Direct Instrumental Nonparametric Estimation of Inverse Regression Functions

Jerome M. Krief, University of Virginia *

March 14, 2014

Abstract

This paper is concerned about estimating the inverse g^{-1} of a monotonic function g which satisfies the restriction $E[Y - g(X)|W] = 0$ almost surely where (Y, X, W) are observable random variables and W is continuously distributed. As far as I know, consistent estimators are available only if $g(X) = E[Y|X]$ almost surely which precludes endogenous models. This paper proposes a direct estimator of g^{-1} when W is not functionally related to X. The method consists of identifying g^{-1} as a solution of a nonlinear integral equation whose underlying operator can be estimated nonparametrically. It is well-known that solving the analog does not deliver consistency due to the ill-posed problem phenomenon, namely the solution is discontinuous in data. To solve this, the estimator is regularized with the Tikhonov technique. Under weak smoothness conditions, I show that the estimator is MSE consistent under a global metric. Furthermore, if g^{-1} is smooth in a certain sense then the MSE rate of convergence is equal to $\aleph[n^{-r/(2r+1)}]$ where $\aleph$ is a continuous function with $\aleph(0) = 0$ and $r > 1$ denotes the number of derivatives assumed for some densities. The analytical expression of $\aleph$ depends on a link function characterizing the smoothness of g^{-1}. This result holds for a broad class of link functions belonging to a suitable space.

Key words: Instrumental nonparametric estimation, Inverse regression function, Nonlinear integral equations, Nonseparable regression, Tikhonov regularization.

*Department of Economics. I would like to thank Marine Carrasco, Peter Hall, Eric Renault, Adam Rosen and Elie Tamer for suggestions and encouragements. Also, I thank the committee members and the participants of the 2011 Info-Metrics Institute Workshop on Shrinkage at American University, Washington, DC.

1 Introduction

This article considers the model

$$E[Y - g(X)|W] = 0 \tag{1}$$

almost surely, where Y is a scalar variable, (X, W) is bivariate continuously distributed, and g is some unknown function strictly monotonic on the support of X. Without loss of generality, g shall be increasing. The multivariate extension (i.e $dim(X, W) > 2$) is delineated in Section 4. It is assumed that W is excluded from X. This paper presents a method to estimate g^{-1} in one step from data. Here are some examples where this methodology applies:

Example 1: Endogenous nonparametric demand-supply regression.

$$Y = g(X) + U,\ E[U|W] = 0,$$

where Y denotes the equilibrium quantity, U is an unobservable variable, and $g(X)$ represents the quantity demanded (respectively supplied) at price X. Standard economic theory dictates that g is strictly decreasing (respectively increasing). Economists are interested in estimating g^{-1} which represents the inverse demand (respectively supply) function. The market clearing forces render U correlated with X. The availability of some external variable W exogenous in the sense of $E[U|W] = 0$ offers a possible identifying relationship for g^{-1} as discussed in Section 2.

Example 2: Endogenous nonseparable regression.

$$X = M(Z, U),\ E[U|W] = 0.$$

where (X, Z, W) is observable, and $M(z, .)$ is strictly increasing in its second argument for all z in the support of Z. The unobservable variable U has an absolutely continuous distribution function, and a symmetrical density with respect to Lebesgue measure conditional on W. Because $M(Z, U) = M[Z, \Gamma^{-1} \circ \Gamma(U)]$ for any invertible Γ, identification of M is only possible up to an increasing function (Matzkin 2003). A convenient choice for this paper

is to select Γ to be some increasing odd function so that one may assume U has support $[-1,1]$[1]. This model can be re-framed as in (1). To understand this duality, write $M^{-1}(z,.)$ the inverse of $M(z,.)$ which yields

$$E\{E[M^{-1}(Z,X)|Z,W]|W\}=0.$$

Since $M(Z,.)$ is the inverse of $M^{-1}(Z,.)$, one can estimate M using the multivariate method described in Section 4. Matzkin (2004) developed an estimator of M assuming that U is independent with Z while Chernozhukov, Gagliardini and Scaillet (2012) and Chernozhukov, Imbens, Newey (2007) proposed estimators of M assuming U is independent with W [2]. The method of this paper can accommodate the case where Z and U are correlated, and allows the distribution of $U|W$ to be heteroscedastic[3].

Example 3: Weighted derivatives estimation for the endogenous nonparametric regression with unknown transformation of latent equation.

$$Y=T[g(X)+U],\ E[U|W]=0.$$

where (Y,X,W) is observable, U is an unobservable error, T is strictly increasing and absolutely continuous, and g is some unknown function. This models notably encompasses the nonparametric transformation of Box and Cox (1964), proportional hazard models, and additive hazard models (see Horowitz 1996). The above model was examined in Jacho-Chavez-Linton Lewebel (2010) when X is uncorrelated with U. Also, Chiappori et al (2014) proposed an estimation method if X contains a 'special' regressor which is independent with U conditional on the remaining components.

[1]For instance, pick Γ to be some continuously differentiable increasing odd function onto $[-1,1]$ with a derivative supported on a compact interval inside of which the density of U is assumed bounded away from the origin. The normalized model is $Y=\tilde{M}(Z,\tilde{U})$ where $\tilde{U}\equiv\Gamma(U)$ and $\tilde{M}(Z,.)\equiv M[Z,\Gamma^{-1}(.)]$. This is convenient because the methodology of Section 4 requires $\tilde{U}$ to be supported on a known compact interval.

[2]If $U\perp W$, the density of $U|W$ does not need to be symmetrical because one can select Γ to be the CDF of U which results, after normalization, in $E\{E[M^{-1}(Z,X)|Z,W]|W\}=1/2$. If this is the case, the estimator described in this paper in effect conducts an extra smoothing step over the identifying relationship examined in Chernozhukov, Imbens, Newey (2007)

[3]Furthermore, M is estimated as a function of both arguments for components of Z which are not redundant in W (see Theorem 1-2bis). In the former papers, $M(z,u)$ is estimated for each fixed u.

Without loss of generality suppose X is a scalar variable with support $[0,1]$. Let $\pi(x)$ denote a positive weight function with compact support $S \subseteq [0,1]$ vanishing on its boundaries. Suppose g and π are continuously differentiable on S. Researchers are often interested in estimating the weighted derivative parameter $\delta \equiv \int_S \pi(x)\{\nabla g(x)\}dx$. This latter can be estimated with the method described in this paper. To appreciate this, introduce $F(u,x) \equiv T[g(x)+u]$ and adopt the normalization $\int \nabla T(z)dz = 1$. Write $D(x,x_o) \equiv \int F(u,x) - F(u,x_o)du$ for any (x,x_o) in $[0,1]^2$. Now observe that the Fubini's theorem and integrating by parts yield

$$\int_{[0,1]}\{\nabla\pi(x)\}D(x,x_o)dx \equiv -\int_{[0,1]}\pi(x)\{\nabla g(x)\}dx.$$

Hence, $\delta = -\int_{[0,1]^2}\{\nabla\pi(x)\}D(x,x_o)dxdx_o$. It follows that δ can be estimated consistently forming the analog replacing $F(u,x)$ by the estimator constructed in a first step as explained in example 2.

If g is known up to some finite dimensional parameter one can estimate g^{-1} consistently using GMM method (Hansen 1982) under mild smoothness conditions. If g does not belong to the parametric family in question, this estimator is inconsistent. A limitation of this parametric approach is that it is solely grounded in analytical easiness because economic theory rarely imposes more than shape restrictions (i.e. monotonicity, concavity-convexity). If g is not assumed to be known up to some finite dimensional parameter, a natural approach for estimating g^{-1} consists of using the generalized inverse of some consistent nonparametric estimator. The consistent estimation of g can be achieved using either the sieves method (Ai and Chen 2003) or the regularization method (Hall and Horowitz 2005, Darolles, Fan, Florens, and Renault 2011). However, inverting these estimators in a second step does not permit to derive a rate of convergence given the metrics available for the above estimators. This paper solves this problem by proposing an instrumental nonparametric estimator of g^{-1} whose rate of convergence is established. Furthermore, the estimator is not produced by inverting a preliminary estimator of g but is directly computed from data.

As explained in section 2, g^{-1} solves a nonlinear integral equation $T\theta = Q$ where the operator T and the function Q are known up to some distributions. T and Q can be estimated from data with nonparametric methods. This paper uses the kernel method although other approaches are possible. Solving the analog does not imply 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). Technically, the ill-posedness arises owing to the fact that the Frechet

derivative of the underlying operator has a discontinuous inverse. Unlike the case where the function of interest is known up to some finite-dimensional parameter (see Hansen 1982, Judge et al 1980), this lack of continuity 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 the estimation losses on the sequence of coefficients characterizing the function are cumulated *ad* infinity.

To overcome this challenge, one may regularize the solution, that is modify it into a continuous mapping. Various regularization schemes are available but this paper applies the Tikhonov method. This technique achieves regularization by penalizing the objective when forming the analog. This latter feature is convenient because it removes the need to work from the first-order conditions, a theoretical challenge due to the ill-posedness previously mentioned. If some sequence of Fourier coefficients representing g^{-1} decays sufficiently fast, and if certain densities admit $r > 1$ derivatives then the proposed estimator say $\hat{\theta}$ satisfies $E||\hat{\theta} - g^{-1}||_H^2 = O\{\aleph[n^{-r/(2r+1)}]\}$ where $\aleph$ is an increasing continuous function such that $\aleph(0) = 0$, and $||.||_H$ is a global metric. The exact analytical expression of $\aleph$ hinges on the type of source conditions, namely on a link function characterizing the speed of decay exhibited by the Fourier coefficients. This paper uses generic source conditions in the sense that the link function is only assumed to belong to a suitable space. This is important because rate-optimality conclusions depend on the link function. The rate of estimation established in this paper is slower than that achieved for estimating T and Q. This deceleration is inevitable in order to remove the bias introduced by the regularization process. The same phenomenon takes place when estimating g (Hall and Horowitz 2005, Carrasco and Florens 2010) with the Tikhonov regularization method. This MSE rate of convergence derived in Theorem 2-bis is the fastest one possible under the assumptions imposed in this paper (Sergei, Pereverzev and Ramlau 2007).

The nonlinear Tikhonov regularization technique has already been employed in econometrics for the nonparametric endogenous quantile regression model (Horowitz and Lee 2007) and the endogenous non-additive regression model (Chernozhukov, Gagliardini and Scaillet 2012). In fact, the structure of the integral equation identifying g^{-1} shares similarities with that examined in these two papers. Estimating the inverse of a monotonic regression function $E[Y|X]$ (i.e. when $W = X$) has been examined in statistics. These include the inverse kernel method (Dette et al. 2005), and the two-stage hybrid method (Tang at al. 2011). As far as I know, there is no method available to estimate the inverse when $g(X)$ does not coincide with $E[Y|X]$. This latter fact frequently arises in

economics due to simultaneousness of data generating processes or omitted variables.

Regularizing is not the only way one can overcome the ill-posed problem. Another strategy consists of working on a compact parameter space. The resulting estimator will not be subject to ill-posedness because the solution is continuous in data by the Arzela's theorem (see Gallant and Nychka 1987, Engl and Kugler 2005, Newey and Powell 2003, Ai and Chen 2003). The main difference with the regularization method deals with the burden of assumptions concerning the function to estimate. With the regularization method, the burden is on the speed of decay of the sequence of Fourier coefficients characterizing the function. On the other hand, the 'compactification' approach hinges on the existence of a sufficient large number of pointwise derivatives for the function.

The rest of the paper is organized as follows. Section 2 provides a summary of the estimation procedure. Section 3 presents the assumptions and results. Section 4 extends the method for the multivariate case. Finally, Section 5 presents a Monte Carlo experiment. All the proofs are located in the appendix located in the back of the paper.

I shall introduce some notations used in the subsequent sections. For any integer $c \geq 1$, $L^2[0,1]^c$ denotes the space of square integrable functions on $[0,1]^c$ with respect to Lebesgue measure. For any f and g belonging to $L^2[0,1]^c$ write $< f,g >\equiv \int_{[0,1]^c} f(x)g(x)dx$ and $||f||_{[0,1]^c} \equiv \sqrt{< f,f >}$. For an integer $\alpha \geq 0$, write $\nabla^{(\alpha)} f$ as the α^{th} weak derivatives whenever the latter exist. To make notations less cumbersome, $\nabla^{(1)} f$ will be simply written as ∇f. Furthermore, for any integer $s \geq 1$, write $H^s[0,1]^c = \{f \in L^2[0,1]^c : \nabla^\alpha f \in L^2[0,1]^c, |\alpha| \leq s\}$ viewed as a completion. Any θ belonging to $H^s[0,1]^c$ with $2s > c$ shall be assumed continuous without loss of generality because one can always redefine θ a.e. by

$$\theta(x_1,....,x_c) \equiv \theta(0,...,0) + \int_0^{x_1} \nabla_1 f(z,x_2,...x_c)dz + \int_0^{x_2} \nabla_2 f(0,z,x_3...,x_c)dz + ... + \int_0^{x_c} \nabla_c f(0,...,0,z)dz.$$

To make notations less cumbersome, I shall write $< f,g >_1 =< f,g > + \sum_{|\alpha|=1} < \nabla^{(\alpha)} f, \nabla^{(\alpha)} g >$ and $||f||_1 \equiv \sqrt{< f,f >_1}$. Moreover, $H^1[0,1]^c$ will be simply written as $H[0,1]^c$. The symbol $\partial[0,1]^c$ denotes the subset of $[0,1]^c$ containing points $(x_1,...x_c)$ such that $x_i \in \{0,1\}$ for some $i = 1...c$. Given a strictly positive deterministic sequence $\{a_n\}_{n\geq 1}$ and a deterministic 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 Estimation Procedure

This section treats the case where X and W are scalar variables. It is assumed that X is supported on a compact interval. Furthermore, I shall suppose for simplicity $(g(X), W) \in [0,1]^2$. If $g(X)$ has a bounded support, using the unit interval may be assumed without loss of generality[4]. The methodology can be adjusted if the support of $g(X)$ is unbounded as discussed in Section 3. Moreover, write $||\theta||_H^2 \equiv ||\theta||^2 + ||\nabla\theta||^2$, and assume $\theta_0 \equiv g^{-1}$ belongs to $\mathcal{H} = \{\theta \in H[0,1] : ||\theta||_H^2 \leq M\}$ where $M < \infty$ is a known real constant.

let $\{Y_i, X_i, W_i\}_{i=1}^n$ denote some i.i.d. sequence of observations. It is assumed that the distribution of (X, W) is absolutely continuous with respect to Lebesgue measure. Furthermore, (Y, W) is assumed to have a density $J(y, w)$ with respect to some sigma-finite product measure $\mu \times \ell$ where ℓ denotes the Lebesgue measure. Write f_{XW} as the density of (X, W), and f as the density of W. Furthermore, let $F[.|w]$ denote the cumulative distribution function (**CDF**) of $X|W = w$.

To appreciate the method observe that (1) implies

$$E[g(X)|W = w] = E[Y|W = w] \tag{2}$$

Furthermore, the Fubini's Theorem yields

$$E[g(X)|W = w] = \int_0^1 P[g(X) > t|W = w]dt = 1 - \int_{[0,1]} F[g^{-1}(t)|w]dt. \tag{3}$$

Let τ denote a weight function with compact support $I \subseteq [0,1]$. The presence of a weight function is introduced for technical reasons which will made more precise in Section 3. Additionally, a weight function can improve the finite-sample properties of the estimator when the density of W is too small close to the boundaries. Furthermore, define

$$\Omega[x, w] \equiv \int_{-\infty}^x f_{XW}(u, w)\tau(w)du, \tag{4}$$

[4]If $g(X)$ is bounded (wp 1) use the transformed model $(\alpha Y + \beta, \alpha g + \beta)$ where (α, β) are real constants so that $\tilde{g}(X) \equiv \alpha g(X) + \beta$ has support in $[0,1]$. Since X has support say $[a, b]$, one may redefine the inverse if needed putting $\tilde{g}^{-1} = a$ if $t < \tilde{g}(a)$ and $\tilde{g}^{-1} = b$ if $t > \tilde{g}(b)$.

and

$$M(w) \equiv \int yJ(y,w)d\mu(y). \tag{5}$$

Write $Q(w) \equiv \tau(w)\{f(w) - M(w)\}$. It follows from (2)-(5) that θ_0 solves the following nonlinear integral equation

$$(T\theta)(w) = Q(w) \tag{6}$$

where,

$$(T\theta)(w) \equiv \int_{[0,1]} \Omega[\theta(t), w]dt.$$

A prerequisite to reach consistency is the uniqueness of a solution in (6) over $\mathcal{H}$. As of today, only results for local identification are available (Chen, Chernozhukov, Lee and Newey 2011)[5]. A necessary (albeit not sufficient) condition for global identification is that X and W exhibit statistical dependency. To understand this point, observe that if X and W are independent then T transforms any function into τf up to a multiplicative scalar. Thus, I assume identification in the sense of $||.||_H$. Sections 3-4 discuss particular models where identification can be achieved using a monotonicity constraint on the parameter space.

To highlight the estimation challenge, suppose T admits a Frechet derivative at θ_0 say $\mathcal{T}$. That is there exists a linear bounded operator $\mathcal{T} : H[0,1] \to L^2[0,1]$ satisfying for all $\theta \in \mathcal{H}$

$$T(\theta) = T(\theta_0) + \mathcal{T}(\theta - \theta_0) + o(||\theta - \theta_0||_H).$$

Write $S(x,w) \equiv f_{XW}(x,w)\tau(w)$. Under mild smoothness conditions given in Section 3 this derivative exists and has the form

$$(\mathcal{T}\xi)(w) = \int_{[0,1]} \xi(t)S[\theta_0(t), w]dt.$$

[5]That is, identification is met over a restricted space of functions 'close enough' to θ_0 in the sense of the Hilbert metric. What constitutes 'close enough' is determined by the magnitude of $L = sup_{x,w}|\partial^2 F(x|w)/\partial x|$ assuming the latter exists, see Chen, Chernozhukov, Lee and Newey (2011), Proposition 5. It is not clear how this result can be implemented since L is unknown.

Introduce $\mathcal{T}^* : L^2[0,1] \to H[0,1]$ the adjoint of $\mathcal{T}$ given by

$$(\mathcal{T}^*\xi)(t) = \mathcal{D}^{-1}\{\int_{[0,1]} \xi(w)S[\theta_0(t), w]dw\}.$$

where $\mathcal{D} : \{u \in H^2[0,1] : \nabla u(0) = \nabla u(1) = 0\} \to L^2[0,1]$ is defined by $\mathcal{D}u \equiv u - \nabla^2 u$. Under standard integrability conditions given in Section 3, $\mathcal{T}^*\mathcal{T}$ is compact, and therefore admits a spectral system (Kress 1999, Theorem 15.16) say $\{\lambda_j^2, \phi_j, j \geq 1\}$ such that

$$(\mathcal{T}^*\mathcal{T})\phi_j = \lambda_j^2\phi_j, j \geq 1.$$

Rearrange the sequence according to $\lambda_1^2 \geq \lambda_2^2 \geq ...$, repeating an eigenvalue by its order of multiplicity if needed. Technically, the ill-posed phenomenon described in Section 1 arises because $\lambda_j \to 0$ as $j \to \infty$ rendering the mapping $(\mathcal{T}^*\mathcal{T})^{-1}$ discontinuous. In that case, forming the sample analog of (6) does not yield consistency since the estimation loss depends on the mapping in question applied on the data estimation error. The Tikhonov method consists of modifying $(\mathcal{T}^*\mathcal{T})^{-1}$ into a new operator which is continuous.

To describe the construction of the estimator, write $I(E) = 1$ if event E is true and $I(E) = 0$ otherwise. Let K denote a kernel function satisfying notably

$$\int K(t)dt = 1, \int t^u K(t)dt = 0 \text{ for } u = 1, ..., r-1 \ \int t^r K(t)dt \neq 0,$$

for some positive integer $r > 1$. The estimator of T is given by

$$(\hat{T}\theta)(w) \equiv \int_{[0,1]} \hat{\Omega}[\theta(t), w]dt$$

with,

$$\hat{\Omega}[x, w] \equiv \frac{1}{nh}\sum_{i=1}^{n} \tau(W_i)I(X_i \leq x)K(\frac{W_i - w}{h}),$$

where h is a deterministic strictly positive sequence satisfying $h \to 0$ as $n \to \infty$. The estimator of Q is given by

$$\hat{Q}(w) \equiv \frac{1}{nh}\sum_{i=1}^{n} \tau(W_i)(1 - Y_i)K(\frac{W_i - w}{h}).$$

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

$$Min_{\theta \in \mathcal{H}} \int_{[0,1]} |(\hat{T}\theta)(w) - \hat{Q}(w)|^2 dw + a_n \{ \int_{[0,1]} |\theta(t)|^2 dt + \int_{[0,1]} |\nabla \theta(t)|^2 dt \}$$

where a_n is a strictly positive deterministic sequence of real numbers meeting $a_n \to 0$ as $n \to \infty$.

To appreciate the presence of the penalty term in the objective, one may write heuristically from the first-order conditions $\hat{\theta} - \theta_0 \approx (\mathcal{T}^*\mathcal{T} + a_n)^{-1}(\Delta)$ where Δ contains the sources of error from estimating T and Q. Hence, adding a penalty term regularizes the solution since $(\mathcal{T}^*\mathcal{T} + a_n)^{-1}$ is now bounded.

3 Assumptions

In the ensuing part of the paper, r denotes a positive integer with $r > 1$.

Assumption 1*: $\{Y_i, X_i, W_i\}_{i=1}^n$ is an i.i.d. sequence of observations from (Y, X, W) where $(g(X), W) \in [0,1]^2$, and $E[Y - g(X)|W] = 0$ almost surely.*

Assumption 2*: g is strictly increasing and continuous (as an extension) on the support of X which is a compact interval. Furthermore, θ_0 belongs to $\mathcal{H}$ as defined in section 2.*

Assumption 3: *$\tau(w)$ is a continuous function on the real line with compact support $I \subseteq [0,1]$. The distribution of (X, W) is absolutely continuous with respect to Lebesgue measure, and (Y, W) has density $J(y, w)$ with respect to some sigma-finite product measure $\mu \times \ell$ where ℓ denotes Lebesgue measure. There is a real constant $C < \infty$ such that, $f(w)\tau(w)$, and $M(w)\tau(w)$ are r-times continuously differentiable with all derivatives bounded in absolute value by C. Furthermore, $\Omega(x, w)$ has two continuous partial derivatives with respect to its first argument and r continuous partial derivatives with respect to its second argument. Moreover, $\int |\tau(w)y|^2 J(y, w) d\mu(y)$, and the derivatives of Ω are bounded in absolute value by C.*

Comments: Assumption 1 may be relaxed when $g(X)$ is only bounded from below[6]. If $g(X)$ is fully supported, the inverse is increasing on the real line which requires notably adjusting the parameter space to insure square integrability is well-defined, and strengthening the regularity conditions[7]. Assumption 2 makes the domain a subspace of the Sobolev space $H[0,1]$ which is an Hilbert space equipped with inner product $<,>_1$. This choice meets the spectral theory laid out in Bissantz, Hohage and Munk (2004) which requires identification over a closed convex subspace of an Hilbert space. Furthermore, using a Sobolev space as opposed to $L^2[0,1]$ is judicious in that it permits, under certain conditions, analyzing identification using known results as discussed subsequently. Assumption 2 requires mild smoothness conditions over the parameter of interest. Notably, this can accommodate the case where θ_0 is not differentiable everywhere. Assumption 2 presumes some knowledge about the norm of θ_0 and its derivative. In practice, this is not too restrictive since one can choose M arbitrary large. Assumption 3 contains notably standard smoothness conditions which allow to estimate the operator with the kernel method. The theoretical motivation for using a weight function τ is to remedy to technical difficulties due to the "edge-effect" which takes place if the density function of W is discontinuous on the boundaries of its support. An alternative if the derivatives are not continuous is to use boundary kernels provided in Hall and Horowitz (2005).

***Assumption 4**: For any θ in $\mathcal{H}$, $||T\theta - T\theta_0|| = 0 \Rightarrow ||\theta - \theta_0||_H = 0$.*

***Assumption 5**: K is supported on $[-1,1]$, symmetrical around the origin, and continuously differentiable. Furthermore,*

$$\int K(t)dt = 1, \int t^u K(t)dt = 0 \textbf{ for } u = 1, ..., r-1 \textbf{ and } \int t^r K(t)dt \neq 0.$$

***Assumption 6**:* (Writing $\delta_n \equiv h^{2r} + 1/nh$). *(a) a_n is a strictly positive deterministic sequence of real numbers satisfying lim $a_n = 0$ as $n \to \infty$. (b) lim $\delta_n / a_n = 0$ as $n \to \infty$.*

[6]This stems from the fact that $\int_{(0,\infty)} P[g(X) > t]dt = \int_{(0,1)} P[g(X) > \chi(t)]\nabla\chi(t)dt$ for a suitable increasing function χ from $(0,1)$ into $(0,\infty)$ meeting the smoothness conditions of Fremlin (2010) Theorem 263D. The methodology is identical if one re-define $\Omega[\theta(t), w]$ with $\{\tau(w)f(w) - \Omega[\theta(t), w]\}\nabla\chi(t)$ and Q with $\tau(w)M(w)$.

[7]In this case, modify T replacing (3) by $E[g(X)|W = w] = \int 1 - F[\theta_0(t)|w] - F[\theta_0(-t)|w]dt$ assuming $E|g(X)|$ exists. Then, work on a weighted Sobolev space $\mathcal{H} = \{\theta : \mathbb{R} \to \mathbb{R} : ||\theta||^2_{\pi,H} \leq M\}$ where $||\theta||^2_{\pi,H} \equiv \int \pi(t)|\theta(t)|^2 dt + \int \pi(t)|\nabla\theta(t)|^2 dt$, and $\pi > 0$ is a weight function with vanishing tails.

Comments: Assumption 4 assumes global identification. There are some models where identification can be achieved by restricting the parameter space to contain increasing functions assuming that the distribution of (X, W) is complete wrt X (Severini and Tripathi 2006) as proven in Lemma ID. Completeness imposes a certain form of functional correlation which will be met, for instance, with the exponential family of distributions (see Newey Powell 2003, Ai and Chen 2003). General class of nonparametric complete distributions are furnished in Andrews (2011). Assumption 5 is met with smooth compactly supported kernels. Using a kernel with a non-compact support is also possible provided some mild integrability conditions are satisfied. Assumption 6 demands the researcher to let collapse the regularization sequence at a rate less rapid than the MSE rate achieved by the nonparametric estimator of T. This is common with the ill-posed literature because a too rapid decay of the regularization sequence would impose dealing with the ill-posed problem again.

Theorem 1 (Convergence)

Let $\hat{\theta} \equiv Argmin_{\theta \in \mathcal{H}} ||\hat{T}\theta - \widehat{Q}||^2 + a_n ||\theta||_H^2$. *Under assumptions 1 through 6,*

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

Furthermore,

$$lim\ Esup\{|\hat{\theta}(t) - \theta_0(t)|^2 : t \in [0,1]\} = 0\ as\ n \to \infty.$$

Comments: Theorem 1 implies lim $E||\hat{\theta} - \theta||^2 = 0$. In practice, the estimator will be computed using a finite-dimensional space dense for $\mathcal{H}$ in the sense of the metric $||.||_1$. A truncated series using some known orthonormal basis of $H[0,1]$ such as Chebyshev's polynomials (Gagliardini and Scaillet 2012) is the easiest choice. 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 a Tikhonov estimator.

Under the Assumptions of Theorem 1, the Frechet derivative of T at θ_0 exists and is given as in Section 2. Furthermore,

$$\int_{[0,1]^2} |S[\theta_0(t), w]|^2 dtdw < \infty.$$

This last integrability makes $\mathcal{T}^*\mathcal{T}$ compact which renders the estimation problem ill-posed since the spectrum of eigenvalues $\{\lambda_j^2\}_{j=1}^{\infty}$ contains 0 as a limit point. Under these circumstances, one cannot derive a rate of convergence unless further smoothness conditions are imposed. These are given next. For any Ψ belonging to $\mathcal{H}$ introduce

$$(DT_\Psi)(\xi)(w) \equiv \int_{[0,1]} \xi(t) S[\Psi(t), w] dt.$$

Assumption 7*: (a) There is a finite real constant $L > 0$ such that for any $(\Psi, \xi) \in \mathcal{H}^2$,*

$$||T(\xi) - T(\Psi) - DT_\Psi(\xi - \Psi)|| \leq L||\xi - \Psi||^2.$$

(b) $\mathcal{T}$ is non singular3[8]*. .*

Comments: Write $B \equiv sup_{x,w}|\partial S(x,w)/\partial x|$. This latter quantity exists under Assumption 3. It is not too difficult to show that Assumption 7(a) is met for any $L > B/2$. However, the existence of B is sufficient but not necessary for Assumption 7(a) to hold. Assumption 7(b) assumes that T_1 is injective which prevents the degenerate case where 0 is an eigenvalue. This extends the full rank condition on the Jacobian necessary for conducting the asymptotic analysis with the GMM method. This injectivity condition fails if X and W are independent[9].

Write ϕ_j the eigenfunction of $\mathcal{T}^*\mathcal{T}$ with associated eigenvalue λ_j^2 for $j \geq 1$. Under Assumption 7, $\{\phi_j\}_{j=1}^{\infty}$ forms a complete orthonormal basis of $H[0,1]$. Hence, θ_0 has the Fourier representation

$$\theta_0 = \sum_{j=1}^{\infty} b_j \phi_j,$$

where

$$b_j \equiv \int_{[0,1]} \theta_0(t)\phi_j(t)dt + \int_{[0,1]} \nabla\theta_0(t)\nabla\phi_j(t)dt, j \geq 1,$$

and

[8]This latter is met if $||\int_{[0,1]} \xi(x)S[\theta_0(x), w]dx|| = 0$ implies $||\xi|| = 0$.

[9]If this is the case, $\mathcal{T}$ maps any function ξ into $< \xi, f_0 > \tau(w)f(w)$ where f_0 is the marginal density of X evaluated at $\theta_0(x)$. Under these circumstances, one can create a function $||\xi|| \neq 0$ meeting $||\mathcal{T}\xi|| = 0$ violating Assumption 7(b). For instance, pick $\xi = \varphi - [< \varphi, f_0 > / < \mu, f_0 >]$ where $\mu(x) = 1[0 < x < 1]$ and $||\varphi|| \neq 0$. Since $||\varphi||_H^2 \geq ||\varphi||^2$ assumption 7(b) is violated.

$$\sum_{j=1}^{\infty} |b_j|^2 < \infty.$$

***Assumption 8**:*

$$\sqrt{\sum_{j=1}^{\infty} \frac{|b_j|^2}{|\lambda_j|^2}} < 1/3L.$$

Furthermore, there are a real constant $\sigma > \lambda_1^2$ and a known real-valued function φ such that $\varphi(0) = 0$, φ is continuous increasing on $[0, \sigma)$, and

$$\sum_{j=1}^{\infty} \frac{|b_j|^2}{\varphi(|\lambda_j|^2)^2} < \infty.$$

Moreover, $\varphi(t)/\sqrt{t}$ is non-decreasing and there is a real constant $C_o > 0$ such that on $(0, \sigma]$,

$$C_o t \leq \varphi(t)\Xi(t),$$

where $\Xi(t) \equiv inf_{\{t \leq \lambda \leq \sigma\}}\{\lambda/\varphi(\lambda)\}$.

***Assumption 9**:* (Write $\Lambda(t) \equiv \varphi(t)\sqrt{t}$) $h \propto n^{-1/(2r+1)}$, *and* $a_n \asymp \Lambda^{-1}(\sqrt{\delta_n})$.

Comments: Assumption 8 is a source condition which demands eigenfunctions having sufficient 'predictive power' to approximate θ_0. Because the regularization process, in effect, estimates an approximation of θ_0, it introduces a deterministic bias. A prerequisite for the bias to vanish at a tractable rate is that the Fourier coefficients exhibit a sufficiently rapid rate of decline in the the sense of the first part of Assumption 8. The more severe the rate of decay of the eigenvalues (i.e. the more "ill-posed" the problem), the less complex θ_0 must be in the sense of having $\theta_0 \simeq \sum_{j=1}^{J} b_j \phi_j$ an accurate approximation for some small positive integer J. Also the smoother T (i.e. the smaller L in Assumption 7), the more complex θ_0 is allowed to be. This highlights the fact that when T is very smooth, the bound $1/3L$ is not likely to be binding. However, if this is not the case then a sharp decline in Fourier coefficients is required. The second part of Assumption 8 introduces extra smoothness conditions which accelerate the rate of collapse of the regularization bias. This is true provided the regularization sequence is chosen according

to Assumption 9. Hence, taking advantage of this extra smoothness is only possible with an a priori-knowledge concerning φ. The last part of Assumption 8 places weak restrictions over the rate of decay of the Fourier coefficients. Assumption 8 is very general because it imposes only 'shape' restrictions over the link function φ. As far as I know, there is no result available to handle the 'rough case' when $\varphi(t)/\sqrt{t}$ is decreasing[10]. Assumption 9 delivers the rate-optimal regularization sequence which balances the regularization bias of size $O(\varphi(a_n))$ and the estimation error incurred on the regularized version having size $O(\sqrt{\delta_n/a_n})$. Selecting $h \propto n^{-1/(2r+1)}$ yields the optimal MSE rate of estimation for the operator, namely $\delta_n \propto n^{-2r/(2r+1)}$. To assess the plausibility of Assumption 8, consider the polynomial link $\varphi(t) = t^u$ for $u \geq 1/2$. Special cases of polynomial source conditions have been examined in Hall-Horowitz (2005) and Horowitz-Lee (2009). Then, Assumption 8 is met for instance if $||(T_1^* T_1)^{-u}\theta_0|| < 1/3L$ for some $u \in [1/2, 1]$ (the first part of Assumption 8 corresponds to $u = 1/2$).

Theorem 2 (Rate of Convergence)

Under the assumptions of Theorem 1 and assumptions 7 through 9,

$$E||\hat{\theta} - \theta_0||_H^2 = O\{\aleph(n^{-r/(2r+1)})\},$$

where $\aleph(t) \equiv \{\varphi[\Lambda^{-1}(t)]\}^2$. *Furthermore,*

$$Esup\{|\hat{\theta}(t) - \theta_0(t)|^2 : t \in [0,1]\} = O\{\aleph(n^{-r/(2r+1)})\},$$

Comments: Theorem 2 implies $E||\hat{\theta} - \theta_0||^2 = O\{\aleph(n^{-r/(2r+1)})\}$. As established in Bissantz, Hohage and Munk (2004), selecting $a_n \asymp \sqrt{\delta_n}$ yields the MSE rate $\sqrt{\delta_n}$ which is no longer optimal whenever $\varphi(t)/\sqrt{t}$ is strictly increasing. To derive a rate of convergence using Theorem 2, one must choose the link function φ. Consider for instance the polynomial link case previously discussed. For the polynomial case, Assumption 9 is met with $a_n \asymp \delta_n^{1/(2u+1)}$. This yields a MSE rate $\delta_n^{2u/(2u+1)}$. It follows from Theorem 2 that $||\hat{\theta} - \theta_0|| = O_p(n^{-2ur/(2r+1)(2u+1)})$. Under the minimal smoothness conditions for the functions of Assumption 3, the rate is $n^{-4u/5(2u+1)}$. Thus, under the minimal assumptions of this paper, the rate for the polynomial case ranges from $n^{-1/5}$ to $n^{-4/15}$ depending on the link smoothness. However, if the functions of Assumption 3 admit enough derivatives, a researcher can construct an estimator converging at a rate close to $n^{-1/4}$ if $u = 1/2$ and close to $n^{-1/3}$ if $u = 1$.

[10]Some results are provided in Tautenhahn and Jin (2003) in the context of polynomial source conditions (i.e. if $0 < u < 1/2$) where T is known suggesting that only further very restrictive smoothness conditions on the Frechet derivative can overcome the 'rough case'.

4 Multivariate extension

This section treats the more general model

$$E[Y - g(X, \tilde{Z})|\tilde{W}] = 0 \tag{7}$$

almost surely, where X is a scalar variable, $\tilde{Z} = (Z, V)$ is 1 by $p = p_z + p_v$ vector, and $\tilde{W} = (V, W)$ is 1 by $p_v + l$ vector with $l \geq c \equiv p_z + 1$. It is assumed that $(g(X, Z, V), Z, V, W)$ has support $[0, 1]^{1+p+l}$ without loss of generality[11]. The unknown function $g(x, \tilde{z})$ is strictly monotonic with respect to its first argument for all $\tilde{z}$ in $[0, 1]^p$. For the sake of clarity, g is taken to be increasing. The goal is to estimate, for each fixed v, $\theta_v(t, z) \equiv g^{-1}(t, z, v)$, which gives the inverse of g as a function of x for any given z. Write $||\theta||^2_{H,s} \equiv \sum_{|\alpha| \leq s} ||\nabla^{(\alpha)}\theta||^2_{[0,1]^c}$, and assume θ_v, for each fixed v, belongs to $\mathcal{H} = \{\theta \in H^s[0, 1]^c : ||\theta||^2_{H,s} \leq M\}$ for some integer $s > 0$ where $M < \infty$ is a known real constant.

Let $\{Y_i, X_i, Z_i, V_i, W_i\}_{i=1}^n$ denote some i.i.d. sequence of observations. It is assumed that the distribution of (X, Z, V, W) is absolutely continuous with respect to Lebesgue measure. Furthermore, (Y, V, W) is assumed to have a density $J(y, v, w)$ with respect to some sigma-finite product measure $\mu \times \ell$ where ℓ denotes the Lebesgue measure in $[0, 1]^{p_v+l}$. Write f_{XZVW} as the density of (X, Z, V, W). Let $f(.|v, w)$ denote the density of $Z|V = v, W = w$, and f_{VW} the density of (V, W). Write $F[.|z, v, w]$ as the **CDF** of $X|Z = z, V = v, W = w$.

To describe this multivariate extension, observe that (7) can be re-written with iterated expectations as

$$\int_{[0,1]^{p_z}} E[g(X, Z, V)|Z = z, V = v, W = w]f(z|v, w)dz = E[Y|V = v, W = w]. \tag{8}$$

Furthermore, the Fubini's Theorem yields

$$E[g(X, Z, V)|Z = z, V = v, W = w] = \int_0^1 P[g(X, Z, V) > t|Z = z, V = v, W = w]dt = 1 - \int_{[0,1]} F[g^{-1}(t, z, v)|z, v, w]dt. \tag{9}$$

[11] For example 2 discussed in Section 1, $g(X, Z, V)$ will satisfy this condition using the transformed model having $Y = 1/2$ and $g = H[M^{-1}]$ where $H(t) \equiv (t + 1)/2$.

Let τ denote a weight function with compact support $I \subseteq [0,1]^l$. The weight function is introduced for similar reasons as those explained in Section 3. Now define

$$\Omega[x,z,v,w] \equiv \int_{-\infty}^{x} f_{XZVW}(u,z,v,w)\tau(w)du, \tag{10}$$

and

$$M(v,w) \equiv \int yJ(y,v,w)d\mu(y). \tag{11}$$

Suppose $\Omega[x,z,v,w]$ and $Q_v(w) \equiv \tau(w)\{f_{VW}(v,w) - M(v,w)\}$ exist for each fixed v in $[0,1]^{p_v}$. It follows from (7)-(11) that $\theta_v(t,z)$ solves the following integral equation

$$(T_v\theta)(w) = Q_v(w) \tag{12}$$

where,

$$(T_v\theta)(w) \equiv \int_{[0,1]^c} \Omega[\theta(t,z),z,v,w]dtdz.$$

Identification of a unique root for (12) over $\mathcal{H}$ will be achieved if (X,Z) and W are sufficiently correlated in the sense of Assumption 4bis.

Under standard integrability conditions given subsequently, the ill-posed problem described in section 3 is still present requiring the usage of the Tikhonov method when solving the analog of (12).

To describe the construction of the estimator, let r denote some integer with $r > 1$ and introduce $\mathcal{K}_r$=$\{f : \mathbb{R} \longrightarrow \mathbb{R}$, f continuous, f supported on $[-1,1]$, f symmetrical, $\int f(t)dt = 1$, $\int t^u f(t)dt = 0$ for $u = 1,...,r-1$ $\int t^r f(t)dt \neq 0\}$. Furthermore for any integer $p \geq 1$ define $\mathcal{K}_{p,r}$ =$\{\kappa : \mathbb{R}^p \to \mathbb{R}$, $\kappa(t_1,...,t_p) \equiv \prod_{j=1}^{p} k(t_j)$, $k \in \mathcal{K}_r$ $\}$.

The estimator of T_v is given by

$$(\hat{T}_v\theta)(w) \equiv \int_{[0,1]^c} \hat{\Omega}[\theta(t,z),z,v,w]dtdz$$

with,

$$\hat{\Omega}[x,z,v,w] \equiv \frac{1}{nh_z^{p_z}h_v^{p_v}h_w^l}\sum_{i=1}^{n}\tau(W_i)I(X_i \leq x)K_z(\frac{Z_i-z}{h_z})K_v(\frac{V_i-v}{h_v})K_w(\frac{W_i-w}{h_w}),$$

where K_z belongs to $\mathcal{K}_{p,r}$, K_v belongs to $\mathcal{K}_{p_v,r}$, K_w belongs to $\mathcal{K}_{l,r}$ while h_z, h_v, and h_w are deterministic strictly positive sequences satisfying $Max(h_z,h_v,h_w) \to 0$ as $n \to \infty$. The estimator of Q_v is given by

$$\hat{Q}_v(w) \equiv \frac{1}{nh_v^{p_v}h_w^l}\sum_{i=1}^{n}\tau(W_i)(1-Y_i)K_v(\frac{V_i-v}{h_v})K(\frac{W_i-w}{h_w}).$$

The Tikhonov estimator $\hat{\theta}_v(t,z)$ solves the following optimization problem

$$Min_{\theta\in\mathcal{H}}\int_{[0,1]^l}|(\hat{T}_v\theta)(w)-\hat{Q}_v(w)|^2dw + a_n\sum_{|\alpha|\leq s}\int_{[0,1]^c}|\nabla^{(\alpha)}\theta(t,z)|^2dtdz,$$

where a_n is a strictly positive deterministic sequence of real numbers meeting $a_n \to 0$ as $n \to \infty$. The regularity conditions for this multivariate setting are given next.

5 Assumptions

***Assumption 1 bis**: $\{Y_i,X_i,Z_i,V_i,W_i\}_{i=1}^n$ is an i.i.d. sequence of observations from (Y,X,Z,V,W). Furthermore, $(g(X,Z,V),Z,V,W) \in [0,1]^{1+p_z+p_v+l}$ and $l \geq c \equiv p_z+1$. Moreover, $E[Y-g(X,Z,V)|V=v,W]=0$ almost surely for each v in $[0,1]^{p_v}$.*

***Assumption 2 bis**: For all (z,v) in $[0,1]^p$, $g(x,z,v)$ is strictly increasing and continuous (as an extension) on the support of X which is a compact interval. Furthermore, $\theta_v(t,z)$ belongs to $\mathcal{H}$ as defined in section 4 for each v in $[0,1]^{p_v}$.*

***Assumption 3 bis**: $\tau(w)$ is a continuous function on $\mathbb{R}^l$ with compact support $I \subseteq [0,1]^l$. The distribution of (X,Z,V,W) is absolutely continuous with respect to Lebesgue measure, and (Y,V,W) has density $J(y,v,w)$ with respect to some sigma-finite product measure $\mu \times \ell$ where ℓ denotes Lebesgue measure in $[0,1]^{p_v+l}$. There is a real constant $C<\infty$ such that, all partial derivatives of order r of $f_{VW}(v,w)\tau(w)$, and $M(v,w)\tau(w)$ are continuous and bounded in absolute value by C. Furthermore, $\Omega[x,z,v,w]$ has two continuous partial derivatives with respect to its first argument and r continuous partial derivatives with respect to its last three arguments. Moreover, $\int|\tau(w)y|^2J(y,v,w)d\mu(y)$, and the derivatives of Ω are bounded in absolute value by C.*

Assumption 4 bis: *For any θ in $\mathcal{H}$, $||T_v\theta_v - T_v\theta||_{[0,1]^l} = 0 \Rightarrow ||\theta_v - \theta||_{H,s} = 0$ for each v in $[0,1]^{p_v}$.*

Assumption 5 bis: *K_z belongs to $\mathcal{K}_{p_z,r}$, K_v belongs to $\mathcal{K}_{p_v,r}$, and K_w belongs to $\mathcal{K}_{l,r}$.*

Assumption 6 bis: (Writing $\delta_n \equiv Max(h_z, h_v, h_w)^{2r} + 1/nh_z^{p_z}h_v^{p_v}h_w^l$). *(a) a_n is a strictly positive deterministic sequence of real numbers satisfying lim $a_n = 0$ as $n \to \infty$. (b) lim $\delta_n/a_n = 0$ as $n \to \infty$.*

Comments: Assumption 4 bis will be met for certain models if the distribution (X, Z, W) is complete wrt (X, Z) conditional on $V = v$ (see Lemma ID2). Consider for example a variant of example 2 with $X = M(V, C+U)$ where $E[U|V, W] = 0$, (X, V, C, W) is observable, and $M(v, .)$ is strictly increasing and continuous in its second argument for all v in the support of V. Clearly, this model is more restrictive than that described in example 2 because V must be exogenous. Since $E[M^{-1}(V, X)|V, W] = E[C|V, W]$, identification of $M(v, .)$ for each fixed v is achieved provided the distribution of (X, W) is complete with respect to X conditional on $V = v$.

Theorem 1bis

Let $\hat{\theta}_v \equiv Argmin_{\theta\in\mathcal{H}}||\hat{T}_v\theta - \hat{Q}_v||^2_{[0,1]^l} + a_n||\theta||^2_{H,s}$. Under assumptions 1bis through 6bis, and for each v in $[0,1]^{p_v}$

$$lim\ E||\hat{\theta}_v - \theta_v||^2_{H,s} = 0\ as\ n \to \infty,$$

Furthermore,

(i) If $2s > c$ then

$$lim\ Esup\{|\hat{\theta}_v(t,z) - \theta_v(t,z)|^2 : (t,z) \in [0,1]^c\} = 0\ as\ n \to \infty,$$

(ii) If $2s \leq c$ then

$$lim\ E\int_{[0,1]^c} |\hat{\theta}_v(t,z) - \theta_v(t,z)|^2 dtdz = 0\ as\ n \to \infty.$$

Comments: Theorem 1 bis implies $||\hat{\theta}_v - \theta_v||_{[0,1]^c} = o_p(1)$ regardless of the smoothness $s \geq 1$. However, (i) shows that achieving strong consistency in the multivariate setting requires a certain number of weak derivatives to overcome the dimensionality of (X, Z). The contract between (i) and (ii) arises because the current level of

mathematical knowledge only permits to assert that the Sobolev norm is stronger than the uniform norm when $2s > c$. The univariate case analyzed in Section 3 does not have this smoothness restriction since $s = 1$ insures the embedding in question.

Write $S(x, z, v, w) \equiv f_{XZVW}(x, z, v, w)\tau(w)$. Under Assumption 3bis, $\mathcal{T}_v : H^s[0,1]^c \to L^2[0,1]^l$ the Frechet derivative of T_v at θ_v exists for each v in $[0,1]^{p_v}$, and has the form

$$(\mathcal{T}_v\xi)(w) = \int_{[0,1]^c} \xi(t,z)S[\theta_v(t,z), z, v, w]dtdz.$$

Write $H_* = \{u \in H^{2s}[0,1]^c : \nabla^\alpha u(x) = \nabla^\alpha u(x) = 0, x \in \partial[0,1]^c, |\alpha| = 2j+1 < 2s, j = 0, ..., s-1\}$ and introduce $\mathcal{T}_v^* : L^2[0,1]^l \to H^s[0,1]^c$ the adjoint of $\mathcal{T}_v$ given by

$$(\mathcal{T}_v^*\xi)(t,z) = \mathcal{C}\mathcal{D}^{-1}\{\int_{[0,1]^l} \xi(w)S[\theta_v(t,z), z, v, w]dw\},$$

where $\mathcal{D} : H_* \to L^2[0,1]^c$ is defined by

$$Du \equiv \sum_{|\alpha| \le s} (-1)^{|\alpha|} \prod_{i=1}^{c} \nabla_i^{2\alpha_i} u,$$

and $\mathcal{C} : H_* \to H^{2s}[0,1]^c$ is defined as the operator such that for any $(u, \theta) \in H_* \times H^s[0,1]^c$

$$\int_{[0,1]^c} (\mathcal{D}u)(t,z)\theta(t,z)dtdz = \sum_{|\alpha| \le s} \int_{[0,1]^c} |\nabla^{(\alpha)}(\mathcal{C}u)(t,z)\nabla^{(\alpha)}\theta(t,z)|^2 dtdz,$$

Under Assumption 3bis $\mathcal{T}_v$ is compact owing to the fact that

$$\int_{[0,1]^{c+l}} |S[\theta_v(t,z), z, v, w]|^2 dtdzdw < \infty.$$

Consequently, the spectrum of eigenvalues $\{\lambda_{j,v}^2\}_{j=1}^\infty$ characterizing $\mathcal{T}_v^*\mathcal{T}_v$ contains 0 as a limit point generating an ill-posed problem similar in nature to that described in Section 3. To derive a rate of convergence, one may introduce additional smoothness conditions. These, which mirror those of Section 3, are given next. For any Ψ belonging to $\mathcal{H}$ introduce

$$(DT_{v,\Psi})(\xi)(w) = \int_{[0,1]^c} \xi(t,z) S[\Psi(t,z), z, v, w] dt dz.$$

Assumption 7 bis*: (a) For each v in $[0,1]^{p_v}$, there is a finite real constant $L_v > 0$ such that for all $(\Psi, \xi) \in \mathcal{H}^2$,*

$$||T_v(\xi) - T_v(\Psi) - DT_{v,\Psi}(\xi - \Psi)||_{[0,1]^l} \leq L_v ||\xi - \Psi||^2_{[0,1]^c}.$$

(b) For each v in $[0,1]^{p_v}$, $\mathcal{T}_v$ is non singular[12]*. .*

Comments: Assumption 7bis(b)is an injectivity condition which demands sufficient 'dependency' between (X, Z) and W as explained in Section 3 conditional on V. Assumption 7bis ensures the existence of a non degenerate spectral system $\{\lambda^2_{j,v}, \phi_{j,v}\}_{j=1}^{\infty}$ for $\mathcal{T}_v^* \mathcal{T}_v$ from which regularization can be performed to derive a rate of convergence for the estimator of θ_v. Furthermore, $\{\phi_{j,v}\}_{j=1}^{\infty}$ forms a complete othornormal basis of $H[0,1]^c$. Hence, θ_v has the Fourier representation

$$\theta_v = \sum_{j=1}^{\infty} b_{j,v} \phi_{j,v}$$

where

$$b_{j,v} \equiv \sum_{|\alpha| \leq s} \int_{[0,1]^c} \nabla^{(\alpha)} \theta_v(t,z) . \nabla^{(\alpha)} \phi_{j,v}(t,z) dt dz, j \geq 1,$$

and

$$\sum_{j=1}^{\infty} |b_{j,v}|^2 < \infty.$$

Assumption 8 bis*:* For each v in $[0,1]^{p_v}$,

$$\sqrt{\sum_{j=1}^{\infty} \frac{|b_{j,v}|^2}{|\lambda_{j,v}|^2}} < 1/3L_v,$$

and

[12]This latter is met if $||\int_{[0,1]^c} \xi(x,z) S[\theta_v(x,z), z, w] dx|| = 0$ implies $||\xi|| = 0$.

$$\sum_{j=1}^{\infty} \frac{|b_{j,v}|^2}{\varphi(|\lambda_{j,v}|^2)^2} < \infty,$$

for some function φ meeting Assumption 8 which does not depend on v.

***Assumption 9 bis**: $h_z \propto n^{-1/(2r+d)}$, $h_v \propto n^{-1/(2r+d)}$,and $h_w \propto n^{-1/(2r+d)}$ where $d \equiv l + p$. Furthermore, $a_n \asymp \Lambda^{-1}(\sqrt{\delta_n})$ where Λ is as defined in Assumption 9.*

Comments: To appreciate Assumption 8 bis introduce

$$B \equiv sup_{x,z,v,w}|\partial S(x,z,v,w)/\partial x|.$$

This latter quantity exists under Assumption 3 bis. It follows that Assumption 8bis is met for any $L_v \geq B/2$. Hence, the same trade-off discussed in Section 3 arises between the smoothness of $S(x,z,v,w)$ and the speed of a decline of the Fourier coefficients. Selecting the bandwidths according to Assumption 9bis yields the optimal MSE rate of estimation on the operator T_v, namely $\delta_n \propto n^{-2r/(2r+d)}$. These bandwidths do not coincide with the classic optimal choice for estimating Q_v with the kernel method. The reason is that the size of the estimation error produced for estimating Ω dominates over that for estimating of Q_v. Thus, there is no benefit asymptotically in using optimal bandwidths for estimating Q_v[13].

Theorem 2 bis

Under the assumptions of Theorem 1bis and assumptions 7bis through 9bis, for each v in $[0,1]^{p_v}$,

$$E||\hat{\theta}_v - \theta_v||^2_{H,s} = O\{\aleph(n^{-r/(2r+d)})\}$$

Furthermore,

(i) if $2s > c$ then

$$Esup\{|\hat{\theta}_v(t,z) - \theta_v(t,z)|^2 : (t,z) \in [0,1]^c\} = O\{\aleph(n^{-r/(2r+d)})\}$$

[13]This is true because of assumption 3 bis which supposes that Ω and M are 'as smooth'. If $\Omega(x,z,v,w)$ admits more derivatives wrt (z,v,w) say $r > r_{vw}$ where r_{vw} denotes the number of derivatives for $M(v,w)$ then there may be a benefit in employing optimal bandwidths for estimating Q_v, namely using instead $h'_v \propto n^{-1/(2r+p_v+l)}$,and $h'_w \propto n^{-1/(2r+p_v+l)}$.

(ii) if $2s \leq c$ *then*

$$E \int_{[0,1]^c} |\hat{\theta}_v(t,z) - \theta_v(t,z)|^2 dtdz = O\{\aleph(n^{-r/(2r+d)})\}$$

Comments: Theorem 2 bis implies $||\hat{\theta}_v - \theta_v||_{[0,1]^c} = O_p\{\sqrt{\aleph(n^{-r/(2r+d)})}\}$. To get some sense about the speed of convergence consider the polynomial link case discussed in Section 3. This yields a MSE rate $\delta_n^{2u/(2u+1)}$. This and Theorem 2 bis show that, under the polynomial link case, the optimal rate of estimation for the multivariate setting is $O_p(n^{-2ru/(2r+d)(2u+1)})$. Thus, $\hat{\theta}$ inherits the same 'curse of dimensionality' as that affecting a nonparametric density estimator with a speed of convergence decelerating rapidly as the number of variables in (Z, V, W) augments. For the endogenous regression model, this means that the speed of convergence is slower as the number of endogenous variables augments since $d \geq 2p_z + 1$ is necessary for identification.

6 Monte Carlo experiments

This section examines the finite sample properties of the estimator of g^{-1} in the context of the model

$$Y = g(X) + \epsilon,\ E[\epsilon|W] = 0$$

where $g(X) = \sqrt{X}$, $\epsilon \sim N(0,1)$, and (X, W) is supported on $[0,1]^2$ with a density function

$$f_{XW}(x,w) \propto \sum_{j=1}^{\infty} (-1)^{j+1} j^{-1} sin(j\pi x) sin(j\pi w).$$

This density, used in Hall and Horowitz (2005), is convenient because it renders the task of deriving the eigensystem of T_1 less cumbersome. In this experiment, T and Q are estimated with

$$K(t) = (15/16) * (1 - t^2)^2 I[|t| \leq 1],$$

which is a kernel of order $r = 2$. It is beyond the scope of this paper to offer optimal bandwidths selection criteria. In this experiment, the bandwidths are selected according to the Silverman's rule of thumb (Silverman 1986),

Table 1: Results of Monte Carlo experiments.

	$n = 200$	$n = 400$
	h——$0.8h$	h——$0.8h$
$C_a = 0.1$	0.197—0.219	0.146—0.171
$C_a = 0.2$	0.182—0.206	0.135—0.152
$C_a = 0.3$	0.174—0.193	0.120—0.154
$C_a = 0.4$	0.218—0.235	0.139—0.183
$C_a = 0.5$	0.236—0.275	0.152—0.234

namely $h = \sigma_w n^{-1/(2r+1)}$ where σ_w denote the empirical standard deviation of W. The regularization sequence is chosen according to $a_n = C_a n^{-r/(2r+1)}$, where C_a is a positive constant.

In this experiment, the weight function is $\tau(w) = I[0 \leq w \leq 1]$. The integrals are computed numerically. The estimator is retrieved with the simulated annealing method using a budget of 250 iterations. Due to the long computational time 200 replications are carried out. The simulations are conducted in Gauss. Table 1 shows the results for $E[||\hat{\theta} - \theta_0||^2]$.

Conclusion

This paper has presented a new estimator for the inverse a monotonic function g satisfying the condition moment restriction $E[Y - g(X)|W] = 0$ where W is excluded from X. The MSE consistency of this estimator has been established and its rate of convergence has been derived under generic source conditions. Furthermore, the multivariate extension was presented offering notably a new method for the endogenous non-additive regression model. There are three questions arising from this article. First, the problem of identification. As discussed in Section 3, the monotonicity of g suggests that one can invoke the concepts of completeness for certain models only. Secondly, more research needs to be pursued concerning the optimal choice for the regularization sequences. Lastly, the choice of the link function defining the source conditions. The rate-optimality result of Theorem 2 presumes that a researcher has some a-priori knowledge over the link function which is often not the case in applied works. As established in Bissantz, Hohage and Munk (2004) selecting $a_n \asymp \sqrt{\delta_n}$ yields the (non-optimal) MSE rate $\sqrt{\delta_n}$. Presumingly, one can use this non-optimal estimator to recover the Fourier coefficients and eigenvalues from data in order to recover a realistic estimate for the link function. This latter may be used to compute a feasible version optimal estimator. Examining the Monte-Carlo properties of this feasible version may suggest some adaptive theory.

References

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

Andrews. D. W. K., 2011. Examples of L2-Complete and Boundedly-Complete Distributions. *Cowles Foundation Discussion Paper No. 1801.*.

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., 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 ISSN 1862-2763.*

Chernozhukov V. , Gagliardini.P. and Scaillet O., 2012. Nonparametric Instrumental Variable Estimation of Structural Quantile Effects. *Econometrica, Vol.80 1533-1562.*.

Chernozhukov V., Imbens.G. and Newey W., 2007. Instrumental variable estimation of nonseparable models. *Journal of Econometrics, Vol.139 4-14.*.

Darolles S. Florens J.P. and Renault.E., 2006. Non parametric instrumental regression. *Working paper, GREMAQ, University of Social Science, Toulouse, France.*

Dette. H., Neumeyer.N., and Pliz K.F, 2005. A note on nonparametric estimation of the effective dose in quantal bioassay. *J.Amer.Statist.Assoc., Vol.100, 503-510.*.

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. Kunisch K. and Neubauer, 1996. Regularization of Inverse Problems *Kluwer Academic Publishers, Dordrecht ,Paperback edition, 2000).*

Engl. H.W and Kugler. P., 2005. Nonlinear Inverse Problems: Theoretical Aspects and Some Industrial Applications. *Multidisciplinary Methods for Analysis Optimization and Control of Complex Systems Mathematics in Industry, Volume 6, Part I, 3-47.* .

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

Horowitz. J., 2007. Asymptotic Normality of a Nonparametric Instrumental Variables Estimator. *International Economic Review, Vol 48,1329-1349..*

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. *John Wiley and Sons, New York..*

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

Lu. S., Pereverzev. S and Ramlau. R., 2007. An analysis of Tikhonov regularization for nonlinear ill-posed problems under a general smoothness assumption. *Inverse Problems, 23, 217-230..*

Newey. W. and Powell. J., 2003. Instrumental variable Estimation of non parametric models. *Econometrica, Vol 71, No 5..*

Rudin. W., 2003. Principles of Mathematical Analysis, 3rd Edition. *McGraw-Hill., Cambridge..*

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

Tang. R., Banerjee M., and Michailidis. G., 2011. A two-stage hybrid procedure for estimating an inverse regression function. *Annals of statistics, Vol 39, No 2.*

Tautenhahn. U., and Jin. Q., 2003. Tikhonov regularization and a posteriori rules for solving nonlinear ill-posed problems. *Inverse Problems, 19, 1-21.*

Appendix

This section provides the proofs. We shall write $A\theta \equiv T\theta - Q$ and $\hat{A}\theta \equiv \hat{T}\theta - \hat{Q}$. For any f and g belonging to $H^s[0,1]$ write $< f, g >_s = \sum_{k=0}^{s} < \nabla^{(k)} f, \nabla^{(k)} g >$ and $||f||_s \equiv \sqrt{< f, f >_s}$. Starting from the proof of Theorem 1, I shall abuse notation using $||.||$ and $<,>$ in lieu of $||.||_1$ and $<,>_1$ for functions belonging to $H[0,1]$.

Lemma 1: Under the assumptions of Proposition 1,

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

(b) $E||\hat{Q} - Q||^2 = O(\delta_n)$.

proof:

(a) $|\hat{T}\theta(w) - T\theta(w)| \leq \int_{[0,1]} |\hat{H}(\theta(x), w) - H(\theta(x), w)|dx$ and a Cauchy-Schwartz's inequality offers

$$||\hat{T}\theta - T\theta||^2 \leq \int_{[0,1]} \int_{[0,1]} |\hat{H}(\theta(x), w) - H(\theta(x), w)|^2 dxdw.$$

Now consider

$$E|\hat{H}(\theta(x), w) - H(\theta(x), w)|^2 = \{E\hat{H}(\theta(x), w) - H(\theta(x), w)\}^2 + Var\hat{H}(\theta(x), w).$$

Under Assumption 3, one can use a change of variable along with a Taylor's expansion to establish

$$E\hat{H}(t, w) - H(t, w) = O(h^r) \text{ uniformly.}$$

Another change of variable yields

$$Var\hat{H}(t, w) = O(1/nh) \text{ uniformly.}$$

Thus, $E|\hat{H}(t, w) - H(t, w)|^2 = O(\delta_n)$ uniformly. This shows the claim.

(b) One can write $Q(w) = \{r(w) - v(w)\}$ and $\hat{Q}(w) = \hat{r}(w) - \hat{v}(w)$. The Proof follows identically as in (a) using the fact that

$$E|\hat{r}(w) - r(w)|^2 = O(\delta_n) \text{ and } E|\hat{v}(w) - v(w)|^2 = O(\delta_n) \text{ uniformly.}$$

Theorem 1:

Under Assumption 2, $\mathcal{H}$ is a closed convex subset of the Sobolev's space $H[0,1]$ which is an Hilbert space equipped with inner product $<,>_1$. Under Assumptions 1-6, $\hat{T}$ is weakly continuous. Since $\mathcal{H}$ is closed and convex $\hat{\theta}$ exists by Bissantz, Hohage and Munk (2004). By definition,

$$||\hat{A\theta}||^2 + a_n||\hat{\theta}||^2 \leq ||\hat{A\theta}_0||^2 + a_n||\theta_0||^2.$$

Also

$$E||\hat{A\theta}_0||^2 = E||\hat{A\theta}_0 - A\theta_0||^2.$$

This establishes $E||\hat{A\theta}_0||^2 = O(\delta_n)$ by Lemma 1.

Consequently

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

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

for some finite real constant M which exists by Assumption 2.

Now a triangular inequality yields

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

with

$$E||A\hat{\theta} - \hat{A\theta}||^2 = O(\delta_n) \text{ by lemma 1, and}$$

$$E||\hat{A\theta}||^2 = O(\delta_n) + O(a_n) \text{ by (1).}$$

It follows from these that

$$E||A\hat{\theta}||^2 = O(\delta_n) + O(a_n) \text{ and } E||\hat{\theta}||^2 \leq O(\delta_n/a_n) + ||\theta_0||^2.$$

This establishes

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

Because T is weakly sequentially closed by Assumptions 1-3, the above result along with Assumption 4 implies $\lim E||\hat{\theta} - \theta_0||^2 = 0$ by Theorem 2 of Bissantz, Hohage and Munk (2004). The proof follows because $||.||_{sup} < M||.||_1$ holds for some constant M on $H[0,1]$ by the Sobolev embedding Theorem (Adams and Fournier 2013). QED.

Theorem 2:

The proof is based upon the argument of Theorem 2.1-2.2 of Lu, Pereverzev and Ramlau (2007). I shall write $D \equiv \mathcal{T}$, and $\psi(t) \equiv \varphi(t)t^{-1/2}$. Under Assumptions 3-7, Theorem 2.42-2.39 of Florens and Carrasco (2007) and the Sobolev embedding theorem, D is a compact non-singular operator on $H[0,1] \subset L^2[0,1]$ Hence, Kress (1999) Theorem 15.16 implies the existence of a spectral system $\{\lambda_j, \phi_j, \psi_j, j \geq 1\}$ with $\lambda_j \neq 0$ for $j = 1, 2, ...$ and

$$D\phi_j, = \lambda_j \psi_j \text{ and } D^*\psi_j = \lambda_j \phi_j,\ j \geq 1.$$

Furthermore, $\{\phi_j, j \geq 1\}$ and $\{\psi_j, j \geq 1\}$ form a complete orthonormal basis of $H[0,1]$ and $\{\psi_j, j \geq 1\}$ form a complete orthonormal basis of $L^2[0,1]$ by Assumption 7b.

Part I: Write $\theta_n \equiv Argmin_{\mathcal{H}}||A\theta||^2 + a_n||\theta||^2$ and $f \equiv \theta_0 - a_n(D^*D + a_n)^{-1}\theta_0$. Note that $\theta_0 = \varphi(D^*D)\varpi$ for some $||\varpi|| < \infty$ by Assumption 8. Furthermore, $v \equiv (DD^*)^{-1}D\theta_0$ meets $||v|| < 1/3L$.

Introduce

$$R(f) \equiv Af - D(f - \theta_0), \text{ and}$$

$$R(\theta_n) \equiv A\theta_n - D(\theta_n - \theta_0).$$

Under Assumption 7, there is a real constant $L < \infty$ such that

$$||R(f)|| \leq L||f - \theta_0||^2, \text{ and}$$

$$||R(\theta_n)|| \leq L||\theta_n - \theta_0||^2.$$

By definition

$$||A\theta_n||^2 + a_n||\theta_n||^2 \leq ||Af||^2 + a_n||f||^2. \tag{3}$$

Note that

$$||\theta_n||^2 = ||\theta_n - \theta_0||^2 + 2 < \theta_n - \theta_0, \theta_0 > + ||\theta_0||^2, \text{ and}$$

$$||f||^2 = ||f - \theta_0||^2 + 2 < f - \theta_0, \theta_0 > + ||\theta_0||^2.$$

These used into (3) give

$$a_n||\theta_n - \theta_0||^2 \leq B_1 + B_2 + B_3 + B_4 + B_5 \tag{4}$$

where

$$B_1 = 2a_n < R(\theta_n), v > \leq 2a_n L||v||.||\theta_n - \theta_0||^2$$

$$B_2 = 2 < D(f - \theta_0), R(f) > \leq 2L||D(f - \theta_0)||.||f - \theta_0||^2$$

$$B_3 = ||R(f)||^2 \leq L^2||f - \theta_0||^4$$

$$B_4 = a_n||f - \theta_0||^2$$

$$B_5 = ||D(f - \theta_0) + a_n v||^2$$

Moreover,

$$||D(f - \theta_0)|| = a_n||D(D^*D + a_n)^{-1}D^*v|| \leq a_n||v||, \text{ and}$$

$$||D(f - \theta_0) + a_n v|| = a_n^2||(D^*D + a_n)^{-1}\psi(D^*D)\varpi||,$$

with the last equality arising because $D(f - \theta_0) + a_n v = (DD^* + a_n)^{-1}v$ with $||(DD^* + a_n)^{-1}v||^2 = ||(D^*D + a_n)^{-1}\psi(D^*D)\varpi||^2$.

These and (4) imply

$$||\theta_n - \theta_0||^2 \leq \{(1 - 2L||v||)\}^{-1}\{(1 + 2L||v||).||f - \theta_0||^2 + a_n^{-1}L^2||f - \theta_0||^4 + a_n^3||(D^*D + a_n)^{-1}\psi(D^*D)\varpi||^2\}.$$

Now we shall bound each norm in the right-hand side of the above inequality.

$$||f - \theta_0|| = a_n||(D^*D + a_n)^{-1}\varphi(D^*D)\varpi|| \leq a_n||\varpi||sup_{j\geq 1}|\varphi(\lambda_j^2)/(a_n + \lambda_j^2)|.$$

This and Assumption 8 imply

$$||f - \theta_0|| = O(\varphi(a_n)) \text{ (i)}$$

Furthermore

$$a_n^3||(D^*D + a_n)^{-1}\psi(D^*D)\varpi||^2 \leq a_n||\varpi||^2 sup_{j\geq 1}|a_n\psi(\lambda_j^2)/(a_n + \lambda_j^2)|^2.$$

This and Assumptions 8-9 imply

$$a_n^3||(D^*D + a_n)^{-1}\psi(D^*D)\varpi||^2 = O(a_n|\psi(a_n)|^2) = O(|\varphi(a_n)|^2) \text{ (ii)}$$

Using (i)-(ii) along with Assumptions 8-9 yield

$$||\theta_n - \theta_0||^2 = O(|\varphi(a_n)|^2) + O(|\varphi(a_n)|^4/a_n) = O(|\varphi(a_n)|^2).$$

Part II. By definition

$$||\hat{A}\hat{\theta}||^2 + a_n||\hat{\theta}||^2 \leq ||\hat{A}\theta_n||^2 + a_n||\theta_n||^2 \tag{5}$$

Adding $2 < \hat{T}\hat{\theta} - \hat{Q}, Q - \hat{T}\theta_n > +||\hat{T}\theta_n - Q||^2 + a_n\{||\theta_n||^2 - 2 < \hat{\theta}, \theta_n >\}$ on both sides in (5) yields

$||\hat{T}\hat{\theta} - \hat{Q} + Q - \hat{T}\theta_n||^2 + a_n||\hat{\theta} - \theta_n||^2$

$\leq$

$||\hat{T}\theta_n - \hat{Q}||^2 + ||\hat{T}\theta_n - Q||^2 + 2 < \hat{T}\hat{\theta} - \hat{Q}, Q - \hat{T}\theta_n > +2a_n < \theta_n, \theta_n - \hat{\theta} > .$

Thus

$a_n||\hat{\theta} - \theta_n||^2$

$\leq ||\hat{Q} - Q||^2 + 2 < T\theta_n - Q, T\theta_n - T\hat{\theta} > +2a_n < \theta_n, \theta_n - \hat{\theta} > +2 < \hat{T}\theta_n - T\theta_n, \hat{T}\theta_n - \hat{T}\hat{\theta} - T\theta_n + T\hat{\theta} > +2 < \hat{T}\theta_n - T\theta_n, T\theta_n - T\hat{\theta} >$
$+2 < T\theta_n - Q, \hat{T}\theta_n - \hat{T}\hat{\theta} - T\theta_n + T\hat{\theta} >.$

Now use the Euler's condition $[DT_{\theta_n}]^*A(\theta_n) + a_n\theta_n = 0$ and Assumption 7 resulting in

$a_n||\hat{\theta} - \theta_n||^2 \leq ||\hat{Q} - Q||^2 + 2 < T\theta_n - Q, T\theta_n + DT_{\theta_n}(\hat{\theta} - \theta_n) - T\hat{\theta} > +2 < \hat{T}\theta_n - T\theta_n, \hat{T}\theta_n - \hat{T}\hat{\theta} - T\theta_n + T\hat{\theta} > +2 < \hat{T}\theta_n - T\theta_n, T\theta_n - T\hat{\theta} >$
$+2 < T\theta_n - Q, \hat{T}\theta_n - \hat{T}\hat{\theta} - T\theta_n + T\hat{\theta} >.$

The Cauchy-Schwartz's inequality offers

$$a_n||\hat{\theta} - \theta_n||^2 \leq ||\hat{Q} - Q||^2 + 2L||T\theta_n - Q||.||\hat{\theta} - \theta_n||^2 + 2U_1 + 2U_2 + 2U_3$$

where

$$U_1 = ||\hat{T}\theta_n - T\theta_n||.||\hat{T}\theta_n - \hat{T}\hat{\theta} - T\theta_n + T\hat{\theta}||$$

$$U_2 = ||\hat{T}\theta_n - T\theta_n||.||T\theta_n - T\hat{\theta}||$$

$$U_3 = ||T\theta_n - Q||.||\hat{T}\theta_n - \hat{T}\hat{\theta} - T\theta_n + T\hat{\theta}||.$$

Furthermore, $||v|| < 1/3L$ by Assumption 8. This and Theorem 2.3 of Tautenhahn and Jin (2003) yield

$$||T\theta_n - Q|| \leq a_n||v||.$$

Consequently

$$\{1 - 2L||v||\}a_n E||\hat{\theta} - \theta_n||^2 \leq E||\hat{Q} - Q||^2 + 2\{EU_1 + EU_2 + EU_3\}$$

Now from Lemma 1

$$E||\hat{Q} - Q||^2 = O(\delta_n).$$

Furthermore, using the Cauchy-Schwartz's inequality along with Lemma 1 yields

$EU_1 \leq \{E[||\hat{T}\theta_n - T\theta_n||^2]E[||\hat{T}\theta_n - \hat{T}\hat{\theta} - T\theta_n + T\hat{\theta}||^2]\}^{1/2} = O(\delta_n)$

$EU_2 \leq \{E[||\hat{T}\theta_n - T\theta_n||^2]E[||T\theta_n - T\hat{\theta}||^2]\}^{1/2} = O(a_n\delta_n^{1/2})$

$EU_3 \leq \{E[||T\theta_n - Q||^2]E[||\hat{T}\theta_n - \hat{T}\hat{\theta} - T\theta_n + T\hat{\theta}||^2]\}^{1/2} = O(a_n\delta_n^{1/2})$

This proves $E||\hat{\theta} - \theta_n||^2 = O(\delta_n/a_n)$ by Assumptions 8-9.

Part III. Consider the inequality

$$E||\hat{\theta} - \theta_0||^2 \leq 2\{E||\hat{\theta} - \theta_n||^2 + ||\theta_n - \theta_0||^2\}.$$

This and combining the results of parts I-II proves the claim.

Theorem 1bis: Under the assumptions of Theorem 1bis

$$sup_{t,z,v,w} E|\hat{\Omega}[t,z,v,w] - \Omega[t,z,v,w]|^2 = O[max(h_z, h_v, h_w)^{2r} + 1/nh_z^{p_z} h_v^{p_v} h_w^l],$$

Furthermore, one can write $Q_v(w) = r(v,w) - m(v,w)$ and $\hat{Q}_v(w) = \hat{r}(v,w) - \hat{m}(v,w)$ where

$$sup_{v,w} E|\hat{r}(v,w) - r(v,w)|^2 = O[Max(h_v, h_w)^{2r}] + 1/nh_v^{p_v} h_w^l, \text{ and}$$

$$sup_{v,w} E|\hat{m}(v,w) - m(v,w)|^2 = O[Max(h_v, h_w)^{2r} + 1/nh_v^{p_v} h_w^l].$$

It follows directly that

$$E \int_{[0,1]^l} |(\hat{T}_v \varphi)(w) - (T_v \varphi)(w)|^2 dw = O[Max(h_z, h_v, h_w)^{2r} + 1/nh_z^{p_z} h_v^{p_v} h_w^l],$$

uniformly over $\mathcal{H}$, and

$$E \int_{[0,1]^l} |\hat{Q}_v(w) - Q_v(w)|^2 dw = O[Max(h_v, h_w)^{2r} + 1/nh_v^{p_v} h_w^l].$$

Write $A_v \equiv T_v - Q_v$ and $\hat{A}_v \equiv \hat{T}_v - \hat{Q}_v$. These last two results yield

$$E \int_{[0,1]^l} |(\hat{A}_v \varphi)(w) - (A_v \varphi)(w)|^2 dw = O[max(h_z, h_v, h_w)^{2r} + 1/nh_z^{p_z} h_v^{p_v} h_w^l],$$

uniformly over $\mathcal{H}$. The proof follows from that of Theorem 1 because $||.||_{sup} < M||.||_1$ for some constant M whenever $2s > c$ by the Sobolev embedding Theorem (Adams and Fournier 2013).

Theorem 2bis: The proof is analogous as that of Theorem 2 and is therefore omitted.

Lemma ID 1: Assume X has support $[a,b]$ and $\mathcal{H} = \{\theta \in H[0,1] : \theta(0) = a, \theta(1) = b, \nabla\theta \geq \mu, ||\theta||^2 + ||\nabla\theta||^2 \leq M\}$ where $M < \infty$ and $\mu > 0$ are known real constants. Furthermore, assume every non-constant random variable $U(X)$ with $E|U(X)|^2 < \infty$ is correlated with some random variable $R(W)$ with $E|R(W)|^2 < \infty$. Under Assumptions 1-3, Assumption 4 is met.

Proof: $||T\theta - Q|| = 0$ implies

$$\int_0^1 P[X > \theta(t)|W = w]dt = \int_0^1 P[X > \theta_0(t)|W = w]dt.$$

By assumption 2, the weak derivatives of θ coincides with its derivative whenever the latter exist. Hence, there is a function $\theta_* = \theta$ a.e which is absolutely continuous. Consequently, for $0 \leq t_1 < t_2 \leq 1$

$$\theta_*(t_1) - \theta_*(t_2) = \int_{[t_1,t_2]} \nabla\theta(t)dt \geq \mu(t_2 - t_1) > 0$$

Thus, θ_* has an inverse on $[a,b]$. The Fubini's Theorem yields

$$E[\theta_*^{-1}(X)|W] = E[\theta_0^{-1}(X)|W] \text{ almost surely.}$$

Now under the assumptions of Lemma ID 1, (X,W) is complete wrt X which insures $\theta_*^{-1}(x) = \theta_0^{-1}(x)$ up to a null set of $[a,b]$. But θ_*^{-1} and θ_0^{-1} are continuous on $[a,b]$ which implies that the null set is empty by Rudin (2003), p 6.2. It follows that $\theta_*^{-1}(x) = \theta_0^{-1}(x)$ everywhere on $[a,b]$ which shows $\theta_*(t) = \theta_0(t)$ everywhere on $[0,1]$. Thus, $\theta = \theta_0$ a.e. and $\nabla\theta = \nabla\theta_0$ a.e.

Q.E.D

Lemma ID 2: Assume X has support $[a,b]$ and assume θ_v belongs to $\mathcal{H} = \{\theta \in H^s[0,1]^c : \theta(0,z) = a, \theta(1,z) = b, \partial\theta/\partial t \geq \mu, \sum_{|\alpha|\leq s} ||\nabla^{(\alpha)}\theta||^2_{[0,1]^c} \leq M\}$ for some integer $s > 0$ where $\mu > 0$ and $M < \infty$ are known real constants. Furthermore, assume for each v in $[0,1]^{p_v}$, every non-constant random variable $U(X,Z)$ with $E|U(X,Z)|^2 < \infty$ is correlated conditional on $V = v$ with some random variable $R(W)$ with $E|R(W)|^2 < \infty$. Under Assumptions 1bis-3bis, Assumption 4bis is met.

Proof: For each v in $[0,1]^{p_v}$, $||T\theta - T\theta_v|| = 0$ on $\mathcal{H}$ implies

$$\int_{[0,1]^{p_z}} \int_0^1 P[X > \theta(t,z)|Z = z, V = v, W = w] f(z|v,w) dt dz = \int_{[0,1]^{p_z}} \int_0^1 P[X > \theta_v(t,z)|W = w, V = v, Z = z] f(z|v,w) dt dz.$$

By assumption 2bis, the weak derivative of θ with respect to its first argument coincides with its derivative whenever the latter exist. Hence, there is a function $\theta_* = \theta$ a.e $[0,1] \times [0,1]^{p_z}$ which is absolutely continuous with respect its first argument. It follows by Lemma I that θ_* has an inverse on $[a,b]$ wrt first argument. Now taking the inverses of both $\theta_*(t,z)$ and $\theta_v(t,z)$, and using the Fubini's Theorem yields

$$\int_{[0,1]^{p_z}} \int_0^1 E[\theta_*^{-1}(X,Z)|W = w, V = v, Z = z] f(z|v,w) dt dz = \int_{[0,1]^{p_z}} \int_0^1 E[\theta_v^{-1}(X,Z)|W = w, V = v, Z = z] f(z|v,w) dt dz.$$

Therefore

$$E[\theta_*^{-1}(X,Z)|V = v, W] = E[\theta_v^{-1}(X,Z)|V = v, W] \text{ almost surely.}$$

This establishes $\theta_*^{-1}(x,z) = \theta_v^{-1}(x,z)$ up to a null set of $[a,b] \times [0,1]^{p_z}$ due to the fact that $(X,Z,W)|V = v$ is complete wrt (X,Z) under the assumptions of Lemma ID 2. But θ_*^{-1} and θ_v^{-1} are continuous on $[a,b]$ which insures that the null set of $[a,b]$ is empty by Rudin (2003), p 6.2. It follows that $\theta_*^{-1}(x,z) = \theta_v^{-1}(x,z)$ everywhere on $[a,b]$ and for almost every z in $[0,1]^{p_z}$. This proves $\theta_*(t,z) = \theta_v(t,z)$ everywhere on $[0,1]$ for almost every z in $[0,1]^{p_z}$. Thus, $\nabla^{(\alpha)}\theta_v = \nabla^{(\alpha)}\theta$ a.e. for $|\alpha| \leq s$.

QED.