimation loss depends on $(T_1^* T_1)^{-1}$ applied on the data estimation error. The Tikhonov methodology proposed in this article consists of modifying $T_1^* T_1$ to a new operator whose inverse is continuous.

To describe the construction of the estimator, let $\{Z_i, W_i\}_{i=1}^n$ be some i.i.d. sequence of observations from (Z, W). Use the notation $I(S = s) = 1$ if $S = s$ and $I(S = s) = 0$ otherwise with the convention $I(S = s) \equiv 1$ whenever $|S| = 0$. The estimator of T is given by

$$(\hat{T}\theta)(w) \equiv \int_{\mathcal{S} \times [0,1]^{|V|}} \rho_\theta(s, v) \hat{f}(s, v, w) d\lambda(s) dv,$$

with

$$\hat{f}(s, v, w) \equiv \frac{1}{n h_w h_v^{|V|}} \sum_{i=1}^n I(S_i = s) K_v(\frac{V_i - v}{h_v}) K_w(\frac{W_i - w}{h_w}),$$

where h_w and h_v are deterministic strictly positive bandwidths satisfying $\lim h_w = 0$ and $\lim h_v = 0$ as $n \to \infty$, $K_v(.)$ any kernel belonging to $\mathcal{K}_{|V|,r}$ if $|V| \geq 1$ and $K_v(.) \equiv 1$ otherwise, while $K_w(.)$ is any kernel belonging to $\mathcal{K}_r$ for some $r \geq 2$.

The Tikhonov estimator $\hat{\theta}$ solves the following problem,

$$Min_{\theta \in \mathcal{H}} ||\hat{T}\theta||^2_{[0,1]} + a_n ||\theta||^2_{\mathcal{X}},$$

where a_n is a strictly positive sequence of real numbers satisfying $\lim a_n = 0$ as $n \to \infty$.

To appreciate the presence of the penalty term $a_n ||\theta||^2_{\mathcal{X}}$ in the objective, one may write heuristically $\hat{\theta} - \theta_0 \approx (T_1^* T_1 + a_n I)^{-1}(e)$ from the first-order conditions, where e contains the sources of error from estimating T. Hence, adding the penalty term in the objective regularizes the solution since $(T_1^* T_1 + a_n I)^{-1}$ is bounded.

Let ϕ_λ denote the eigenfunction with eigenvalue λ. Under a certain injectivity condition given in Section 3, the sequence of eigenfunctions forms an orthonormal basis of $L^2(\mathcal{X}, Leb)$. Hence, θ_0 has the representation

$$\theta_0 = \sum_{\lambda \in \sigma} b_\lambda \phi_\lambda,$$

where $b_\lambda \equiv < \theta_0, \phi_\lambda >_{\mathcal{X}}$ is the generalized Fourier coefficient associated with eigenfunction ϕ_λ. Because the regularization process, in effect, seeks to estimate an approximation of θ_0, it introduces a deterministic bias. A prerequisite for the bias to vanish asymptotically is that the Fourier coefficients exhibit a sufficiently rapid rate of decline in the the sense that

$$\sum_{\lambda \in \sigma} \frac{b_\lambda^2}{\lambda} < C_0, \tag{3}$$

for some real constant $C_0 < \infty$ whose exact expression is given in section 3. Now assume (3) is true. Write $c \equiv |V|+1$ and $\delta_n \equiv n^{-2r/(2r+c)}$. Suppose the researcher selects the smoothing parameters according to $h_v \propto n^{-1/(2r+c)}$, $h_w \propto n^{-1/(2r+c)}$, and $a_n \asymp \sqrt{\delta_n}$. When $f(s, v, w)$, for each $s \in \mathcal{S}$ which is at least twice differentiable, Proposition 3 will establish

$$E||\hat{\theta} - \theta_0||_{\mathcal{X}}^2 = O(\sqrt{\delta_n}) \tag{4}$$

The result from (4) implies $||\hat{\theta} - \theta_0||_{\mathcal{X}} = O_p(n^{-r/(4r+2c)})$. As explained in Section 3, obtaining $||\hat{\theta} - \theta_0||_{\mathcal{X}} = O_p(n^{-1/4})$ is in general not achievable unless $f(s, v, w)$ is sufficiently differentiable.

3 Assumptions

Assumption 1:

(a) *W admits a continuous density with respect to Lebesgue measure, denoted $f_W(.)$, with bounded support $\mathcal{W} \subset \mathbb{R}$.*

(b) *$Z \in \mathbb{R}^{|Z|}$ for some natural number $|Z| \geq 1$ and its support $\mathcal{Z}$ is bounded. Also, $Z|W = w$ admits a probability density denoted $p(.|w)$ with respect to some sigma-finite measure denoted μ (Leb a.e. w). Furthermore, ($\mathcal{Z}$,$\mathcal{B}$, μ) is a complete measure space where $\mathcal{B}$ refers to the Borel field on which μ is countably additive.*

Assumption 2:

θ_0 *belongs to* $\mathcal{H} \equiv \{\theta \in L^2(\mathcal{X}, Leb) : ||\theta||_{\mathcal{X}} \leq M\}$ *where* $M < \infty$ *is a given real number and* $\mathcal{X}$ *a given compact subset of the real line.*

Assumption 3:

For any $\theta \in \mathcal{H}$,

$\rho_\theta(.)$ *is a Borel measurable real-valued function from* $\mathcal{Z}$ *with* $E|\rho_\theta(Z)| < \infty$.

Assumption 4:

Define $f_{ZW}(z, w) \equiv p(z|w)f_W(w)$ whenever this latter exists. Also, for any $g : \mathcal{Z} \to \mathbb{R}$ with $g\mu$-measurable, write $||g||_{\mathcal{Z}}^2 \equiv \int_{\mathcal{Z}} |g|^2 d\mu$ whenever the integral exists. Finally, use $\mu \times \ell$ for the product measure on the Borel field of $\mathcal{Z} \times \mathcal{W}$.

(a) $\int_{\mathcal{Z}\times\mathcal{W}} |f_{ZW}|^2 d(\mu \times \ell) < \infty$.

(b) $||\rho_\theta||_{\mathcal{Z}} < \infty$ *for all* $\theta \in L^2(\mathcal{X}, Leb)$.

Assumption 5:

(a) $E[\rho_{\theta_0}(Z)|W = w] = 0$ *Leb a.e.* w.

(b) For any $\varphi \in L^2(\mathcal{X}, Leb)$ *with* $||\varphi||_{\mathcal{X}} \neq 0$, *there exists* $D \in \mathcal{L}(L^2(\mathcal{X}, Leb), L^2([0, 1], Leb))$ *such that* $T(\theta_0 + \varphi) = D(\varphi)$.

(c) $Inf_{\{||\theta||_{\mathcal{X}}=1\}} ||D(\theta)||_{\mathcal{W}} > 0$.

Comments: Assumption 1(a) requires a continuous instrument. Assumption 1(b) allows very general variables in Z, notably a mixture of discrete and continuous variables with bounded supports. Assumption 2 makes the domain of T closed and convex which is convenient in order to derive the consistency of the first stage estimator using the theory laid out in Bissantz, Hohage, and Munk (2004). The relevance of Assumption 2 for applied works is that θ_0 is allowed to be discontinuous at many points. Assumption 2 can be modified by working on the domain $\mathcal{H} \equiv \{\theta \in L^2(\mathcal{X}, Leb) : \theta \geq \underline{C}, ||\theta||_{\mathcal{X}} \leq M\}$ for some given strictly positive real number $\underline{C}$ which is more natural, for

instance, in models when θ_0 represents a density function or the derivative of an increasing function as suggested by economic theory. The identification condition of Assumption 5(b) is made only up to a Lebesgue null set since in most cases the concept of completeness (see Newey and Powell 2003) permits only identification in that sense. Assumption 5 (b) is a form of mean value expansion for a nonlinear operator with D being some linear representor. Assumption 5 (c) is an injectivity condition which may be viewed as the generalization of the rank condition for the infinite-dimensional case.

Proposition 1 (Identification)

Under Assumptions 1 through 5,

$$T\theta = 0 \text{ Leb a.e. } w \text{ implies } \theta = \theta_0 \text{ Leb a.e. } x$$

Assumption 6:

(a) For any sequence $\{\theta_n\} \subseteq \mathcal{H}$, $\theta_n \rightsquigarrow \theta$ *implies* $\rho_{\theta_n} \rightsquigarrow \rho_\theta$ *in* $L^2(\mathcal{Z}, \mu)$.

(b) $sup_{\theta \in \mathcal{H}} ||\rho_\theta||_{\mathcal{Z}} < \infty$.

Comments: Assumption 6(b) is technical, allowing a uniform rate of convergence for the estimator of T in the sense of the $L^2(\mathcal{X}, Leb)$ induced metric. Assumption 6(a) permits weak continuity in the sense of the Hilbert inner products. This and Assumption 2 ensures that T is weakly sequentially closed (Bissantz, Hohage, and Munk 2004). The weak continuity assumption plays a key role in establishing both the existence of a solution for the minimization problem and its consistency.

Assumption 7:

There exists a partition of $Z' = (S', V')$ *such that*

(a) (V', W) *admits a density with respect to Lebesgue measure denoted* $f_{VW}(.,.)$ *with support* $\mathcal{V} \times \mathcal{W} \subset [0,1]^{|V|+1}$.

(b) $S|V = v, W = w$ *admits a probability density with respect to counting measure denoted* $p(.|v,w)$ *Leb a.e.* v, w. *Furthermore, the support of* S *denoted* $\mathcal{S} \subset \mathbb{R}^{|S|}$ *is bounded.*

(c) $\{Z_i, W_i\}_{i=1}^n$ *is an i.i.d. sequence from* (Z, W).

Assumption 8:

(a) As functions of (v, w)*, for each* $s \in \mathcal{S}$ *and for some* $r \geq 2$*, all partial derivatives of order not greater than* r *of* $f_{VW}(v, w)$ *and* $p(s|v, w)$ *are continuous and bounded.*

(b) K_v *belongs to* $\mathcal{K}_{|V|,r}$ *and* K_w *belongs to* $\mathcal{K}_r$

Assumption 9:

Writing $\delta_n \equiv Max(h_w, h_v)^{2r} + 1/nh_w h_v^{|V|}$,

(a) a_n *is a sequence of strictly positive real numbers satisfying* $\lim a_n = 0$ *as* $n \to \infty$*. (b)* $\lim \delta_n / a_n = 0$ *as* $n \to \infty$*.*

Comments: Assumption 7(a) is imposed to simplify the proofs concerning the asymptotic behavior of the estimators. The results remain valid if (V, W) has unbounded support if one imposes smoothness assumptions on the last derivatives for Assumption 8(a) similar to those in Robinson (1988) and Florens et al. (2005). Additionally, if the derivatives of Assumption 8(a) are not continuous on the boundaries of the support, the results remain true using the boundary kernels provided in Hall and Horowitz (2005). Assumption 9 demands the researcher to let the regularization sequence collapse at a rate less rapid than the MSE rate achieved by the nonparametric estimator of f_{VW}.

Proposition 2 (Convergence in Norm)

Let $\hat{\theta} \equiv Argmin_{\theta \in \mathcal{H}} ||\hat{T}\theta||^2_{[0,1]} + a_n ||\theta||^2_{\mathcal{X}}$*. Under Assumptions 1 through 9,*

$$\lim E||\hat{\theta} - \theta_0||^2_{\mathcal{X}} = 0 \text{ as } n \to \infty.$$

Comments: One may pick, in theory, a different norm for the regularization term. That is, replace the right-hand side of the objective with $a_n||\theta||^2$ where $||\theta||$ is induced by some other inner product (Chernozhukov, Gagliardini, and Scaillet 2009, Chen and Pouzo 2008) whenever $||\theta_0||$ exists. In practice, the estimator will be computed using a finite-dimensional space dense for $\mathcal{H}$ in the sense of the metric $||.||_{\mathcal{X}}$. A truncated series using some known orthonormal basis of $L^2(\mathcal{X}, Leb)$, such as the sine or cosine basis, is the most natural choice. Under slightly stronger assumptions than those imposed in this paper, one can choose from a richer set of alternatives to select the finite-dimensional space. For instance, one may compute the estimator relying on a high-dimensional nonlinear

space such as the sigmoid Artifical Neural Network (Chen and White 1999), provided the Fourier transform of θ_0 satisfies some mild integrability conditions. Apart from Neubaeur (1987), little formal results are available to understand the numerical effect, in a finite sample, of using a space of large dimension, albeit finite, for computing the estimator.

Assumption 10:

(a) There exists $T_1 \in \mathcal{L}(L^2(\mathcal{X}, Leb), L^2([0,1], Leb))$ *and a real number* $L < \infty$ *such that:*

$$\lim \frac{||T(\theta_0 + \varphi) - T_1(\varphi)||_{[0,1]}}{||\varphi||_{\mathcal{X}}} = 0,$$

as $\varphi \to 0$ *and for all* $\varphi \in \mathcal{H}$*,*

$$||T(\varphi) - T_1(\varphi - \theta_0)||_{[0,1]} \leq L||\varphi - \theta_0||^2_{\mathcal{X}}.$$

Furthermore, writing $||.||_{ess}$ for the essential supremum on $[0,1]$,

$$sup_{||\varphi||_{\mathcal{X}}=1}||T_1(\varphi)||_{ess} < \infty,$$

Denoting by $F(f_{ZW}, \theta_0)(x,w)$ the real valued function satisfying $T_1(\varphi) =< \varphi, F(f_{ZW}, \theta_0)(.,w) >_{\mathcal{X}}$ Leb a.e. w, which exists by Assumption 10(a)[1], also assume:

(b)$\int_{[0,1]} \int_{\mathcal{X}} |F(f_{ZW}, \theta_0)(x,w)|^2 dxdw < \infty$*.*

(c)$\int_{\mathcal{X}} \phi(x) F(f_{ZW}, \theta_0)(x,w)dx = 0$ *Leb a.e.* w *implies* $\phi(x) = 0$ *Leb a.e.* x*.*

Comments: Assumption 10 demands the existence of the Fréchet derivative at θ_0 with a Lipschitz continuity on the Fréchet derivative in the sense of the operator norm. Fréchet differentiability at θ_0 is stronger than Gâteaux differentiability at θ_0 since the latter is implied by the former. From an applied perspective, one needs to nail down this operator T_1. It is useful to use the fact that if the Fréchet derivative at θ_0 exists it is then given by

$$T_1(\varphi)(w) \equiv \lim_{\varepsilon \to 0} \frac{T(\theta_0 + \varepsilon\varphi)(w)}{\varepsilon},$$

[1]This is a direct consequence of the Riesz representation theorem since the linear functional $L_w(\varphi) \equiv T_1(\varphi)(w)$ is bounded for almost every w by Assumption 10(a). Thus, for almost every w there exists a unique $\varrho_w \in L^2(\mathcal{X}, Leb)$ such that $T_1(\varphi)(w) =< \varphi, \varrho_w >_{\mathcal{X}}$ for all $\varphi \in L^2(\mathcal{X}, Leb)$. This $\varrho_w(x)$ is naturally denoted by $F(f_{ZW}, \theta_0)(x,w)$ since it depends implicitly on f_{ZW} and θ_0 via some functional F.

for any $\varphi \in L^2(\mathcal{X}, Leb)$. Applying the Lebesgue Dominated Convergence Theorem along with some mean value expansion usually permits recovering the functional F for the problem at hand using the above formula. In fact this is how the Fréchet derivative at θ_0 will be found in practice, that is, compute the above limit and then check whether the suggested operator meets the second criteria of Assumption 10(a). Assumption 10(c) is the classic injectivity requirement whose plausibility will depend on how much the kernel $F(f_{ZW}, \theta_0)(x, w)$ varies once w is fixed, loosely speaking.

Assumption 11:

$$\sqrt{\sum_{j=1}^{\infty} \frac{| < \theta_0, \phi_j > |^2}{|\lambda_j|^2}} < 1/2L,$$

where L is the Lipschitz constant from Assumption 10(a).

Assumption 12:

$\sqrt{\delta_n}/a_n \asymp 1$

Comments: Assumption 11 is a source condition (see Bissantz, Hohage, and Munk 2004). The more severe is the rate of decay of the eigenvalues (i.e., the more ill-posed is the problem), the less complex θ_0 must be in the sense of having $\theta_0 \simeq \sum_{j=1}^{J} < \theta_0, \phi_j >_{\mathcal{X}} \phi_j$ an accurate approximation for some small natural number J. Also, the smoother the non linear operator (i.e., the smaller is L in Assumption 10), the more complex θ_0 is allowed to be. Notice that for linear models, T and its Fréchet derivative coincide. In that case, any positive L can be chosen and assumption 11 is the classic source condition of Carrasco and Florens (2007) which only requires the series to exist. Finally, Assumption 12 is met for instance by selecting $a_n = \sqrt{\delta_n}$.

Proposition 3 (L^2 Rate of Convergence)

With the assumptions of Proposition 2 and Assumptions 10 through 12,

$$E||\hat{\theta} - \theta_0||_{\mathcal{X}}^2 = O(\sqrt{\delta_n})$$

Comments: To get some feeling for the speed of convergence, consider the case where $h_v = h_w = h$ with $h \propto n^{-1/(2r+c)}$ where $c \equiv |V| + 1$ which yields the optimal MSE rate $\delta_n \propto n^{-2r/(2r+c)}$ for $\hat{f}_{ZW}$. Then, Proposition

3 implies that $||\hat{\theta} - \theta_0||_{\mathcal{X}} = O_p(n^{-r/(4r+2c)})$. From this, one can understand that $\hat{\theta}$ inherits the same curse of dimensionality affecting a nonparametric density estimator with a speed of convergence decelerating rapidly as the number of continuous variables in Z increases. However, one may obtain $||\hat{\theta} - \theta_0||_{\mathcal{X}} = O_p(n^{-1/4})$ whenever the functions of Assumptions 8 are infinitely smooth. It is not clear whether this rate is optimal under the polynomial source conditions of Assumption 11 because the optimal rate derived in Chen and Pouzo (2008) in the polynomial case is based upon a different set of assumptions.

4 Monte Carlo experiments

This section examines the finite sample properties of the Tikhonov estimator in the context of the model

$$Y = (G \circ \theta_0)(X) + \epsilon \text{ with } E[\epsilon|W] = 0 \text{ a.s.}$$

where G is a known "link function". In this experiment $G(t) \equiv t - 0.5t^2$. The triplet (ϵ, X, W) is supported on $[-1, 1] \times [0, 1]^2$ with a density function meeting

$$f(e, x, w) \propto 2.45 \sum_{j=1}^{\infty} j^{-a} \frac{(2j+1)}{4j}(1 - e^{2j})\phi_j(x)\phi_j(w) + 2^{-5/2} \sum_{j=1}^{\infty} j^{-d}\phi_j(e+1)\phi_j(2x)\phi_j(w),$$

where $\phi_j(t) \equiv \sqrt{2}\sin(j\pi t)$, $a = 2.45$ and $d = 10$. This density, used in Horowitz and Lee (2007), is convenient because it renders the task of deriving the eigensystem of T_1 less cumbersome. In this Monte Carlo experiment, two designs are considered:

Design 1:

$$\theta_0(x) \equiv \sum_{j=1}^{\infty} (-1)^{j+1} j^{-2a} \phi_j(x).$$

Design 2:

$$\theta_0(x) \equiv \{\sum_{j=1}^{\infty} (-1)^{j+1} j^{-2a} \phi_j(x)\} \mathbf{1}_{[|x-0.5|>0.2]} + 1.3\mathbf{1}_{[|x-0.5|\leq 0.2]}.$$

The function of design 1 is continuous while the function to estimate in design 2 is discontinuous at $x = 0.5 \pm 0.2$. This model has $T\theta \equiv \mathcal{T}\theta - M$ where,

$$(\mathcal{T}\theta)(w) \equiv \int_{[0,1]} (G \circ \theta)(x) f_{XW}(x, w) dx,$$

$$M(w) \equiv E[Y|W = w]f_W(w)$$

and,

$$(T_1\varphi)(w) \equiv \int_{[0,1]} \varphi(x)(G^{(1)} \circ \theta_0)(x) f_{XW}(x, w)dx.$$

Here, $c = 2$ and the Fréchet kernel is given by $F(f_{XW}, \theta_0)(x, w) \equiv (G^{(1)} \circ \theta_0)(x) f_{XW}(x, w)$ which satisfies the conditions imposed in this paper. The estimation of T is conducted with

$$k(t) = (15/16)(1 - t^2)^2 \mathbf{1}_{[|t|\leq 1]},$$

which is a kernel of order $r = 2$.

The bandwidths conditions of this article are only qualitative. This paper explores a Silverman's like rule of thumb (Silverman 1986) to estimate the density function of (X, W) and $M(w)$. Let R_z denote the interquartile range between the third and first quartile of sample z. The rule of thumb consists of using $h_x = min(\hat{\sigma}_x, R_x/1.37)n^{-1/(2r+c)}$ and, $h_w = min(\hat{\sigma}_w, R_w/1.37)n^{-1/(2r+c)}$ where $\hat{\sigma}_x$ is the sample standard deviation of $\{X_i\}_{i=1..n}$ and, $\hat{\sigma}_w$ the sample standard deviation of $\{W_i\}_{i=1..n}$. Even though this rule for the bandwidths does not a priori satisfy any optimal criteria in the context of our specific problem, it has the benefit of being easy to implement while performing reasonably well compared to other choices used in preliminary experiments.

The regularization sequence $a_n = n^{-r/(2r+c)}$ meets the conditions of section 3. The Tikhonov estimator is computed iteratively using the Levenberg-Marquardt algorithm given in Engl and Kügler (2005) with $\phi_1(x)$ as the starting function. Due to the long computational time required, the series are truncated with 100 terms and 100 replications are carried out. The simulations are conducted in Gauss. Table 1 shows $IMSE \equiv E\int_{[0,1]}\{\hat{\theta}(x) - \theta_0(x)\}^2 dx$ and $IBS \equiv \int_{[0,1]}\{E\hat{\theta}(x) - \theta_0(x)\}^2 dx$. The numerical results yield approximately a 20 percent decline in IMSE for a quadrupling of the sample size. This agrees with proposition 3 which gives $IMSE = O(n^{-1/3})$ when $r = 2$ and $c = 2$. As shown in Figures 2-9, the Tikhonov estimator is fairly close in shape to the true parameter on average.

Figure 1: Marginal density function of X or W (green line), density function of $X|W = 0.4$ (blue line) and density function of $X|W = 0.6$ (red line).

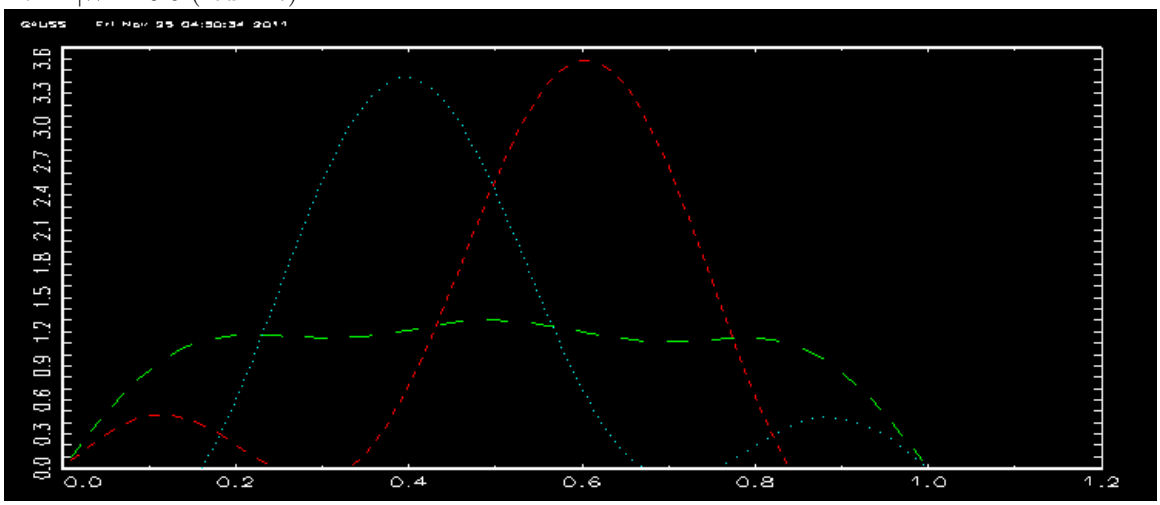


Table 1: Losses of Monte Carlo Experiments

n=100	**Design 1**	**Design 2**
	IMSE—IBS	**IMSE—IBS**
a_n	0.0344—0.0034	0.0360—0.0058
$2a_n$	0.0427—0.0040	0.0321—0.0055
n=400		
a_n	0.0247—0.0027	0.0285—0.0056
$2a_n$	0.0261—0.0028	0.0238—0.0050

Figure 2: Design 1,n=100 and a_n: $\theta_0(x)$ (dashed line) Versus $E\hat{\theta}(x)$ (dotted line).

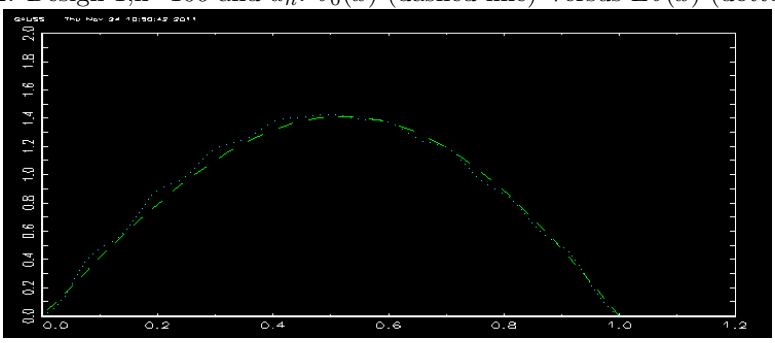


Figure 3: Design 1,n=100 and $2a_n$: $\theta_0(x)$ (dashed line) Versus $E\hat{\theta}(x)$ (dotted line).

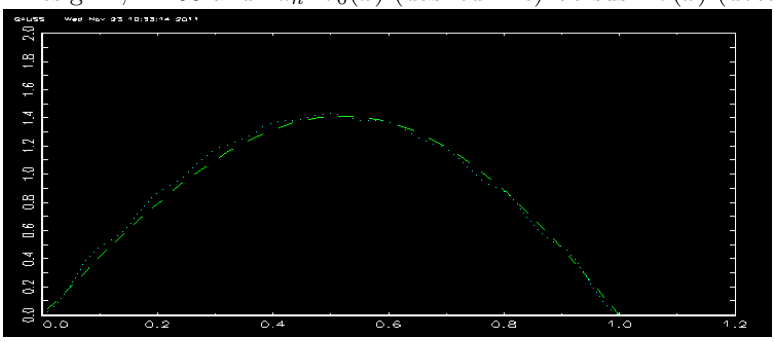


Figure 4: Design 1,n=400 and a_n: $\theta_0(x)$ (dashed line) Versus $E\hat{\theta}(x)$ (dotted line).

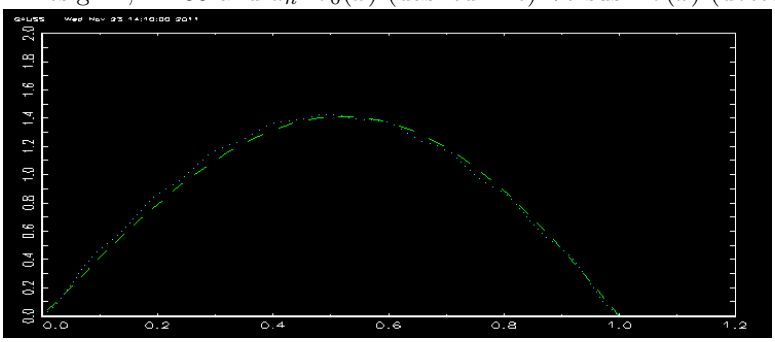


Figure 5: Design 1,n=400 and $2a_n$: $\theta_0(x)$ (dashed line) Versus $E\hat{\theta}(x)$ (dotted line).

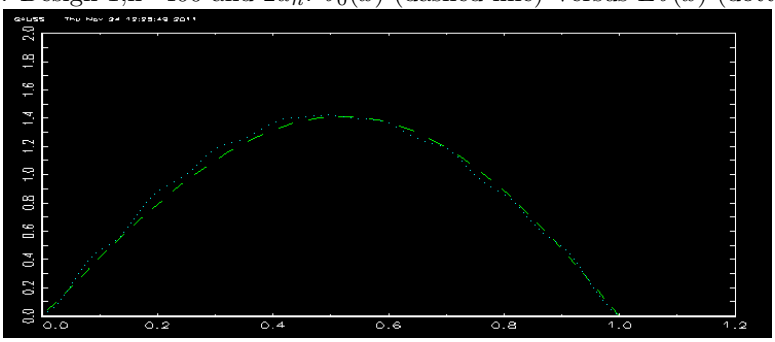


Figure 6: Design 2,n=100 and a_n: $\theta_0(x)$ (dashed line) Versus $E\hat{\theta}(x)$ (dotted line).

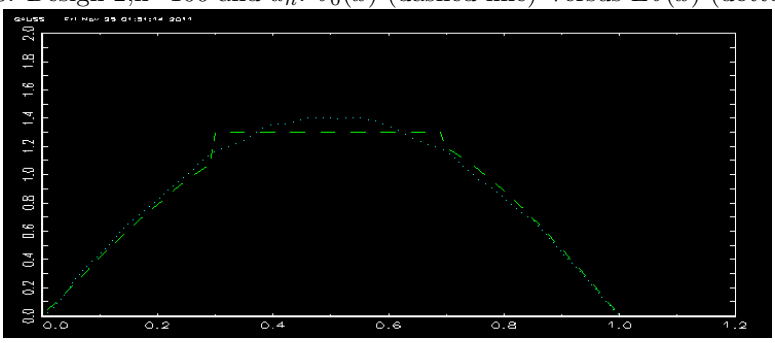


Figure 7: Design 1,n=100 and $2a_n$: $\theta_0(x)$ (dashed line) Versus $E\hat{\theta}(x)$ (dotted line).

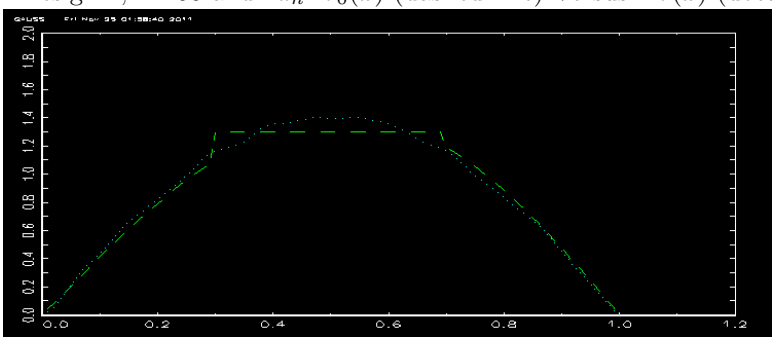


Figure 8: Design 2,n=400 and a_n: $\theta_0(x)$ (dashed line) Versus $E\hat{\theta}(x)$ (dotted line).

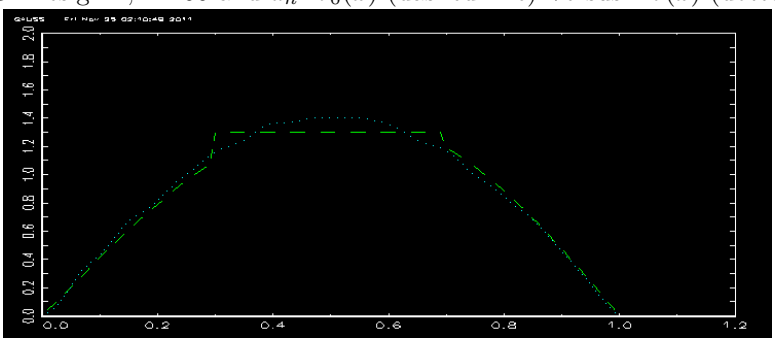
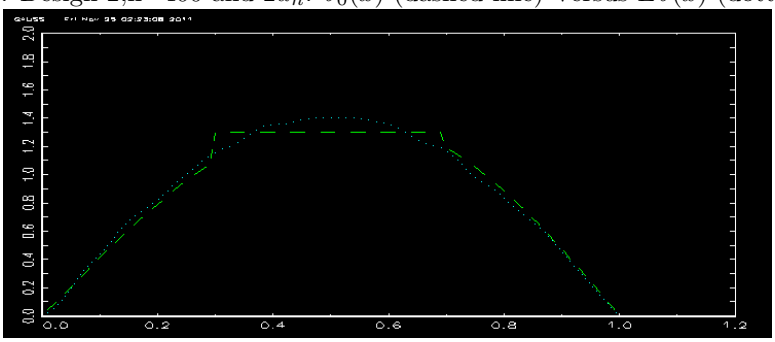


Figure 9: Design 2,n=400 and $2a_n$: $\theta_0(x)$ (dashed line) Versus $E\hat{\theta}(x)$ (dotted line).



Conclusion

This paper has presented a methodology for estimating an infinite dimensional parameter where the parameter is identified by a conditional moment restriction using a continuous instrument. There are two questions arising from this paper. First, more research needs to be pursued concerning the optimal choice for the regularization sequence. Another important problem deals with the numerical effect, in a finite sample, of using a space of large albeit finite dimension for computing the estimator. Some asymptotic results for Sieves estimation given in Chen (2007) suggest that the quality of the numerical approximation will be partly determined by the smoothness of the objective, the norm of the Fréchet derivative in that case, and the type of smoothness satisfied by the unknown parameter since this latter influences how accurately the solution can be approximated.

References

Ai, C., and Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* Vol. 71, No. 6.

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

Carrasco, M., Florens, J.P., and Renault, E., 2007. Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. *Handbook of Econometrics Vol. 6*, E.E. Leamer and J.J. Heckman, (eds), Amsterdam: North-Holland.

Carrasco, M., and Florens, J.P., 2010. A spectral method for deconvolving a density. *Econometric Theory*, available on CJO 11 Oct 2010.

Chen X., Pouzo D., 2008. On nonlinear ill-posed inverse problems with applications to pricing of defaultable bonds and option pricing. *Science in China. Series A, Mathematics.*

Chernozhukov, V., Gagliardini, P., and Scaillet, O., 2009. *Nonparametric Instrumental Variable Estimation of Quantile Structural Effects.* Swiss Finance Institute Research Paper No. 08-03.

Darolles, S., Florens, J.P. and Renault, E., 2006. *Non Parametric Instrumental Regression.* Working paper, GREMAQ, University of Social Science, Toulouse, France.

Engl, H.W., Kunisch, K., and Neubauer, 1989. Convergence rates for Tikhonov regularization of non linear ill-posed problems. *Inverse Problems* 5, 523-540.

Engl, H.W., and Kügler, P., 2005. Nonlinear inverse problems: Theoretical aspects and some industrial applications. In: *Multidisciplinary Methods for Analysis Optimization and Control of Complex Systems Mathematics in Industry, Volume 6, Part I*, pp. 3–47.

Florens, J.P., Johannes, J., and Van Bellegem, S. 2005. *Instrumental Regression in Partially Linear Models.* CORE Discussion Paper.

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

Horowitz, J. and Lee, S., 2007. Non parametric instrumental variables estimation of a quantile regression model. *Econometrica* Vol. 75, No. 4.

Judge, G., Griffiths, W., Hill, R.C., Lutkepohl, H., and Lee, T.-C., 1980. *The Theory and Practice of Econometrics.* New York: Wiley.

Kress, R., 1999. *Linear Integral Equations. 2nd edition.* Berlin: Springer-Verlag.

Newey, W., and Powell, J., 2003. Instrumental variable estimation of nonparametric models. *Econometrica* Vol. 71, No. 5.

Pagan, A., and Ullah, A., 1999. *Non Parametric Econometrics.* Cambridge Univ. Press.

Robinson, P., 1988. Root N consistent semi parametric regression. *Econometrica* Vol. 156.

Silverman, B.W., 1986. *Density Estimation.* London: Chapman and Hall.

Appendix

This section provides the proofs of the propositions. The notation $||.||$ will refer to either $||.||_{\mathcal{X}}$ or $||.||_{[0,1]}$ depending on the context. Likewise for the inner product $<,>$. The notation lim is to be understood as n approaches infinity.

Proposition 1

First, notice that by Assumptions 1–4 $T : \mathcal{H} \to L^2(\mathcal{W}, Leb)$ defined by $T(\theta)(w) \equiv \int_{\mathcal{Z}} \rho_\theta(z) f_{ZW}(z,w) d\mu(z)$ exists since by the Cauchy–Schwartz inequality and Fubini's Theorem:

$$||T\theta||^2 \leq \int_{\mathcal{Z}} |\rho_\theta(z)|^2 d\mu(z) \int_{\mathcal{W}} \int_{\mathcal{Z}} |f_{ZW}(z,w)|^2 d\mu(z) dw < \infty,$$

because $\int_{\mathcal{Z}} |\rho_\theta(z)|^2 d\mu(z) < \infty$ by Assumption 4(b) and $\int_{\mathcal{W}} \int_{\mathcal{Z}} |f_{ZW}(z,w)|^2 d\mu(z) dw < \infty$ by Assumption 4(a).

Now let $\theta \equiv \theta_0 + \Delta$ where $||\Delta|| \neq 0$. This implies $T(\theta) = D(\Delta)$ for some linear operator D by Assumption 5(b). But $||T(\theta)|| = ||\Delta||.||D(\frac{\Delta}{||\Delta||})||$ which shows $||T(\theta)|| > 0$ by Assumption 5(c). The proposition follows directly since $T(\theta_0)(w) = 0$ for almost every w in $\mathcal{W}$ by Assumption 5(a).

Lemma 1: Under the assumptions of Proposition 2,

(a) $E||\hat{T}\theta - T\theta||^2 = O(\delta_n)$ uniformly over $\mathcal{H}$.

(b) T is weakly sequentially closed on $\mathcal{H}$.

proof:

(a) Using $|\hat{T}\theta(w) - T\theta(w)| \leq \int_{\mathcal{S}\times[0,1]^{|V|}} |\rho_\theta(s,v)||\hat{f}(s,v,w) - f(s,v,w)| d\lambda(s) dv$ and the Cauchy–Schwartz inequality yields

$||\hat{T}\theta - T\theta||^2 \leq sup_{\{\theta\in\mathcal{H}\}} ||\rho_\theta||^2_{\mathcal{S}\times[0,1]^{|V|}} \int_{[0,1]} \int_{\mathcal{S}} \int_{[0,1]^{|V|}} |\hat{f}(s,v,w) - f(s,v,w)|^2 dv d\lambda(s) dw.$

Now $E|\hat{f}(s,v,w) - f(s,v,w)|^2 = \{E\hat{f}(s,v,w) - f(s,v,w)\}^2 + Var\hat{f}(s,v,w)$. By Assumption 8, one can use a change of variable along with a multivariate Taylor expansion to rapidly establish that $E\hat{f}(s,v,w) - f(s,v,w) = O(Max(h_v, h_w)^r)$ uniformly over $\mathcal{S} \times [0,1]^{|V|+1}$. Another change of variable using the fact that $f(s,v,w)$ is bounded by Assumption 8 yields $Var\hat{f}(s,v,w) = O(1/nh_w h_v^{|V|})$ uniformly over $\mathcal{S} \times [0,1]^{|V|+1}$. Thus, $E|\hat{f}(s,v,w) - f(s,v,w)|^2 = O(\delta_n)$ uniformly over $\mathcal{S} \times [0,1]^{|V|+1}$. This shows the claim because $||\rho_\theta||^2_{\mathcal{S}\times[0,1]^{|V|}} < \infty$ by Assumption 6(b) and $\lambda(\mathcal{S}) < \infty$ by Assumption 7(a).

(b) It is clear that $\mathcal{H}$ is convex. Also, $\mathcal{H}$ is closed with respect to $||.||$ since any sequence $\{\theta_n\}_{n\geq 1} \subseteq \mathcal{H}$ converging in the mean to some θ will satisfy $M \geq \lim ||\theta_n|| = ||\theta||$. Hence, T will be weakly sequentially closed if it is also weakly continuous. To show that T is weakly continuous, consider any sequence $\{\theta_n\}_{n\geq 1} \subseteq \mathcal{H}$ such that $\theta_n \rightsquigarrow \theta$ and $T(\theta_n) \rightsquigarrow m$. It is easy to show that the weak convergence of θ_n leads to $\lim T\theta_n - T\theta = 0$ up to a Lebesgue null set by Assumption 6(a). Hence, $\lim < T\theta_n - T\theta, \varpi >= 0$ for any $\varpi \in L^2(\mathcal{W}, Leb)$ by the Lebesgue Dominated Convergence Theorem whose dominance condition holds under the assumption of Proposition 1 due to Assumption 6(b). This last result shows that T is weakly continuous, which concludes the proof.

Proposition 2:

Using the same reasoning as in Lemma 1 (b), one can establish that $\hat{T}$ is also weakly continuous. Since $\mathcal{H}$ is closed and convex by Lemma 1, this ensures that $\hat{\theta}$ exists by Bissantz, Hohage, and Munk (2004). By definition, $||\hat{T}\hat{\theta}||^2 + a_n||\hat{\theta}||^2 \leq ||\hat{T}\theta_0||^2 + a_n||\theta_0||^2$. Also, $E||\hat{T}\theta_0||^2 = E||\hat{T}\theta_0 - T\theta_0||^2$ which gives $E||\hat{T}\theta_0||^2 = O(\delta_n)$ by Lemma 1 and consequently

$$E||\hat{T}\hat{\theta}||^2 \leq O(\delta_n) + a_n M, \tag{1}$$

$$a_n E||\hat{\theta}||^2 \leq O(\delta_n) + a_n M, \tag{2}$$

for some real number M which exists by Assumption 2(a). Now the triangle inequality yields

$$E||T\hat{\theta}||^2 \leq 2\{E||T\hat{\theta} - \hat{T}\hat{\theta}||^2 + E||\hat{T}\hat{\theta}||^2\},$$

with $E||T\hat{\theta} - \hat{T}\hat{\theta}||^2 = O(\delta_n)$ by Lemma 1 and $E||\hat{T}\hat{\theta}||^2 = O(\delta_n) + O(a_n)$ by (1). It follows from these that $E||T\hat{\theta}||^2 = O(\delta_n) + O(a_n)$ and $E||\hat{\theta}||^2 \leq O(\delta_n/a_n) + ||\theta_0||^2$. Thus, invoking Assumption 9(a) yields

$$\lim E||T\hat{\theta}||^2 = 0 \text{ and limsup } E||\hat{\theta}||^2 \leq ||\theta_0||^2.$$

Because T is weakly sequentially closed by Lemma 1, the above result ensures that $\lim E||\hat{\theta} - \theta_0||^2 = 0$ by Theorem 2 of Bissantz, Hohage, and Munk (2004).

Proposition 3:

The proof is based upon the argument of Theorem 3(ii) of Bissantz, Hohage, and Munk (2004).

By definition, $||\hat{T}\hat{\theta}||^2 + a_n||\hat{\theta}||^2 \leq ||\hat{T}\theta_0||^2 + a_n||\theta_0||^2$. Adding $a_n||\hat{\theta} - \theta_0||^2 - a_n||\hat{\theta}||^2$ to both sides yields

$$||\hat{T}\hat{\theta}||^2 + a_n||\hat{\theta} - \theta_0||^2 \leq ||\hat{T}\theta_0||^2 + a_n[||\theta_0||^2 + ||\hat{\theta} - \theta_0||^2 - ||\hat{\theta}||^2],$$

and simplifying $||\theta_0||^2 + ||\hat{\theta} - \theta_0||^2 - ||\hat{\theta}||^2$ using $||\hat{\theta} - \theta_0||^2 = ||\hat{\theta}||^2 + ||\theta_0||^2 - 2 < \hat{\theta}, \theta_0 >$ yields

$$||\hat{T}\hat{\theta}||^2 + a_n||\hat{\theta} - \theta_0||^2 \leq ||\hat{T}\theta_0||^2 + 2a_n < \theta_0, \theta_0 - \hat{\theta} >, \tag{3}$$

Introduce $\varpi \equiv (T_1 T_1^*)^{-1} T_1 \theta_0$. Using the spectral decomposition of T_1 gives

$$\varpi = \sum_{j=1}^{\infty} \frac{< \theta_0, \phi_j >}{\lambda_j} \psi_j,$$

and consequently,

$$||\varpi|| = (\sum_{j=1}^{\infty} \frac{| < \theta_0, \phi_j > |^2}{|\lambda_j|^2})^{1/2} < 1/2L$$

for some $L > 0$ which exists by Assumption 11. Thus, ϖ belongs to $L^2([0,1], Leb)$ and since, by Assumption 10(c), the operator T_1 is injective,

$$T_1^* \varpi = \theta_0 \text{ with } ||\varpi|| \leq 1/2L,$$

(4)

Using (4) and $< T_1^* \varpi, \theta_0 - \hat{\theta} >=< \varpi, T_1(\theta_0 - \hat{\theta}) >$ permits us to rewrite (3):

$$||\hat{T}\hat{\theta}||^2 + a_n||\hat{\theta} - \theta_0||^2 \leq ||\hat{T}\theta_0||^2 + 2a_n < \varpi, T_1(\theta_0 - \hat{\theta}) > .$$

(5)

Now $| < \varpi, T_1(\theta_0 - \hat{\theta}) > | \leq ||\varpi||.||T_1(\theta_0 - \hat{\theta})||$ by the Cauchy–Schwartz inequality. Also, $T_1(\hat{\theta} - \theta_0) = T(\hat{\theta}) - R$ with $||R|| \leq L||\hat{\theta} - \theta_0||^2$ by Assumption 10(a). These two facts along with the Minkowski's inequality give

$$| < \varpi, T_1(\theta_0 - \hat{\theta}) > | \leq ||\varpi||.[||T\hat{\theta}|| + L||\theta_0 - \hat{\theta}||^2].$$

(6)

Using (6) in (5) yields

$$||\hat{T}\hat{\theta}||^2 + a_n||\hat{\theta} - \theta_0||^2 \leq ||\hat{T}\theta_0||^2 + 2a_n||\varpi||.||T\hat{\theta}|| + 2a_n L||\varpi||.||\theta_0 - \hat{\theta}||^2,$$

(7)

and the triangle inequality, $||T\hat{\theta}|| \leq ||T\hat{\theta} - \hat{T}\hat{\theta}|| + ||\hat{T}\hat{\theta}||$ used in the right hand side of (7) yields

$$||\hat{T}\hat{\theta}||^2 + a_n||\hat{\theta} - \theta_0||^2 \leq ||\hat{T}\theta_0||^2 + 2a_n||\varpi||.||T\hat{\theta} - \hat{T}\hat{\theta}|| + 2a_n||\varpi||.||\hat{T}\hat{\theta}|| + 2a_n L||\varpi||.||\theta_0 - \hat{\theta}||^2$$

(8)

This last result can be condensed:

$$A_n + C_n \leq B_n,$$

where

$A_n \equiv ||\hat{T}\hat{\theta}||^2 - 2a_n||\varpi||.||\hat{T}\hat{\theta}||$

$B_n \equiv ||\hat{T}\theta_0||^2 + 2a_n||\varpi||.||T\hat{\theta} - \hat{T}\hat{\theta}||$

$C_n \equiv a_n||\hat{\theta} - \theta_0||^2 - 2a_n L||\varpi||.||\theta_0 - \hat{\theta}||^2$

Now factor $C_n = a_n||\hat{\theta} - \theta_0||^2(1 - 2||\varpi||L)$ and use elementary algebra to get $A_n = (||\hat{T}\hat{\theta}|| - a_n||\varpi||)^2 - a_n^2||\varpi||^2$ resulting in

$$(||\hat{T}\hat{\theta}|| - a_n||\varpi||)^2 + a_n||\hat{\theta} - \theta_0||^2(1 - 2||\varpi||L) \leq a_n^2||\varpi||^2 + ||\hat{T}\theta_0||^2 + 2a_n||\varpi||.||T\hat{\theta} - \hat{T}\hat{\theta}||$$

(9)

From (9), use $\xi \equiv 1 - 2||\varpi||L > 0$ by (4) and take the expected value on both sides:

$$a_n E||\hat{\theta} - \theta_0||^2 \leq \xi^{-1}[a_n^2||\varpi||^2 + E||\hat{T}\theta_0||^2 + 2a_n||\varpi||.E||T\hat{\theta} - \hat{T}\hat{\theta}||],$$

Since, by Lemma 1, $E||\hat{T}\theta_0||^2 = E||\hat{T}\theta_0 - T\theta_0||^2 = O(\delta_n)$ and $E||T\hat{\theta} - \hat{T}\hat{\theta}|| = O(\sqrt{\delta_n})$, using Assumptions 12 and 9(a) permits us to finally conclude

$$E||\hat{\theta} - \theta_0||^2 \leq \xi^{-1}[a_n||\varpi||^2 + O(\delta_n/a_n) + O(\sqrt{\delta_n})],$$

Thus, $E||\hat{\theta} - \theta_0||^2 \leq \sqrt{\delta_n}[O(a_n/\sqrt{\delta_n}) + O(\sqrt{\delta_n}/a_n) + O(1)]$ and Proposition 2 follows since $a_n/\sqrt{\delta_n} \asymp 1$ by Assumption 12.