Nonparametric Instrumental Regression[1]

Serge Darolles[2] Yanqin Fan[3] Jean-Pierre Florens[4]
Eric Renault[5]

January 21, 2010

[1]We first want to thank our coauthors on papers strongly related with this one: M. Carrasco, C. Gouriéroux, J. Johannes, J. Heckman, C. Meghir, S. Van Bellegem, A. Vanhems and E. Vytlacil. We also acknowledge helpful comments from the editor, the four referees and D. Bosq, X. Chen, L. Hansen, P. Lavergne, J.M. Loubes, W. Newey and J.M. Rolin. We thank the participants to conferences and seminars in Chicago, Harvard-MIT, London, Louvain-la-Neuve, Montréal, Paris, Princeton, Santiago, Seattle, Stanford, Stony Brook and Toulouse. We also thank R. Lestringand who performed the numerical illustration given in Section 5.

[2]Lyxor Asset Management and CREST.
[3]Vanderbilt University.
[4]Toulouse School of Economics.
[5]UNC at Chapel Hill, CIRANO and CIREQ.

Abstract

The focus of the paper is the nonparametric estimation of an instrumental regression function φ defined by conditional moment restrictions stemming from a structural econometric model: $E[Y - \varphi(Z) \mid W] = 0$, and involving endogenous variables Y and Z and instruments W. The function φ is the solution of an ill-posed inverse problem and we propose an estimation procedure based on Tikhonov regularization. The paper analyses identification and overidentification of this model and presents asymptotic properties of the estimated nonparametric instrumental regression function.

Keywords: Instrumental Variables, Integral Equation, Ill-posed Problem, Tikhonov Regularization, Kernel Smoothing.

Classification JEL: C14, C30.

Résumé

Nous nous intéressons à l'estimation nonparamétrique d'une fonction de régression instrumentale φ. Cette fonction est définie à l'aide de conditions de moment provenant d'un modèle économétrique structurel de la forme $E[Y - \varphi(Z) \mid W] = 0$, où les Y et Z sont des variables endogènes et les W des instruments. La fonction φ est alors la solution d'un problème inverse mal posé, et nous proposons une procédure d'estimation utilisant la régularisation de Tikhonov. Le papier analyse l'identification et la suridentification du modèle et donne les propriétés asymptotiques de l'estimateur de la régression instrumentale non paramétrique.

Mots clés: Variables instrumentales, Equation intégrale, problème mal posé, Régularisation de Tikhonov, Lissage par noyau.

Classification JEL : C14, C30.

1 Introduction

An economic relationship between a response variable Y and a vector Z of explanatory variables is often represented by an equation:

$$Y = \varphi(Z) + U, \tag{1.1}$$

where the function φ should define the relationship of interest while U is an error term[1]. The relationship (1.1) does not characterize the function φ if the residual term is not constrained. This difficulty is solved if it is assumed that $E[U \mid Z] = 0$, or equivalently $\varphi(Z) = E[Y \mid Z]$. However, in numerous structural econometric models, the conditional expectation function is not the parameter of interest. The structural parameter is a relation between Y and Z, where some of the Z components are endogenous. This is for example the case in various situations: simultaneous equations, error-in-variables models, treatment models with endogenous selection, ...

This paper considers an instrumental variables treatment of the endogeneity. The introduction of instruments may be done in several ways. Our framework is based on the introduction of a vector W of instruments such that φ is defined as the solution of:

$$E[U \mid W] = E[Y - \varphi(Z) \mid W] = 0. \tag{1.2}$$

Instrumental variables estimation may be also introduced using control functions (for a systematic treatment see Newey, Powell and Vella (1999)) or local instrumental variables (see e.g. Florens, Heckman, Meghir and Vytlacil (2008)).

Equation (1.2) characterizes φ as the solution of a Fredholm integral equation of the first kind and this inverse problem is known to be ill-posed and needs a regularization method. The connection between instrumental variables estimations and ill-posed inverse problems has been pointed out by Florens (2000) who proposed to address this question using a Tikhonov regularization approach, also used in Carrasco and Florens (2000) to treat GMM estimation with an infinite number of moment conditions. The Tikhonov approach has also been adopted by Hall and Horowitz (2005), while Newey and Powell (2003) have resorted a different analysis based on sieve estimation under regularization by compactness.

The literature on ill-posed inverse problems is huge, in particular in numerical analysis and image processing. The deconvolution problem is one of the main uses of inverse problems in statistics (see Carrasco and Florens (2009)). The main features of the instrumental variables estimation are coming from the necessity of the estimation of the equation itself (and not

[1] We remain true to the tradition in Econometrics of additive error terms. See e.g. Florens (2005), Horowitz and Lee (2007), Imbens and Newey (2009) for alternative structural approaches.

only the right hand side) and from the combination between parametric and nonparametric rates of convergence. The theory of inverse problems introduces in Econometrics a different albeit related class of concepts of regularity of functions. Source conditions extend standard differentiability assumptions used for example in kernel smoothing. Even if the present paper is self contained, we refer to Carrasco, Florens and Renault (2007) for a general discussion on inverse problem in Econometrics.

This paper is organized as follows. In Section 2 the instrumental regression problem (1.2) is precisely defined and the identification of φ is discussed. Section 3 discusses the ill-posedness and presents regularization methods and regularity spaces. The estimator is defined in Section 4 and consistency and rate of convergence are analyzed. Section 5 briefly considers practical questions about the implementation of our estimator and displays some simulations. Some extensions are suggested in the conclusion section. Two appendices collect proofs: Appendix A contains the proofs of the theorems and Appendix B shows that our set of assumptions may be derived from more primitive conditions on the DGP.

Throughout the rest of this paper, all the limits are taken as the sample size N goes to infinity, unless otherwise stated. We will use $f_A(\cdot)$, $f_{A,B}(\cdot,\cdot)$ to denote the density function of the random variable A and the joint density function of the random variables A, B. In addition, we will use $f_{A|B}(\cdot|b)$ and $f_{A|B,C}(\cdot|b,c)$ to denote the conditional density functions of A given $B=b$ and $B=b, C=c$ respectively. For two numbers α, β, we let $\alpha \wedge \beta = \min(\alpha, \beta)$.

2 The instrumental regression and its identification

2.1 Definition

We denote by $S=(Y,Z,W)$ a random vector partitioned into $Y \in \mathbf{R}$, $Z \in \mathbf{R}^p$ and $W \in \mathbf{R}^q$. The probability distribution on S is characterized by its joint cumulative distribution function (*cdf*) F. We assume that the first coordinate of S, Y is square integrable. This condition is actually a condition on F and $\mathcal{F}$ denotes the set of all *cdf*s satisfying this integrability condition. For a given F, we consider the Hilbert space L^2_F of square integrable functions of S and we denote by $L^2_F(Y)$, $L^2_F(Z)$, $L^2_F(W)$ the subspaces of L^2_F of real valued functions depending on Y, Z or W only. We denote by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ the norm and scalar product in these spaces. Typically F is the true distribution function from which the observations are generated and these L^2_F spaces are related to this distribution.

In this section no additional restriction is maintained on the functional spaces but more conditions are necessary, in particular for the analysis of

the asymptotic properties. These restrictions will only be introduced when necessary.

Definition 2.1: *We call instrumental regression any function* $\varphi \in L_F^2(Z)$ *which satisfies the condition:*

$$Y = \varphi(Z) + U, \qquad E[U \mid W] = 0. \tag{2.1}$$

Equivalently φ corresponds to any solution of the following functional equation:

$$E[Y - \varphi(Z) \mid W] = 0. \tag{2.2}$$

If Z and W are identical, φ is equal to the conditional expectation of Y given Z, and then it is uniquely defined. In the general case, additional conditions are required in order to identify uniquely φ by (2.1) or (2.2).

Example 2.1: *We assume that* $S \sim N(\mu, \Sigma)$ *and we restrict our attention to linear instrumental functions* φ, $\varphi(z) = Az + b$. *Conditions* (2.1) *are satisfied if and only if* $A\Sigma_{ZW} = \Sigma_{YW}$, *where* $\Sigma_{ZW} = cov(Z, W)$ *and* $\Sigma_{YW} = cov(Y, W)$. *If* Z *and* W *have the same dimension and if* Σ_{ZW} *is non singular, then* $A = \Sigma_{YW}\Sigma_{ZW}^{-1}$ *and* $b = \mu_Y - A\mu_Z$. *We will see later that this linear solution is the unique solution of* (2.2) *in the normal case. If* Z *and* W *do not have the same dimension, more conditions are needed for existence and uniqueness of* φ.

It will be useful to introduce the two following notations:

i) $T : L_F^2(Z) \to L_F^2(W) \quad \varphi \to T\varphi = E[\varphi(Z) \mid W]$,

ii) $T^* : L_F^2(W) \to L_F^2(Z) \quad \psi \to T^*\psi = E[\psi(W) \mid Z]$.

These two linear operators satisfy:

$$\begin{aligned} \langle \varphi(Z), \psi(W) \rangle &= E[\varphi(Z)\psi(W)] = \langle T\varphi(W), \psi(W) \rangle \\ &= \langle \varphi(Z), T^*\psi(Z) \rangle, \end{aligned}$$

and then T^* is the adjoint (or dual) operator of T, and reciprocally. Using these notations, φ corresponds to any solution of the functional equation:

$$A(\varphi, F) = T\varphi - r = 0, \tag{2.3}$$

where $r(W) = E[Y \mid W]$. This implicit definition of the parameter of interest φ as a solution of an equation depending on the data generating process is the main characteristic of the structural approach in econometrics. In our case note that equation (2.3) is linear in φ.

If the joint *cdf* F is characterized by its density $f(y,z,w)$ w.r.t. the Lebesgue measure, equation (2.3) is an *integral Fredholm type I equation*:

$$\int \varphi(z) \frac{f_{Z,W}(z,w)}{f_W(w)} dz = r(w), \tag{2.4}$$

where $r(w) = \int y \frac{f_{Y,W}(y,w)}{f_W(w)} dy$.

The estimation of a function by solving an integral equation is a usual problem in nonparametric statistics. The simpler issue of nonparametric estimation of a density function is actually an ill-posed inverse problem. From the empirical counterpart of the cumulative distribution function, we have a root$-n$ consistent estimator of the integral of the density function on any interval of the real line. It is precisely the necessary regularization of the ill-posed characterization of the density function, which leads to nonparametric rates of convergence for density estimation (see e.g. Hardle and Linton (1994) and Vapnik (1998)).

The inverse problem (2.4) is an even more difficult issue since its inputs for statistical estimation of φ are nonparametric estimators of the functions $f_{Z,W}$, f_W, and r, which also involve nonparametric speeds of convergence. However, a contribution of this paper will be to show that the dimension of W has no negative impact on the resulting speed of convergence of the estimator of φ. Roughly speaking, increasing the dimension of W increases the speed of convergence. The usual dimensionality curse in nonparametric estimation is only dependent on the dimension of Z.

2.2 Identification

The *cdf* F and the regression function r are directly identifiable from the random vector S. Our objective is then to study the identification of the function of interest φ. The solution of equation (2.3) is unique if and only if T is one to one (or equivalently the null space $\mathcal{N}(T)$ of T is reduced to zero). This abstract condition on F can be related to a probabilistic point of view using the fact that T is a conditional expectation operator.

This concept is well-known in statistics and corresponds to the notion of a complete statistic[2] (see Lehman and Scheffe (1950), Basu (1955)). A systematic study is made in Florens and Mouchart (1986), and Florens, Mouchart and Rolin (1990) Chapter 5, under the name of strong identification (in a L^2 sense) of the σ-field generated by the random vector Z by the σ-field generated by the random vector W.

The characterization of identification in terms of "*completeness of the conditional distribution function of Z given W*" was already provided by Newey and Powell (2003). They also discussed the particular case detailed

[2] A statistic t is complete in a probability model depending on θ if $E[\lambda(t) \mid \theta] = 0 \ \forall\theta$ implies $\lambda(t) = 0$.

in Example 2.2 below. Actually, the strong identification assumption can be interpreted as a nonparametric rank condition as it is shown in the following example dealing with the normal case.

Example 2.2: *Following Example 2.1, let us consider a random normal vector (Z, W). The vector Z is strongly identifiable by W if one of the three following equivalent conditions is satisfied (see Florens, Mouchart and Rolin (1993)):*

i) $\mathcal{N}(\Sigma_{ZZ}) = \mathcal{N}(\Sigma_{WZ})$;

ii) $\mathcal{N}(\Sigma_{WZ}) \subset \mathcal{N}(\Sigma_{ZZ} - \Sigma_{ZW}\Sigma_{WW}^{-1}\Sigma_{WZ})$;

iii) $\text{Rank}(\Sigma_{ZZ}) = \text{Rank}(\Sigma_{WZ})$.

In particular, if Σ_{ZZ} is non singular, the dimension of W must be greater than or equal to the dimension of Z. If the joint distribution of (Y, Z, W) is normal and if a linear instrumental regression is uniquely defined as in Example 2.1, then it is the unique instrumental regression.

The identification condition can be checked in specific models (see e.g. Blundell, Chen and Kristensen (2007)). It is also worth interpreting it in terms of the adjoint operator T^* of T.

Proposition 2.1: *The three following conditions are equivalent:*

i) φ *is identifiable;*

ii) T^*T *is one-to-one;*

iii) $\overline{\mathcal{R}(T^*)} = L_F^2(Z)$, *where* $\overline{E}$ *is the closure of* $E \subset L_F^2(Z)$ *in the Hilbert sense and* $\mathcal{R}(T^*)$ *is the range of* T^*.

We will now introduce an assumption which is only a regularity condition when Z and W have no element in common. However, this assumption cannot be satisfied if there are some elements in common between Z and W. For an extension, see Feve and Florens (2009).

Assumption A.1: *The joint distribution of (Z, W) is dominated by the product of its marginal distributions, and its density is square integrable w.r.t. the product of margins.*

Assumption A.1 amounts to assume that T and T^* are Hilbert Schmidt operators, and is a sufficient condition of compactness of T, T^*, TT^* and T^*T (see Lancaster (1968), Darolles, Florens and Renault (1998)). Therefore, there exists a singular values decomposition, i.e. a sequence of non negative real numbers $\lambda_0 = 1 \geq \lambda_1 \geq \lambda_2 \cdots$ and two sequences of functions φ_i, $i \geq 0$, and ψ_j, $j \geq 0$, such that (see Kress (1999), 15.4):

Singular Values Decomposition (SVD)

i) φ_i, $i \geq 0$, is an orthonormal sequence of $L_F^2(Z)$ (i.e. $\langle \varphi_i, \varphi_j \rangle = \delta_{ij}$, $i, j \geq 0$, where δ_{ij} is the Kronecker symbol) and ψ_j, $j \geq 0$, is an orthonormal sequence of $L_F^2(W)$;
ii) $T\varphi_i = \lambda_i \psi_i$, $i \geq 0$;
iii) $T^*\psi_i = \lambda_i \varphi_i$, $i \geq 0$;
iv) $\varphi_0 = 1$, $\psi_0 = 1$;
v) $\langle \varphi_i, \psi_j \rangle = \lambda_i \delta_{ij}$, $i, j \geq 0$;
vi) $\forall g \in L_F^2(Z)$, $g(z) = \sum_{i=0}^{\infty} \langle g, \varphi_i \rangle \varphi_i(z) + \bar{g}(z)$, where $\bar{g} \in \mathcal{N}(T)$;
vii) $\forall h \in L_F^2(W)$, $h(w) = \sum_{i=0}^{\infty} \langle h, \psi_i \rangle \psi_i(w) + \bar{h}(w)$, where $\bar{h} \in \mathcal{N}(T^*)$.

Thus:

$$T[g(Z)](w) = E[g(Z) \mid W = w] = \sum_{i=0}^{\infty} \lambda_i < g, \varphi_i > \psi_i(w),$$

and:

$$T^*[h(W)](z) = E[h(W) \mid Z = z] = \sum_{i=0}^{\infty} \lambda_i < h, \psi_i > \varphi_i(z).$$

The strong identification assumption of Z by W can be characterized in terms of the singular values decomposition of T. Actually, since φ is identifiable if and only if T^*T is one-to-one, we have:

Corollary 2.1: *Under assumption A.1, φ is identifiable if and only if 0 is not an eigenvalue of T^*T.*

Note that the two operators T^*T and TT^* have the same non null eigenvalues λ_i^2, $i \geq 0$. But, for example, if W and Z are jointly normal, 0 is an eigenvalue of TT^* as soon as $\dim W > \dim Z$ and Σ is non singular[3]. But if Σ_{WZ} is of full-column rank, 0 is not an eigenvalue of T^*T.

The strong identification assumption corresponds to $\lambda_i > 0$ for any i. It means that there is a sufficient level of nonlinear correlation between the two sets of random variables Z and W. Then, we can directly deduce the Fourier decomposition of the inverse of T^*T from the one of T^*T by inverting the λ_is.

Note that, in these Fourier decompositions, the sequence of eigenvalues, albeit all positive, decrease fast to zero due to the Hilbert-Schmidt property. It should be stressed that the compactness (and the Hilbert Schmidt) assumption are <u>not</u> simplifying assumptions but describe a realistic framework (we can consider for instance the normal case). These assumptions formalize the decline to zero of the spectrum of the operator and make the inverse

[3] In this case $a'\Sigma_{WZ} = 0 \Longrightarrow T^*(a'W) = 0$.

problem ill-posed, and then more involved for statistical applications. Assuming that the spectrum is bounded from below may be relevant for other econometric applications, but is not a realistic assumption for the continuous nonparametric IV estimation.

We conclude this section by a result illustrating the role of the instruments in the decline of the λ_j. The following theorem shows that increasing the number of instruments increases the singular values and then the dependence between the Z and the W.

Theorem 2.1: *Let us assume that* $W = (W_1, W_2) \in R^{q_1} \times R^{q_2} (q_1 + q_2 = q)$ *and denote by* T_1 *the operator:*

$$\varphi \in L_F^2(Z) \to E[\varphi \mid W_1] \in L_F^2(W_1),$$

and T_1^* *its dual. Then* T_1 *is still an Hilbert Schmidt operator and the eigenvalues of* $T_1^* T_1, \lambda_{j,1}^2$*, satisfy:*

$$\lambda_{j,1} \leq \lambda_j,$$

where the eigenvalues are ranked as a non decreasing sequence and each eigenvalue is repeated according to its multiplicity order.

Example 2.3: *Consider the case* $(Z, W_1, W_2) \in R^3$ *endowed with a joint normal distribution with a zero mean and a variance* $\begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & 0 \\ \rho_2 & 0 & 1 \end{pmatrix}$*. The operator* T^*T *is a conditional expectation operator characterized by:*

$$Z \mid u \sim N\left[\left(\rho_1^2 + \rho_2^2\right) u, 1 - \left(\rho_1^2 + \rho_2^2\right)^2\right],$$

and its eigenvalues λ_j^2 *are* $(\rho_1^2 + \rho_2^2)^j$*. The eigenvectors of* T^*T *are the Hermite polynomials of the invariant distribution of this transition, i.e. the* $N\left(0, \frac{1-\left(\rho_1^4+\rho_2^4\right)}{1-\left(\rho_1^2+\rho_2^2\right)}\right)$*. The eigenvalues of* $T_1^* T_1$ *are* $\lambda_{j,1}^2 = \rho_1^{2j}$ *and the eigenvectors are the Hermite polynomials of the* $N(0,1)$ *distribution.*

3 Existence of the instrumental regression: an ill-posed inverse problem

The focus of our interest in this section is to characterize the solution of the IV equation (2.3):

$$T\varphi = r, \tag{3.1}$$

under the maintained identification assumption that T is one-to-one. The following result is known as the Picard theorem (see e.g. Kress (1999)):

Proposition 3.1: *r belongs to the range $\mathcal{R}(T)$ if and only if the series $\sum_{i\geq 0}\frac{1}{\lambda_i} < r, \psi_i > \varphi_i$ converges in $L^2_F(Z)$. Then $r = T\varphi$ with:*

$$\varphi = \sum_{i\geq 0}\frac{1}{\lambda_i} < r, \psi_i > \varphi_i.$$

Although Proposition 3.1 ensures the existence of the solution φ of the inverse problem (3.1), this problem is said ill-posed because a noisy measurement of r, $r+\delta\psi_i$ say (with δ arbitrarily small), will lead to a perturbed solution $\varphi + \frac{\delta}{\lambda_i}\varphi_i$ which can be infinitely far from the true solution φ since λ_i can be arbitrarily small ($\lambda_i \to 0$ as $i \to \infty$). This is actually the price to pay to be nonparametric, that is not to assume a priori that $r = E[Y \mid W]$ is in a given finite dimensional space.

While in finite dimensional case, all linear operators are continuous, the inverse of the operator T, albeit well-defined by Proposition 3.1 on the range of T, is not a continuous operator. Looking for one regularized solution is a classical way to overcome this problem of non-continuity.

A variety of regularization schemes are available in the literature[4] (see e.g Kress (1999) and Carrasco, Florens and Renault (2007) for econometric applications) but we focus in this paper on the Tikhonov regularized solution:

$$\varphi^\alpha = (\alpha I + T^*T)^{-1}T^*r = \sum_{i\geq 0}\frac{\lambda_i}{\alpha+\lambda_i^2} < r, \psi_i > \varphi_i, \tag{3.2}$$

or equivalently:

$$\varphi^\alpha = arg\ min_\varphi \left[\|r - T\varphi\|^2 + \alpha\|\varphi\|^2\right]. \tag{3.3}$$

By comparison with the exact solution of Proposition 3.1, the intuition of the regularized solution (3.2) is quite clear. The idea is to control the decay of eigenvalues λ_i (and implied explosive behavior of $\frac{1}{\lambda_i}$) by replacing $\frac{1}{\lambda_i}$ with $\frac{\lambda_i}{\alpha+\lambda_i^2}$. Equivalently, this result is obtained by adding a penalty term $\alpha\|\varphi\|^2$ to the minimization of $\|T\varphi - r\|^2$ which leads to (non continuous) generalized inverse. Then α will be chosen positive and converging to zero with a speed well tuned with respect to both the observation error on r and the convergence of λ_i. Actually, it can be shown (see Kress (1999), p. 285) that:

$$\lim_{\alpha\to 0}\|\varphi - \varphi^\alpha\| = 0.$$

[4]More generally, there is a large literature on ill-posed inverse problems (see e.g Wahba (1973), Nashed and Wahba (1974), Tikhonov and Arsenin (1977), Groetsch (1984), Kress (1999) and Engl, Hanke and Neubauer (2000)). For other econometric applications see Carrasco and Florens (2000), Florens (2000), Carrasco, Florens and Renault (2007) and references therein.

Note that the regularization bias is:

$$\begin{aligned} \varphi - \varphi^{\alpha} &= \left[I - (\alpha I + T^*T)^{-1} T^*T \right] \varphi \\ &= \alpha(\alpha I + T^*T)^{-1} \varphi. \end{aligned} \tag{3.4}$$

In order to control the speed of convergence to zero of the regularization bias $\varphi - \varphi^{\alpha}$, it is worth restricting the space of possible values of the solution φ. This is the reason why we introduce the spaces Φ_{β}^{F}, $\beta > 0$.

Definition 3.1: *For any positive β, Ψ_{β}^{F} (resp. Φ_{β}^{F}) denotes the set of functions $\psi \in L_F^2(W)$ (resp. $\varphi \in L_F^2(Z)$) such that:*

$$\sum_{i \geq 0} \frac{\langle \psi, \psi_i \rangle^2}{\lambda_i^{2\beta}} < +\infty, \quad \left(\text{resp. } \sum_{i \geq 0} \frac{\langle \varphi, \varphi_i \rangle^2}{\lambda_i^{2\beta}} < +\infty \right).$$

It is then clear that:

i) $\beta \leq \beta' \Longrightarrow \Psi_{\beta}^{F} \supset \Psi_{\beta'}^{F}$ *and* $\Phi_{\beta}^{F} \supset \Phi_{\beta'}^{F}$*;*

ii) $T\varphi = r$ *admits a solution* $\Rightarrow r \in \Psi_1^F$*;*

iii) $r \in \Psi_{\beta}^{F}$, $\beta > 1 \Longrightarrow \varphi \in \Phi_{\beta-1}^{F}$*;*

iv) $\Phi_{\beta}^{F} = \mathcal{R}\left[(T^*T)^{\frac{\beta}{2}} \right]$ *and*[5] $\Psi_{\beta}^{F} = \mathcal{R}\left[(TT^*)^{\frac{\beta}{2}} \right]$.

The condition $\varphi \in \Phi_{\beta}^{F}$ is called "*source condition*" (see e.g. Engl, Hanke and Neubauer (2000)). It involves both the properties of the solution φ (through its Fourier coefficients $< \varphi, \varphi_i >$) and of the conditional expectation operator T (through its singular values λ_i). As an example, Hall and Horowitz (2005) assume $< \varphi, \varphi_i > \sim \frac{1}{i^a}$ and $\lambda_i \sim \frac{1}{i^b}$. Then $\varphi \in \Phi_{\beta}^{F}$ if $\beta < \frac{1}{b}\left(a - \frac{1}{2}\right)$. However, it can be shown that choosing b is akin to choose the degree of smoothness of the joint probability density function of (Z, W). This is the reason why we will rather maintain here a high-level assumption $\varphi \in \Phi_{\beta}^{F}$, without being tightly constrained by specific rates. Generally speaking, it can be shown that the maximum value allowed for β depends on the degrees of smoothness of the solution φ (rate of decay of $< \varphi, \varphi_i >$) as well as on the degree of ill-posedness of the inverse problem (rate of decay of singular values λ_i)[6].

[5] The fractional power of an operator is trivially defined through its spectral decomposition, as in the elementary matrix case.

[6] A general study of the relationship between smoothness and Fourier coefficients is beyond the scope of this paper. It involves the concept of Hilbert scale (see Engl, Hanke and Neubauer (2000), Chen and Reiss (2007) and Johannes, Van Bellegem, Vanhems (2007)).

Assumption A.2: *For some real β, we have $\varphi \in \Phi_\beta^F$.*

The main reason why the spaces Φ_β^F are worthwhile to consider is the following result (see Carrasco, Florens, Renault (2007) p. 5679):

Proposition 3.2*: If $\varphi \in \Phi_\beta^F$ for some $\beta > 0$ and $\varphi^\alpha = (\alpha I + T^*T)^{-1}T^*T\varphi$, then $\|\varphi - \varphi^\alpha\|^2 = O(\alpha^{\beta \wedge 2})$ when α goes to zero.*

Even though the Tikhonov regularization scheme will be the only one used in all the theoretical developments of this paper, its main drawback is obvious from Proposition 3.2. It cannot take advantage of a degree of smoothness β for φ larger than 2: its so-called "*qualification*" is 2 (see Engl, Hanke and Neubauer (2000) for more details about this concept). However, iterating the Tikhonov regularization allows to increase its qualification. Let us consider the following sequence of iterated regularization schemes:

$$\left\{ \begin{array}{lcl} \varphi_{(1)}^\alpha & = & (\alpha I + T^*T)^{-1}T^*T\varphi \\ \varphi_{(k)}^\alpha & = & (\alpha I + T^*T)^{-1}\left[T^*T\varphi + \alpha\varphi_{(k-1)}^\alpha\right] \\ ... & & \end{array} \right. .$$

Then, it can be shown (see Engl, Hanke and Neubauer (2000), p. 123) that the qualification of $\varphi_{(k)}^\alpha$ is $2k$, that is: $\|\varphi - \varphi_{(k)}^\alpha\| = O(\alpha^{\beta \wedge 2k})$. To see this, note that:

$$\varphi_{(k)}^\alpha = \sum_{i \geq 0} \frac{(\lambda_i^2 + \alpha)^k - \alpha^k}{\lambda_i(\alpha + \lambda_i^2)^k} < \varphi, \varphi_i > \varphi_i.$$

Another way to increase the qualification of the Tikhonov regularization is to replace the norm of φ in (3.3) by a Sobolev norm (see Florens, Johannes, Van Bellegem (2007)).

4 Statistical inverse problem

4.1 Estimation

In order to estimate the regularized solution (3.2) by a Tikhonov method, we need to estimate T, T^*, and r. In this section, we introduce the kernel approach. We assume that Z and W take respectively values in $[0,1]^p$ and $[0,1]^q$. This assumption is not really restrictive, up to some monotone transformations. We start by introducing univariate generalized kernel functions of order l.

Definition 4.1: *Let $h \equiv h_N \to 0$ denote a bandwidth*[7] *and $K_h(\cdot,\cdot)$ denote a univariate generalized kernel function with the properties: $K_h(u,t) =$*

[7] We will use h and h_N interchangebly in the rest of this paper.

0 *if* $u > t$ *or* $u < t-1$*; for all* $t \in [0,1]$*,*

$$h^{-(j+1)} \int_{t-1}^{t} u^j K_h(u,t)\, du = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } 1 \leq j \leq l-1 \end{cases}.$$

We call $K_h(\cdot,\cdot)$ *a univariate generalized kernel function of order* l*.*

The following example is taken from Muller (1991). Specific examples of $K_+(\cdot,\cdot)$ and $K_-(\cdot,\cdot)$ are provided in Muller (1991).

Example 4.1: *Define:*

$$\mathcal{M}_{0,l}([a_1,a_2]) = \left\{ g \in \text{Lip}([a_1,a_2]),\ \int_{a_1}^{a_2} x^j g(x)\, dx = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } 1 \leq j \leq l-1 \end{cases} \right\},$$

where $\text{Lip}([a_1,a_2])$ *denotes the space of Lipschitz continuous functions on* $[a_1,a_2]$*. Define* $K_+(\cdot,\cdot)$ *and* $K_-(\cdot,\cdot)$ *as follows:*
(i) the support of $K_+(x,q')$ *is* $[-1,q'] \times [0,1]$ *and the support of* $K_-(x,q')$ *is* $[-q',1] \times [0,1]$*;*
(ii) $K_+(\cdot,q') \in \mathcal{M}_{0,l}([-1,q'])$ *and* $K_-(\cdot,q') \in \mathcal{M}_{0,l}([-q',1])$*.*
We note that $K_+(\cdot,1) = K_-(\cdot,1) = K(\cdot) \in \mathcal{M}_{0,l}([-1,1])$*. Now let:*

$$K_h(u,t) = \begin{cases} K_+(u,1) & \text{if } h \leq t \leq 1-h \\ K_+\left(\frac{u}{h}, \frac{t}{h}\right) & \text{if } 0 \leq t \leq h \\ K_-\left(\frac{u}{h}, \frac{1-t}{h}\right) & \text{if } 1-h \leq t \leq 1. \end{cases} \tag{4.1}$$

Then we can show that $K_h(\cdot,\cdot)$ *is a generalized kernel function of order* l*.*

A special class of multivariate generalized kernel functions of order l is given by that of products of univariate generalized kernel functions of order l. Let $K_{Z,h}$ and $K_{W,h}$ denote two generalized multivariate kernel functions of respective dimensions p and q. First we estimate the density functions $f_{Z,W}(z,w)$, $f_W(w)$, and $f_Z(z)$ [8]:

$$\widehat{f}_{Z,W}(z,w) = \frac{1}{Nh^{p+q}} \sum_{n=1}^{N} K_{Z,h}(z - z_n, z) K_{W,h}(w - w_n, w), \tag{4.2}$$

$$\widehat{f}_W(w) = \frac{1}{Nh^q} \sum_{n=1}^{N} K_{W,h}(w - w_n, w), \tag{4.3}$$

$$\widehat{f}_Z(z) = \frac{1}{Nh^p} \sum_{n=1}^{N} K_{Z,h}(z - z_n, z). \tag{4.4}$$

[8] For simplicity of notation and exposition, we use the same bandwidth to estimate $f_{Z,W}$, f_W, and f_Z. This can obviously be relaxed.

Then the estimators of T, T^* and r are:

$$(\hat{T}\varphi)(w) = \int \varphi(z) \frac{\widehat{f}_{Z,W}(z,w)}{\widehat{f}_W(w)} dz, \tag{4.5}$$

$$(\hat{T}^*\psi)(z) = \int \psi(w) \frac{\widehat{f}_{Z,W}(z,w)}{\widehat{f}_Z(z)} dw, \tag{4.6}$$

and

$$\hat{r}(w) = \frac{\sum_{n=1}^{N} y_n K_{W,h}(w - w_n, w)}{\sum_{n=1}^{N} K_{W,h}(w - w_n, w)}. \tag{4.7}$$

Note that $\hat{T}$ (resp. $\hat{T}^*$) is a finite rank operator from $L_F^2(Z)$ into $L_F^2(W)$ (resp. $L_F^2(W)$ into $L_F^2(Z)$). Moreover, $\hat{r}$ belongs to $L_F^2(W)$ and thus $\hat{T}^*\hat{r}$ is a well defined element of $L_F^2(Z)$. However, $\hat{T}^*$ is <u>not</u> in general the adjoint operator of $\hat{T}$. In particular, while T^*T is a nonnegative self-adjoint operator and thus $\alpha I + T^*T$ is invertible for any nonnegative α, it may not be the case for $\alpha I + \hat{T}^*\hat{T}$. Of course, for α given and consistent estimators $\hat{T}$ and $\hat{T}^*$, invertibility of $\alpha I + \hat{T}^*\hat{T}$ will be recovered for N sufficiently large. The estimator of φ is then obtained by estimating T^*, T and r in the first order condition (3.2) of the minimization (3.3).

Definition 4.2: *For $(\alpha_N)_{N>0}$ given sequence of positive real numbers, we call estimated instrumental regression function the function* $\hat{\varphi}^{\alpha_N} = (\alpha_N I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{r}$.

This estimator can be basically obtained by solving a linear system of N equations with N unknowns $\hat{\varphi}^{\alpha_N}(z_i)$, $i = 1, \ldots, N$, as it will be explained in Section 5 below.

4.2 Consistency and rate of convergence

Estimation of the instrumental regression as defined in Section 4.1 above requires consistent estimation of T^*, T and $r^* = T^*r$. The main objective of this section is to derive the statistical properties of the estimated instrumental regression function from the statistical properties of the estimators of T^*, T and r^*. Following Section 4.1, we use kernel smoothing techniques to simplify the exposition, but we could generalize the approach and use any other nonparametric techniques (for a sieve approach, see Ai and Chen (2003)). The crucial issue is actually the rate of convergence of nonparametric estimators of T^*, T and r^*. This rate is specified by Assumptions A.3 and A.4 below in relation with the bandwidth parameter chosen for all kernel estimators. We propose in Appendix B a justification of high

level Assumptions A.3 and A.4 through a set of more primitive sufficient conditions.

Assumption A.3: *There exists $\rho \geq 2$ such that:*

$$\|\hat{T} - T\|^2 = O_P\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right), \|\hat{T}^* - T^*\|^2 = O_P\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right),$$

where the norm in the equation is the supremum norm ($\|T\| = \sup_\varphi \|T\varphi\|$ with $\|\varphi\| \leq 1$).

Assumption A.4: $\|\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi\|^2 = O_P\left(\frac{1}{N} + h_N^{2\rho}\right)$.

Assumption A.4 is not about estimation of $r = E[Y \mid W]$ but only about estimation of $r^* = E[E[Y \mid W] \mid Z]$. The situation is even more favorable since we are not really interested in the whole estimation error about r^* but only one part of it:

$$\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi = \hat{T}^*[\hat{r} - \hat{T}\varphi].$$

The smoothing step by application of $\hat{T}^*$ allows us to get a parametric rate of convergence $1/N$ for the variance part of the estimation error.

We can then state the main result of the paper.

Theorem 4.1: *Under Assumptions A.1-A.4, we have:*

$$\|\hat{\varphi}^{\alpha_N} - \varphi\|^2 = O_P\left[\frac{1}{\alpha_N^2}\left(\frac{1}{N} + h_N^{2\rho}\right) + \left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right)\alpha_N^{(\beta-1)\wedge 0} + \alpha_N^{\beta\wedge 2}\right].$$

Corollary 4.1: *Under Assumptions A.1-A.4, if:*

- $\alpha_N \to 0$ *with* $N\alpha_N^2 \to \infty$,
- $h_N \to 0$ *with* $\left\{ \begin{array}{c} Nh_N^{p+q} \to \infty \\ Nh_N^{2\rho} \to c < \infty \end{array} \right.$,

  and

- $\beta \geqslant 1$ *or* $Nh_N^{p+q}\alpha_N^{1-\beta} \to \infty$,

Then:

$$\|\hat{\varphi}^{\alpha_N} - \varphi\|^2 \to 0.$$

To simplify the exposition, Corollary 4.1 is stated under the maintained assumption that $h_N^{2\rho}$ goes to zero at least as fast as $1/N$. Note that this assumption, jointly with the condition $Nh_N^{p+q} \to \infty$, implies that the degree ρ of regularity (order of differentiability of the joint density function of (Z, W) and order of the kernel) is larger than $\frac{p+q}{2}$. This constant is very little binding. For instance, it is fulfilled with $\rho = 2$ when considering $p = 1$ explanatory variable and $q = 2$ instruments.

The main message of Corollary 4.1 is that it is only when the relevance of instruments, that is the dependence between explanatory variables Z and instruments W is weak ($\beta < 1$) that consistency of our estimator takes more than the standard conditions on bandwidth (for the joint distribution of (Z, W)) and a regularization parameter.

Moreover, the cost of the nonparametric estimation of conditional expectations (see terms involving the bandwidth h_N) will under very general conditions be negligible in front of the two other terms $\frac{1}{N\alpha_N^2}$ and $\alpha_N^{\beta\wedge 2}$. To see this, first note that the optimal trade-off between these two terms leads to choose:

$$\alpha_N \propto N^{-\frac{1}{(\beta\wedge 2)+2}}.$$

The two terms are then equivalent:

$$\frac{1}{N\alpha_N^2} \sim \alpha_N^{\beta\wedge 2} \sim N^{-\frac{\beta\wedge 2}{(\beta\wedge 2)+2}},$$

and in general dominate the middle term:

$$\left[\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right] \alpha_N^{(\beta-1)\wedge 0} = O\left(\frac{\alpha_N^{(\beta-1)\wedge 0}}{Nh_N^{p+q}}\right),$$

under the maintained assumption $h_N^{2\rho} = O\left(\frac{1}{N}\right)$. More precisely, it is always possible to choose a bandwidth h_N such that:

$$\frac{1}{Nh_N^{p+q}} = O\left(\frac{\alpha_N^{\beta\wedge 2}}{\alpha_N^{(\beta-1)\wedge 0}}\right).$$

For $\alpha_N \propto N^{-\frac{1}{(\beta\wedge 2)+2}}$, it takes:

$$\frac{1}{h_N^{p+q}} = \begin{cases} O\left(N^{\frac{\beta+1}{\beta+2}}\right) \text{ when } \beta < 1, \\ O\left(N^{\frac{2}{(\beta\wedge 2)+2}}\right) \text{ when } \beta \geq 1, \end{cases}$$

which simply reinforce the constraint[9] $Nh_N^{p+q} \to \infty$. Since we maintain the assumption $h_N^{2\rho} = O\left(\frac{1}{N}\right)$, it simply takes:

$$\frac{p+q}{2\rho} \leq \begin{cases} \frac{\beta+1}{\beta+2} \text{ if } \beta < 1, \\ \frac{2}{(\beta\wedge 2)+2} \text{ if } \beta \geq 1. \end{cases}$$

To summarize, we have proved:

Corollary 4.2: *Under Assumptions A.1-A.4, if one of the two following conditions is fulfilled:*

(i) $\beta \geq 1$ *and* $\rho \geqslant [(\beta \wedge 2) + 2]\,\frac{p+q}{4}$,

(ii) $\beta < 1$ *and* $\rho \geqslant \left(\frac{\beta+2}{\beta+1}\right)\left(\frac{p+q}{2}\right)$.

Then, for α_N *proportional to* $N^{-\frac{1}{(\beta\wedge 2)+2}}$, *there exist bandwidth choices such that:*

$$\|\hat{\varphi}^{\alpha_N} - \varphi\|^2 = O_P\left[N^{-\frac{\beta\wedge 2}{(\beta\wedge 2)+2}}\right].$$

In other words, while the condition $\rho \geqslant \frac{p+q}{2}$ was always sufficient for the validity of Theorem 4.1, the stronger condition $\rho \geqslant p+q$ is always sufficient for Corollary 4.2.

Remark 4.1: We have presented the estimation part in the framework of kernel smoothing. However our result is more general and the rate of convergence given in Corollary 4.2 is actually minimax when the only maintained assumptions are Assumptions A.2, A.3 and A.4. This minimax property is easy to derive from the following heuristic argument, showing

[9] In fact, the stronger condition: $\left(Nh_N^{p+q}\right)^{-1}\log N \to 0$ (see Assumption B.4 in Appendix B) is satisfied with this choice of h_N.

that the bound given by Corollary 4.2 is sharp, that is, it may be reached in some circumstances. To show this result, there is no cost to assume that the operator T is known since we have seen that the estimation error on T does not play any role in the optimal rate of convergence. When T is known, the decomposition in the proof of Theorem 4.1 involves only two terms. The first term is due to the estimation error on r:

$$(\alpha_N I + T^*T)^{-1}(T^*\hat{r} - T^*T\varphi) = (\alpha_N I + T^*T)^{-1}T^*(\hat{r} - r), \tag{4.8}$$

and the second term is the regularization bias:

$$\varphi^{\alpha_N} - \varphi.$$

Let us denote:

$$T^*(\hat{r} - r) = \frac{\varepsilon}{\sqrt{N}},$$

where ε is a zero mean random element in $L^2_F(Z)$. It is consistent with Assumption A.4 to imagine that the (random) Fourier coefficients of ε with respect to the orthonormal system (φ_j) have a variance independent of N:

$$\rho_j^2 = E[< \varepsilon, \varphi_j >^2].$$

The key is then to relate the variance of the estimation error (4.8) to a Tikhonov regularization bias on a function $\Lambda = \sum_{j=1}^{\infty} \rho_j \varphi_j$:

$$E\left[[(\alpha_N I + T^*T)^{-1}T^*(\hat{r} - r)]^2\right] = \frac{1}{N}\sum_{j=1}^{\infty} \frac{\rho_j^2}{(\alpha_N + \lambda_j^2)^2} = \frac{1}{N\alpha_N^2}\left\|\Lambda^{\alpha_N} - \Lambda\right\|^2.$$

Hence, when $\Lambda \in \Phi_\gamma^F$ for some $\gamma > 0$, we have:

$$E\left[[(\alpha_N I + T^*T)^{-1}T^*(\hat{r} - r)]^2\right] = O\left(\frac{\alpha_N^\gamma}{N\alpha_N^2}\right).$$

However, from a minimax point of view, we cannot maintain any lower bound on $\gamma > 0$ and we can only say that the variance of the estimation error is at most $o\left(\frac{1}{N\alpha_N^2}\right)$. By contrast, we have the maintained Assumption A.2 ensuring that (under $\beta \leq 2$):

$$\left\|\varphi^{\alpha_N} - \varphi\right\|^2 = O(\alpha_N^\beta).$$

We then deduce the minimax rate of convergence by equalizing the speeds of the two parts of the above decomposition, namely $\frac{1}{N\alpha_N^2}$ and α_N^β. This minimax rate, reached for $\alpha_N^{\beta+2} = 1/N$, is $N^{-\frac{\beta}{\beta+2}}$, that is precisely the rate

given by Corollary 4.2.

Note that in the case of Hall and Horowitz (2005) with $\beta = \frac{1}{b}\left(a - \frac{1}{2}\right)$ (see comment after Definition 3.1) the minimax rate $\frac{\beta}{\beta+2} = \frac{a-\frac{1}{2}}{a+2b-\frac{1}{2}}$ provides a rate of convergence slower than the minimax rate in Hall and Horowitz (2005). This is due to the fact that they characterize the minimax rate within a more restricted family of errors $\hat{r} - r$.

In our presentation, we have two distinct regularity conditions: the differentiability of the joint density needed to control the kernel estimation properties, and the source condition on φ related to the singular value decomposition of T^*T. In some cases, these two kinds of assumptions may be linked (using in particular an Hilbert scale approach).

5 Numerical implementation and examples

Let us come back on the computation of the estimator $\hat{\varphi}^{\alpha_N}$. This estimator is a solution of the equation[10]:

$$(\alpha_N I + \hat{T}^*\hat{T})\varphi = \hat{T}^*\hat{r}, \tag{5.1}$$

where the estimators of T^*, T are linear forms of $\varphi \in L^2_F(Z)$ and $\psi \in L^2_F(W)$:

$$\hat{T}\varphi(w) = \sum_{n=1}^{N} a_n(\varphi) A_n(w),$$

$$\hat{T}^*\psi(z) = \sum_{n=1}^{N} b_n(\psi) B_n(z),$$

and

$$\hat{r}(w) = \sum_{n=1}^{N} y_n A_n(w),$$

with

$$a_n(\varphi) = \int \varphi(z) \frac{1}{h^p} K_{Z,h}(z - z_n, z) dz,$$

$$b_n(\psi) = \int \psi(w) \frac{1}{h^q} K_{W,h}(w - w_n, w) dw,$$

$$A_n(w) = \frac{K_{W,h}(w - w_n, w)}{\sum_{k=1}^{N} K_{W,h}(w - w_k, w)},$$

$$B_n(z) = \frac{K_{Z,h}(z - z_n, z)}{\sum_{k=1}^{N} K_{Z,h}(z - z_k, z)}.$$

[10] A more detailed presentation of the practice of nonparametric instrumental variable is given in Feve and Florens (2009).

Equation (5.1) is then equivalent to:

$$\alpha_N\varphi(z)+\sum_{m=1}^{N} b_m \left(\sum_{n=1}^{N} a_n(\varphi)A_n(w)\right) B_m(z) = \sum_{m=1}^{N} b_m \left(\sum_{n=1}^{N} y_n A_n(w)\right) B_m(z). \tag{5.2}$$

This equation is solved in two steps: first integrate the previous equation multiplied by $\frac{1}{h^p}K_{Z,h}(z-z_n,z)$ to reduce the functional equation to a linear system where the unknowns are $a_l(\varphi)$, $l=1,...,n$:

$$\alpha_N a_l(\varphi)+\sum_{m,n=1}^{N} a_n(\varphi)b_m\left(A_n(w)\right)a_l\left(B_m(z)\right) = \sum_{m,n=1}^{N} y_n b_m\left(A_n(w)\right)a_l\left(B_m(z)\right),$$

or

$$\alpha_N\overline{a} + EF\overline{a} = EF\overline{y},$$

with

$$\begin{aligned}
\overline{a} &= (a_l(\varphi))_l,\\
\overline{y} &= (y_n)_n,\\
E &= \left(b_m\left(A_n(w)\right)\right)_{n,m},\\
F &= \left(a_l\left(B_m(z)\right)\right)_{l,m}.
\end{aligned}$$

The last equation can be solved directly to get the solution $\overline{a} = (\alpha_N + EF)^{-1}EF\overline{y}$.

In a second step, Equation (5.2) is used to compute φ at any value of z. These computations can be simplified if we use the approximation $a_l(\varphi) \simeq \varphi(z_l)$ and $b_l(\psi) \simeq \psi(w_l)$. Equation (5.2) is then a linear system where the unknowns are the $\varphi(z_n)$, $n=1,...,N$.

In order to illustrate the power of our approach and its simplicity we present the following simulated example. The data generating process is:

$$\begin{cases} Y = \varphi(Z) + U \\ Z = 0.1W_1 + 0.1W_2 + V, \end{cases}$$

where:

$$\begin{aligned}
W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right)\\
V &\sim N\left(0,(0.27)^2\right)\\
U = -0.5V + \varepsilon \qquad & \varepsilon \sim N\left(0,(0.05)^2\right)\\
W, V, \varepsilon &\text{ mutually independent.}
\end{aligned}$$

The function $\varphi(Z)$ is chosen equal to Z^2 (which represents a maximal order of regularity in our model, i.e. $\beta = 2$) or $e^{-|Z|}$ which is highly irregular. The bandwidths for kernel estimation are chosen equal to .45 (kernel on Z variable) or .9 (kernel on W variable) for $\varphi(Z) = Z^2$ and .45 and .25 in the $e^{-|Z|}$ case. For each selection of φ we show the estimation for α_N varying in a very large range and selection of this parameter appears naturally. All the kernels are Gaussian[11]. For $\varphi(Z) = Z^2$, we present in Graph 1 the set of data ($N = 1000$) in the (Z, Y) space, the true function, the kernel estimation of the regression and our estimation. In Graph 2, we show the evolution of our estimator for different values of α_N and in Graph 3 a Monte Carlo analysis is performed: a sample is generated 150 times and the estimation of φ is performed with the same bandwidths and same regularization parameter as in Graph 1. All these curves are plotted and give an illustration of their distribution. Finally Graph 4 is identical to Graph 1 with $\varphi(Z) = e^{-|Z|}$ and Graph 5 corresponds to Graph 2 in this case.

[Insert here Figure 1: Numerical Implementation]

Let us stress that the endogeneity bias in the estimation of the regression by kernel smoothing clearly appears. The estimated φ curve is not obviously related to the sample of Z and Y and depends on the instrumental variables W. Even though they cannot be represented, the instruments play a central role in the estimation.

The main question about the practical use of nonparametric instrumental variables estimation is the selection of the bandwidth and of the α_N parameter. This question is complex and the construction of a data driven procedure for the simultaneous selection of h_N and α_N is still an open question. We propose the following sequential method:

i) Fix first the bandwidths for the estimation of r and of the joint density of Z and W (for the estimation of T and T^*) by usual methods. Note that these bandwidths do not need to be equal for the two estimations.

ii) Select α_N by a data driven method. We suggest the following method based on a residual approach extending the discrepancy principle of Morozov (1993).

We consider the "extended residuals" of the model defined by:

$$\varepsilon^{\alpha_N} = \hat{T}^* \hat{r} - \hat{T}^* \hat{T} \hat{\varphi}_{(2)}^{\alpha_N},$$

[11] Note that for simplicity we have not used generalized kernels in the simulation.

where $\hat{\varphi}_{(2)}^{\alpha_N}$ is the iterated Tikhonov estimation of order 2. Then:

$$\|\varepsilon^{\alpha_N}\| \leq \|\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi\| + \|\hat{T}^*\hat{T}\varphi - \hat{T}^*\hat{T}\hat{\varphi}_{(2)}^{\alpha}\|.$$

To simplify the exposition, let's assume $h_N^{2\rho}$ goes to zero at least as fast as $1/N$. Then Assumption A.4 implies that the first term on the right hand side of the above displayed inequality is $O_P(\frac{1}{\sqrt{N}})$. Under the previous assumptions it can be shown that $\|\hat{T}^*\hat{T}(\hat{\varphi}_{(2)}^{\alpha_N} - \varphi)\|^2 = \left\|\left(\hat{T}^*\hat{T}\varphi\right)_{(2)}^{\alpha_N} - \hat{T}^*\hat{T}\varphi\right\|^2 = O_P(\alpha_N)$. This last property requires a regularization method of qualification at least 4 in order to characterize a β not greater than 2, and this is the motivation for the use of an iterated Tikhonov estimation at the first stage. Then we have:

$$\frac{1}{\alpha_N^2}\|\varepsilon^{\alpha_N}\|^2 = O_P(\frac{1}{\alpha_N^2 N} + \alpha_N^{(\beta+2)\wedge 4}),$$

and a minimization with respect to α_N of this value gives an α_N with an optimal speed $(N^{-\frac{1}{\beta+2}})$ for the use in a non iterated Tikhonov estimation. In practice $\frac{1}{\alpha_N^2}\|\varepsilon^{\alpha_N}\|^2$ may be computed for different values of α_N and the minimum can be selected. We give in Graph 6 this curve in the example of $\varphi(Z) = e^{-|Z|}$.

6 Conclusion

This paper has considered the nonparametric estimation of a regression function in presence of a simultaneity problem. We have established a set of general sufficient conditions to ensure consistency and asymptotic normality of our nonparametric instrumental variables estimator. The discussion of rates of convergence emphasizes the crucial role of the degree of ill-posedness of the inverse problem whose unique solution defines the regression function. A Monte Carlo illustration shows that our estimator is rather easy to implement and able to correct for the simultaneity bias displayed by the naive kernel estimator. This paper treats essentially the purely nonparametric basic model and is in particular focused on kernel-based estimators. Numerous extensions are possible and relevant for the practical implementation of this procedure.

1. A first extension is to analyze the case where the explanatory variables Z contain exogenous variables also included in the instrumental variables W. These variables may be introduced in a nonparametric way or semi nonparametrically ($\varphi(Z)$ becomes $\varphi(Z) + X'\beta$ with X exogenous). In the general case, results are essentially the same as in our

paper by fixing these variables (see Hall and Horowitz (2005)). In the semi parametric case the procedure is described in Feve and Florens (2009).

2. The treatment of semi parametric models (additive, partially linear, index models,...) (see Florens, Johannes and Van Bellegem (2005), Ai and Chen (2003)) or nonparametric models with constraints is helpful to reduce the curse of dimensionality.

3. We need to improve and to study more deeply the adaptive selection of the bandwidths and of the regularization parameter.

4. The structure L^2 of the spaces may be modified. In particular Sobolev spaces may be used and the penalty norm may incorporate the derivatives (see Gagliardini and Scaillet (2006)). This approach is naturally extended in terms of Hilbert scales (see Florens, Johannes and Van Bellegem (2007)).

5. Separable models may be extended to non separable models or more generally to non linear problems (duration models, auctions, GMM, dynamic models) (see Ai and Chen (2003)).

6. A Bayesian approach to the nonparametric instrumental variables estimation (Florens and Simoni (2007)) enhances a use of gaussian process prior similar to machine learning.

7. In a preliminary version of this paper we give a proof of the asymptotic normality of $\langle \widehat{\varphi}^{\alpha_N} - \varphi, \delta \rangle$. This result is now exposed in a separate paper.

APPENDIX

A Proofs

A.1 Proof of Proposition 2.1

$i) \Longleftrightarrow ii)$: $ii)$ implies $i)$. Conversely, let us consider φ such that:

$$T^*T[\varphi(Z)] = E[E[\varphi(Z) \mid W] \mid Z] = 0.$$

Then:

$$\begin{aligned} E\left[E[\varphi(Z) \mid W]^2\right] &= E[\varphi(Z)\, E[\varphi(Z) \mid W]] \\ &= E[\varphi(Z)\, E[E[\varphi(Z) \mid W] \mid Z]] = 0. \end{aligned}$$

We obtain $E[\varphi(Z) \mid W] = 0$ and $\varphi = 0$ using the strong identification condition.

$i) \Longleftrightarrow iii)$: This property can be deduced from Florens-Mouchart-Rolin (1990), Theorem 5.4.3 or Luenberger (1969), Theorem 3 section 6.3. Since $\mathcal{R}(T^*) = \mathcal{N}(T)^{\perp}, \overline{\mathcal{R}(T^*)} = L_F^2(Z)$ is tantamount to $\mathcal{N}(T) = \{0\}$.

A.2 Proof of Theorem 2.1

Let us first remark that:

$$\begin{aligned} &\int \frac{f_{Z,W_1}^2(z,w_1)}{f_Z^2(z)\, f_{W_1}^2(w_1)} f_Z(z)\, f_{W_1}(w_1) dz dw_1 \\ &= \int \left\{ \int \frac{f_{Z,W}(z,w_1,w_2)}{f_Z(z)\, f_W(w_1,w_2)} f_{W_2|W_1}(w_2 \mid w_1) dw_2 \right\}^2 f_Z(z)\, f_{W_1}(w_1) dz dw_1 \\ &\leq \int \frac{f_{Z,W}^2(z,w_1,w_2)}{f_Z^2(z)\, f_W^2(w_1,w_2)} f_Z(z) f_W(w_1,w_2) dz dw_1 dw_2, \end{aligned}$$

by Jensen's inequality for conditional expectations. The first term is the Hilbert Schmidt norm of $T_1^*T_1$ and the last one is the Hilbert Schmidt norm of $T^*\ T$. Then $T_1^*T_1$ is an Hilbert Schmidt operator and $\sum_j \lambda_{j,1}^2 \leq \sum_j \lambda_j^2$.

The eigenvalues may be compared pairwise. Using the Courant theorem

(see Kress (1999), 15), we get:

$$\begin{aligned}
\lambda_j^2 &= \min_{\rho_0,\rho_1,\ldots,\rho_{j-1}\in L_z^2} \max_{\substack{\|\varphi\|=1 \\ \varphi\perp(\rho_0,\rho_1,\ldots,\rho_{j-1})}} \langle T^*T\varphi,\varphi\rangle \\
&= \max_{\substack{\|\varphi\|=1 \\ \varphi\perp(\rho_0,\rho_1,\ldots,\rho_{j-1})}} \|E(\varphi|w)\|^2 \\
&\geq \max_{\substack{\|\varphi\|=1 \\ \varphi\perp(\rho_0,\rho_1,\ldots,\rho_{j-1})}} \|E(\varphi|w_1)\|^2 \\
&\geq \min_{\rho_0,\rho_1,\ldots,\rho_{j-1}\in L_z^2} \max_{\substack{\|\varphi\|=1 \\ \varphi\perp(\rho_0,\rho_1,\ldots,\rho_{j-1})}} \left\langle T^{1*}T^1\varphi,\varphi\right\rangle \\
&= \lambda_{j,1}^2.
\end{aligned}$$

A.3 Proof of Theorem 4.1

The proof of Theorem 4.1 is based upon the decomposition:

$$\hat{\varphi}^{\alpha_N} - \varphi = A_1 + A_2 + A_3,$$

with

$$\begin{aligned}
A_1 &= (\alpha_N I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{r} - (\alpha_N I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\varphi \\
A_2 &= (\alpha_N I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\varphi - (\alpha_N I + T^*T)^{-1}T^*T\varphi \\
A_3 &= (\alpha_N I + T^*T)^{-1}T^*T\varphi - \varphi
\end{aligned}$$

By Proposition 3.2:

$$\|A_3\|^2 = O\left(\alpha_N^{\beta\wedge 2}\right),$$

and, by virtue of Assumption A.4, we have directly:

$$\|A_1\|^2 = O_P\left[\frac{1}{\alpha_N^2}\left(\frac{1}{N} + h_N^{2\rho}\right)\right].$$

To assess the order of A_2, it is worth rewriting it as:

$$\begin{aligned}
A_2 &= \alpha_N\left[(\alpha_N I + \hat{T}^*\hat{T})^{-1} - (\alpha_N I + T^*T)^{-1}\right]\varphi \\
&= -\alpha_N(\alpha_N I + \hat{T}^*\hat{T})^{-1}(\hat{T}^*\hat{T} - T^*T)(\alpha_N I + T^*T)^{-1}\varphi \\
&= -\left(B_1 + B_2\right),
\end{aligned}$$

with

$$\begin{aligned}
B_1 &= \alpha_N(\alpha_N I + \hat{T}^*\hat{T})^{-1}\hat{T}^*(\hat{T} - T)(\alpha_N I + T^*T)^{-1}\varphi \\
B_2 &= \alpha_N(\alpha_N I + \hat{T}^*\hat{T})^{-1}(\hat{T}^* - T^*)T(\alpha_N I + T^*T)^{-1}\varphi
\end{aligned}$$

By Assumption A.3:

$$\|\hat{T} - T\|^2 = O_P\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right),$$

$$\|\hat{T}^* - T^*\|^2 = O_P\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right),$$

and, by Proposition 3.2:

$$\left\|\alpha_N(\alpha_N I + T^*T)^{-1}\varphi\right\|^2 = O\left(\alpha_N^{\beta\wedge 2}\right),$$

$$\left\|\alpha_N T(\alpha_N I + T^*T)^{-1}\varphi\right\|^2 = O\left(\alpha_N^{(\beta+1)\wedge 2}\right),$$

while

$$\left\|(\alpha_N I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\right\|^2 = O_P\left(\frac{1}{\alpha_N}\right),$$

$$\left\|(\alpha_N I + \hat{T}^*\hat{T})^{-1}\right\|^2 = O_P\left(\frac{1}{\alpha_N^2}\right).$$

Therefore

$$\begin{aligned}
\|A_2\|^2 &= O_P\left[\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right)\left(\frac{\alpha_N^{\beta\wedge 2}}{\alpha_N} + \frac{\alpha_N^{(\beta+1)\wedge 2}}{\alpha_N^2}\right)\right] \\
&= O_P\left[\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right)\left(\alpha_N^{(\beta-1)\wedge 1} + \alpha_N^{(\beta-1)\wedge 0}\right)\right] \\
&= O_P\left[\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right)\alpha_N^{(\beta-1)\wedge 0}\right].
\end{aligned}$$

B A discussion

The objective of this appendix is to give a set of *primitive* conditions which imply the main assumptions of the paper for the kernel estimator. For notational compactness, in this appendix, we will suppress the subscripts in $\widehat{f}_W(w)$, $\widehat{f}_Z(z)$, and $\widehat{f}_{Z,W}(z,w)$ and the corresponding pdfs. They will be distinguished by their arguments. We will also suppress the subscript in h_N. We use C to denote a generic positive constant which may take different values in different places and adopt the following assumptions.

Assumption B.1: (i) *The data* (y_n, z_n, w_n), $n = 1, ..., N$, *define an i.i.d sample of* (Y, Z, W); (ii) *The pdf* $f(z,w)$ *is* d *times continuously differentiable in the interior of* $[0,1]^p \times [0,1]^q$.

Assumption B.2: *The pdf* $f(z,w)$ *is bounded away from zero on the support* $[0,1]^p \times [0,1]^q$.

Assumption B.3: *Both multivariate kernels* $K_{Z,h}$ *and* $K_{W,h}$ *are product kernels generated from the univariate generalized kernel function* K_h *satisfying: (i) the kernel function* $K_h(\cdot,\cdot)$ *is a generalized kernel function of order* l*; (ii) for each* $t \in [0,1]$*, the function* $K_h(h\cdot, t)$ *is supported on* $[(t-1)/h, t/h] \cap \mathcal{K}$*, where* $\mathcal{K}$ *is a compact interval not depending on* t *and:*

$$\sup_{h>0, t\in[0,1], u\in\mathcal{K}} |K_h(hu,t)| < \infty.$$

Assumption B.4: *The smoothing parameter satisfies:* $h \to 0$ *and* $(Nh^{p+q})^{-1} \log N \to 0$.

The independence assumption is a simplifying assumption and could be extended to weakly dependent (stationary mixing) observations. Assumption B.3 is the same as A.5 in Hall and Horowitz (2005). We first provide a result on the uniform convergence of $\widehat{f}(w)$, $\widehat{f}(z)$, and $\widehat{f}(z,w)$ with rates. For density functions with compact support, uniform convergence of kernel density estimators using ordinary kernel functions must be restricted to a proper subset of the compact support. Using generalized kernel functions, we show uniform convergence over the entire support. A similar result is provided in Proposition 2 (ii) in Rothe (2009). However, the assumptions in Rothe (2009) differ from our assumptions and no proof is provided in Rothe (2009).

Lemma B.1: *Suppose Assumptions B.1-B.4 hold. Let* $\rho = \min\{l, d\}$. *Then:*
(i)

$$\sup_{w\in[0,1]^q} \left|\widehat{f}(w) - f(w)\right| = O_P\left(\left[(Nh^q)^{-1}\log N\right]^{1/2} + h^\rho\right) = o_P(1);$$

(ii)

$$\sup_{z\in[0,1]^p, w\in[0,1]^q} \left|\widehat{f}(z,w) - f(z,w)\right| = O_P\left(\left[\left(Nh^{p+q}\right)^{-1}\log N\right]^{1/2} + h^{\rho}\right) = o_P(1);$$

(iii)

$$\sup_{z\in[0,1]^p} \left|\widehat{f}(z) - f(z)\right| = O_P\left(\left[(Nh^{p})^{-1}\log N\right]^{1/2} + h^{\rho}\right) = o_P(1).$$

Proof. We provide a proof of (i) only. First we evaluate the bias of $\widehat{f}(w)$. Let $w = (w_1, ..., w_q)'$. Then:

$$\begin{aligned}
&E\left(\widehat{f}(w)\right) \\
&= \frac{1}{h^q} E\left[K_{W,h}(w - w_n, w)\right] \\
&= \frac{1}{h^q}\int_{[0,1]^q} K_{W,h}(w - v, w) f(v)\, dv \\
&= \int_{\Pi_{j=1}^q \left[\frac{w_j-1}{h}, \frac{w_j}{h}\right]} K_{W,h}(hv, w) f(w - hv)\, dv \\
&= \int_{\Pi_{j=1}^q \left[\frac{w_j-1}{h}, \frac{w_j}{h}\right]} K_{W,h}(hv, w) \left[\begin{array}{c} f(w) + (-h)\sum_{j=1}^q \frac{\partial f(w)}{\partial w_j} v_j + \cdots \\ +\frac{1}{\rho!}\sum_{j_1=1}^q \cdots \sum_{j_\rho=1}^q \frac{\partial^\rho f(w^*)}{\partial w_{j_1}\cdots\partial w_{j_\rho}} (-h)^\rho v_{j_1}\cdots v_{j_\rho} \end{array}\right] dv,
\end{aligned}$$

where w^* lies between w and $(w - hv)$. Now making use of Assumptions B.2-B.4, we get:

$$\sup_{w\in[0,1]^q} \left|E\left(\widehat{f}(w)\right) - f(w)\right| \leq Ch^{\rho}\left[\sup_{h>0, t\in[0,1], u\in\mathcal{K}} |K_h(hu, t)|\right]^q = O(h^{\rho}).$$

It remains to show: $\sup_{w\in[0,1]^q}\left|\widehat{f}(w) - E\left[\widehat{f}(w)\right]\right| = O_P\left(\left[(Nh^q)^{-1}\log N\right]^{1/2}\right)$. This can be shown by the standard arguments in the proof of uniform consistency of kernel density estimators based on ordinary kernel functions, see e.g., Hansen (2008) and references therein.

The next lemma shows that Assumption A.3 is satisfied under the previous conditions. Actually, this lemma proves a stronger result as the one needed for Assumption A.3 because the convergence is proved in Hilbert Schmidt norm which implies the convergence for the supremum norm.

Lemma B.2: *Suppose Assumptions B.1-B.4 hold. Then:*
(i) $\left\|\widehat{T} - T\right\|_{HS}^2 = O_P\left(\left(Nh^{p+q}\right)^{-1} + h^{2\rho}\right)$,

(ii) $\left\|\widehat{T}^{*}-T^{*}\right\|_{HS}^{2}=O_{P}\left(\left(Nh^{p+q}\right)^{-1}+h^{2\rho}\right)$,
where $\|\cdot\|_{HS}$ *denotes the Hilbert-Schmidt norm, i.e.:*

$$\begin{aligned}\left\|\widehat{T}-T\right\|_{HS}^{2} &= \int_{[0,1]^{q}}\int_{[0,1]^{p}}\frac{\left[\widehat{f}\left(z|w\right)-f\left(z|w\right)\right]^{2}}{f^{2}\left(z\right)}f\left(z\right)f\left(w\right)dzdw \\ &= \int_{[0,1]^{q}}\int_{[0,1]^{p}}\left[\frac{\widehat{f}\left(z,w\right)}{\widehat{f}\left(w\right)}-\frac{f\left(z,w\right)}{f\left(w\right)}\right]^{2}\frac{f\left(w\right)}{f\left(z\right)}dzdw.\end{aligned}$$

Proof of (i). Let $\int\int\cdot\,dzdw=\int_{[0,1]^{q}}\int_{[0,1]^{p}}\cdot\,dzdw$. Note that:

$$\begin{aligned}&\left\|\widehat{T}-T\right\|_{HS}^{2} \\ &=\int\int\left[\frac{\widehat{f}\left(z,w\right)}{\widehat{f}\left(w\right)}-\frac{f\left(z,w\right)}{f\left(w\right)}\right]^{2}\frac{f\left(w\right)}{f\left(z\right)}dzdw \\ &=\int\int\left[\frac{\widehat{f}\left(z,w\right)f\left(w\right)-f\left(z,w\right)\widehat{f}\left(w\right)}{\widehat{f}\left(w\right)f\left(w\right)}\right]^{2}\frac{f\left(w\right)}{f\left(z\right)}dzdw \\ &\leq\frac{1}{\inf_{w\in[0,1]^{q}}\left[\widehat{f}\left(w\right)\right]^{2}}\int\int\left[\frac{\widehat{f}\left(z,w\right)f\left(w\right)-f\left(z,w\right)\widehat{f}\left(w\right)}{f\left(w\right)}\right]^{2}\frac{f\left(w\right)}{f\left(z\right)}dzdw \\ &=O_{P}\left(1\right)\int\int\left[\widehat{f}\left(z,w\right)-f\left(z,w\right)-\frac{f\left(z,w\right)\left[\widehat{f}\left(w\right)-f\left(w\right)\right]}{f\left(w\right)}\right]^{2}\frac{f\left(w\right)}{f\left(z\right)}dzdw \\ &=O_{P}\left(1\right)\int\int\frac{\left[\widehat{f}\left(z,w\right)-f\left(z,w\right)\right]^{2}}{f\left(z\right)}f\left(w\right)dzdw+O_{P}\left(1\right)\int\int\frac{f^{2}\left(z,w\right)\left[\widehat{f}\left(w\right)-f\left(w\right)\right]^{2}}{f\left(w\right)f\left(z\right)}dzdw \\ &\equiv O_{P}\left(1\right)\left(A_{1}+A_{2}\right),\end{aligned}$$

where we have used the fact that $\frac{1}{\inf_{w\in[0,1]^{q}}\left[\widehat{f}(w)\right]^{2}}=O_{P}\left(1\right)$ implied by Assumptions B.2, B.4, and Lemma B.1. Now, we show:

$$A_{1}=O_{P}\left(\left(Nh^{p+q}\right)^{-1}+h^{2\rho}\right)\text{ and }A_{2}=O_{P}\left(\left(Nh^{q}\right)^{-1}+h^{2\rho}\right).$$

As a result, we obtain $\left\|\widehat{T}-T\right\|_{HS}^{2}=O_{P}\left(\left(Nh^{p+q}\right)^{-1}+h^{2\rho}\right)$.

We prove the result for A_1. Note that:

$$\begin{aligned} E\left(|A_1|\right) &= \int\int \frac{E\left[\widehat{f}\left(z,w\right)-f\left(z,w\right)\right]^2}{f\left(z\right)} f\left(w\right) dzdw \\ &= \int\int Var\left(\widehat{f}\left(z,w\right)\right)\frac{f\left(w\right)}{f\left(z\right)}dzdw \\ &+ \int\int \left[E\left(\widehat{f}\left(z,w\right)\right)-f\left(z,w\right)\right]^2 \frac{f\left(w\right)}{f\left(z\right)}dzdw \\ &= O\left(\left(Nh^{p+q}\right)^{-1}\right)+O\left(h^{2\rho}\right). \end{aligned}$$

This follows from the standard arguments for evaluating the first term and the proof of Lemma B.1 for the second term. By Markov inequality, we obtain $A_1 = O_P\left(\left(Nh^{p+q}\right)^{-1}+h^{2\rho}\right)$.

The next lemma shows that Assumption A.4 is satisfied under the primitive conditions.

Lemma B.3 *Suppose Assumptions B.1-B.4 hold. In addition, we assume* $E\left(U^2|W=w\right)$ *is uniformly bounded in* $w\in[0,1]^q$. *Then:*
$\left\|\widehat{T}^*\widehat{r}-\widehat{T}^*\widehat{T}\varphi\right\|^2 = O_P\left(N^{-1}+h^{2\rho}\right)$.

Proof. By definition:

$$\begin{aligned} \left(\widehat{T}^*\widehat{r}-\widehat{T}^*\widehat{T}\varphi\right)(z) &= \left[\widehat{T}^*\left(\widehat{r}-\widehat{T}\varphi\right)\right](z) \\ &= \int\left(\widehat{r}-\widehat{T}\varphi\right)(w)\frac{\widehat{f}(z,w)}{\widehat{f}(z)}dw \\ &= \int\left(\widehat{r}\left(w\right)-\int\varphi\left(z'\right)\frac{\widehat{f}(z',w)}{\widehat{f}(w)}dz'\right)\frac{\widehat{f}(z,w)}{\widehat{f}(z)}dw \\ &= \int\left(\frac{1}{Nh^q}\sum_{n=1}^{N}y_nK_{W,h}(w-w_n,w)-\int\varphi\left(z'\right)\widehat{f}(z',w)dz'\right)\frac{\widehat{f}(z,w)}{\widehat{f}(z)\widehat{f}(w)}dw \\ &\equiv \int A_N\left(w\right)\frac{\widehat{f}(z,w)}{\widehat{f}(z)\widehat{f}(w)}dw. \end{aligned}$$

Similar to the proof of Lemma B.2, we can show by using Lemma B.1 that uniformly in $z\in[0,1]^p$, the following holds:

$$\left(\widehat{T}^*\widehat{r}-\widehat{T}^*\widehat{T}\varphi\right)(z) = \int A_N\left(w\right)\frac{f(z,w)}{f(z)f(w)}dw+o_P\left(\int A_N\left(w\right)\frac{f(z,w)}{f(z)f(w)}dw\right).$$

Thus, it suffices to show that $\left\|\int A_N(w)\frac{f(z,w)}{f(z)f(w)}dw\right\|^2 = O_P\left(N^{-1}+h^{2\rho}\right)$. Writing $A_N(w)$ as:

$$\begin{aligned}A_N(w) &= \frac{1}{Nh^q}\sum_{n=1}^{N}U_nK_{W,h}(w-w_n,w)\\ &\quad+\frac{1}{Nh^q}\sum_{n=1}^{N}\left[\varphi(z_n)-\frac{1}{h^p}\int\varphi(z')K_{Z,h}(z'-z_n,z')dz'\right]K_{W,h}(w-w_n,w)\\ &\equiv A_{N1}(w)+A_{N2}(w),\end{aligned}$$

we obtain:

$$\begin{aligned}&E\left[\left\|\int A_N(w)\frac{f(z,w)}{f(z)f(w)}dw\right\|^2\right]\\ &\le 2E\left[\left\|\int A_{N1}(w)\frac{f(z,w)}{f(z)f(w)}dw\right\|^2+\left\|\int A_{N2}(w)\frac{f(z,w)}{f(z)f(w)}dw\right\|^2\right]\\ &= 2\int\int\int E\left[A_{N1}(w)A_{N1}(w')\right]\frac{f(z,w)f(z,w')}{f(z)f(w)f(w')}dwdw'dz\\ &+2\int\int\int E\left[A_{N2}(w)A_{N2}(w')\right]\frac{f(z,w)f(z,w')}{f(z)f(w)f(w')}dwdw'dz\\ &= 2B_{N1}+2B_{N2}.\end{aligned}$$

Below, we will show that $B_{N1}=O\left(N^{-1}\right)$ and $B_{N2}=O\left(N^{-1}+h^{2\rho}\right)$. First consider the term B_{N1}:

$$\begin{aligned}B_{N1} &= \frac{1}{Nh^{2q}}\int\int\int E\left[U_n^2K_{W,h}(w-w_n,w)K_{W,h}(w'-w_n,w')\right]\frac{f(z,w)f(z,w')}{f(z)f(w)f(w')}dwdw'dz\\ &= \frac{1}{N}\int E\left[\begin{array}{c}\int_{-w_n/h}^{(1-w_n)/h}\int_{-w_n/h}^{(1-w_n)/h}U_n^2K_{W,h}(hw,w_n+hw)K_{W,h}(hw',w_n+hw')\\ \frac{f(z,w_n+hw)f(z,w_n+hw')}{f(z)f(w_n+hw)f(w_n+hw')}dwdw'\end{array}\right]dz\\ &\le CN^{-1}\left[\sup_{h>0,t\in[0,1],u\in\mathcal{K}}|K_h(hu,t)|\right]^{2q}\\ &= O\left(N^{-1}\right),\end{aligned}$$

under the conditions of Lemma B.3.

Now for B_{N2}, letting $B(z_n)=\varphi(z_n)-\frac{1}{h^p}\int\varphi(z')K_{Z,h}(z'-z_n,z')dz'$,

we get:

$$
\begin{aligned}
&B_{N2} \\
&= \frac{1}{\left(Nh_N^q\right)^2} \sum \sum_{n \neq n'} \int \int \int \\
&E\left[B\left(z_n\right) K_{W,h}(w - w_n, w)\right] E\left[B\left(z_{n'}\right) K_{W,h}(w' - w_{n'}, w')\right] \frac{f(z,w) f\left(z,w'\right)}{f(z) f(w) f\left(w'\right)} dw dw' dz \\
&+ \frac{1}{Nh^{2q}} \int \int \int E\left[\left[B\left(z_n\right)\right]^2 K_{W,h}(w - w_n, w) K_{W,h}(w' - w_n, w')\right] \frac{f(z,w) f\left(z,w'\right)}{f(z) f(w) f\left(w'\right)} dw dw' dz \\
&= \frac{1}{\left(Nh^q\right)^2} \sum \sum_{n \neq n'} \int \left[\int E\left[B\left(z_n\right) K_{W,h}(w - w_n, w)\right] \frac{f\left(z,w\right)}{f\left(w\right)} dw\right]^2 \frac{1}{f(z)} dz \\
&+ \frac{1}{Nh^{2q}} \int \int \int E\left[\left[B\left(z_n\right)\right]^2 K_{W,h}(w - w_n, w) K_{W,h}(w' - w_{n'}, w')\right] \frac{f(z,w) f\left(z,w'\right)}{f(z) f(w) f\left(w'\right)} dw dw' dz \\
&= \frac{1}{\left(Nh^q\right)^2} \sum \sum_{n \neq n'} \int \left[\int E\left[B\left(z_n\right) K_{W,h}(w - w_n, w)\right] \frac{f\left(z,w\right)}{f\left(w\right)} dw\right]^2 \frac{1}{f(z)} dz \\
&+ O\left(N^{-1}\right),
\end{aligned}
$$

where the second term on the right hand side of the second last equation can be shown to be $O\left(N^{-1}\right)$ by change of variables and by using Assumptions B.1-B.4. Let B_{N21} denote the first term, i.e.,

$$
B_{N21} = \frac{1}{\left(Nh^q\right)^2} \sum \sum_{n \neq n'} \int \left[\int E\left[B\left(z_n\right) K_{W,h}(w - w_n, w)\right] \frac{f\left(z,w\right)}{f\left(w\right)} dw\right]^2 \frac{1}{f(z)} dz.
$$

Then it suffices to show that $B_{N21} = O(h^{2\rho})$. Similar to the proof of Lemma B.1, we note that, uniformly in $z \in [0,1]^p$, we get:

$$
\begin{aligned}
&\frac{1}{h^q} \int E\left[B\left(z_n\right) K_{W,h}(w - w_n, w)\right] \frac{f\left(z,w\right)}{f\left(w\right)} dw \\
&= \frac{1}{h^q} \int E\left[\varphi\left(z_n\right) K_{W,h}(w - w_n, w)\right] \frac{f\left(z,w\right)}{f\left(w\right)} dw \\
&- \int \int \varphi\left(z'\right) E\left[\frac{1}{h^{p+q}} K_{Z,h}(z' - z_n, z') K_{W,h}(w - w_n, w)\right] dz' \frac{f\left(z,w\right)}{f\left(w\right)} dw \\
&= O\left(h^{2\rho}\right).
\end{aligned}
$$

As a result, we obtain $B_{N1} + B_{N2} = O\left(N^{-1} + h^{2\rho}\right)$ or $E\left[\left\|\widehat{T}^* \widehat{r} - \widehat{T}^* \widehat{T} \varphi\right\|^2\right] = O\left(N^{-1} + h^{2\rho}\right)$. By Markov inequality, we obtain the result in Lemma B.3.

REFERENCES

Ai, C. and X. Chen (2003), *Efficient Estimation of Conditional Moment Restrictions Models Containing Unknown Functions*, Econometrica, **71**, 1795-1843.

Basu, D. (1955), *On Statistics Independent of a Sufficient Statistic*, Sankhya, **15**, 377-380.

Blundell, R., X. Chen and D. Kristensen (2007), *Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves*, Econometrica, **75**, 1613-1669.

Blundell, R. and J. Horowitz (2007), *A Non Parametric Test of Exogeneity*, Review of Economic Studies, **74**, 1035-1058.

Carrasco, M. and J.P. Florens (2000), *Generalization of GMM to a Continuum of Moment Conditions,* Econometric Theory, **16**, 797-834.

Carrasco, M. and J.P. Florens (2009), *Spectral Method for Deconvolving a Density,* forthcoming, Econometric Theory.

Carrasco, M., J.P. Florens and E. Renault (2007), *Estimation Based on Spectral Decomposition and Regularization*, Handbook of Econometrics, J.J. Heckman and E. Leamer, eds, **Vol. 6**, Elsevier, North Holland.

Chen, X. and M. Reiss (2007), *On Rate Optimality for Ill-posed Inverse Problems in Econometrics*, forthcoming, Econometric Theory.

Darolles, S., J.P. Florens and E. Renault (1998), *Nonlinear Principal Components and Inference on a Conditional Expectation Operator*, mimeo, CREST.

Engl, H.W., M. Hank and A. Neubauer (2000), *Regularization of Inverse Problems*, Kluwer.

Feve, F. and Florens, J. P (2009), *The Practice of Nonparametric Estimation by Solving Inverse Problem: The Example of Transformation Models*, Discussion paper.

Florens, J. P. (2000), *Inverse Problems and Structural Econometrics: The Example of Instrumental Variables*, Initial communication at the Econometric Society World Meeting (Seattle), published in Advances in Economics and Econometrics: Theory and Applications, Dewatripont, M., Hansen, L.P. and Turnovsky, S.J., eds, **2**, 284-311, Cambridge University Press.

Florens, J. P. (2005), *Endogeneity in Non Separable Models. Application to Treatment Effect Models where Outcomes are Durations*, Discussion Paper.

Florens, J.P., J. Heckman, C. Meghir and E. Vytlacil (2008), *Identification of Treatment Effects using Control Function in Model with Continuous Endogenous Treatment and Heterogenous Effects*, Econometrica, **76**, 1191-1206. .

Florens, J.P., J. Johannes, and S. Van Bellegem (2005), *Instrumental Regression in Partially Linear Models*, forthcoming, Econometric Journal.

Florens, J.P., J. Johannes, and S. Van Bellegem (2007), *Identification and Estimation by Penalization in Nonparametric Instrumental Regression*, forthcoming, Econometric Theory.

Florens, J.P. and M. Mouchart (1986), *Exhaustivité, Ancillarité et Identification en Statistique Bayésienne*, Annales d'Economie et de Statistique, **4**, 63-93.

Florens, J.P., M. Mouchart and J.M. Rolin (1990), *Elements of Bayesian Statistics*, Dekker, New York.

Florens, J.P., M. Mouchart and J.M. Rolin (1993), *Noncausality and Marginalization of Markov Process*, Econometric Theory, **9**, 241-262.

Florens, J.P. and A. Simoni (2007), *Regularized Posteriors in Linear Ill-posed Inverse Problems*, Discussion paper.

Foldes, A. and P. Revesz (1974), *A General Method for Density Estimation*, Sc. Math. Hung., **7**, 90-94.

Gagliardini, C. and O. Scaillet (2006), *Tikhonov regularisation for Functional Minimum Distance Estimators*, Discussion paper.

Gavin, J., S. Haberman, and R. Verall (1995), *Graduation by Kernel and Adaptive Kernel Methods with a Boundary Correction*, Transaction of Society of Actuaries, **47**, p. 173-209.

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

Hall, P. and J. Horowitz (2005), *Nonparametric Methods for Inference in the Presence of Instrumental Variables*, Annals of Statistics, **33**, 2904-2929.

Hall, P. and J. Horowitz (2007), *Methodology and Convergence Rates for Functional Linear Regression*, Annals of Statistics, **35**, 70-91.

Hansen, B. E. (2008), *Uniform convergence rates for kernel estimation with dependent data*, Econometric Theory, **24**, 726-748.

Härdle, W., and O. Linton (1994), *Applied Nonparametric Methods,* Handbook of Econometrics, **Vol. 4**, 2295-2339.

Horowitz, J. (2006), *Testing a Parametric Model Against a Nonparametric Alternative with Identification through Instrumental Variables*, Econometrica, **74**, 521-538.

Horowitz, J. and S. Lee (2007), *Non parametric Instrumental Variables Estimation of a Quantile Regression Model*, Econometrica, **75**, 1191-1208.

Imbens, G. and J. Angrist (1994), *Identification and Estimation of Local Average Treatment Effects*, Econometrica, **62**, 467-476.

Imbens, G. and W. Newey (2009), *Identification and Inference in Triangular Simultaneous Equations Models without Additivity*, forthcoming, Econometrica.

Johannes, J., S. Van Bellegem and A. Vanhems (2007), *A Unified Approach to Solve Ill-posed Inverse Problem in Econometrics*, forthcoming, Econometric Theory.

Jones, M.C. (1993), *Simple Boundary Correction for Kernel Density Estimation*, Statistics and Computing, 135-146.

Kress, R. (1999), *Linear Integral Equations,* Springer.

Lancaster, H. (1968), *The Structure of Bivariate Distributions,* Ann. Math. Statist., **29**, 719-736.

Lehmann, E.L. and H. Scheffe (1950), *Completeness Similar Regions and Unbiased Tests Part I*, Sankhya, **10**, 305-340.

Luenberger (1969), *Optimization By Vector Space Methods*, Wiley.

Morozov V.A. (1993), *Regularization Methods for Ill-posed Problems*, CRC Press.

Nashed, M.Z. and G. Wahba (1974), *Generalized Inverse in Reproducing Kernel Spaces: an Approach to Regularization of Linear Operator Equations*, SIAM Journal of Mathematical Analysis, Vol **5**, 974-987.

Newey, W. (1997), *Convergence Rates and Asymptotic Normality for Series Estimators,* Journal of Econometrics, **79**, 147-168.

Newey, W., F. Hsieh and M. Robins (2004), *Twicing Kernels and Small Bias Property of Semi Parametric Estimators*, Econometrica, **72**, 947-962.

Newey, W. and J. Powell (2003), *Instrumental Variables Estimation of Nonparametric Models,* Econometrica, **71**, 1565-1578.

Newey, W., J. Powell and F. Vella (1999), *Nonparametric Estimation of Triangular Simultaneous Equations Models*, Econometrica, **67**, 565-604.

Muller, H.-G. (1991), *Smooth optimum kernel estimators near endpoints*, Biometrika, **78**, 521-530.

Rice, J.A. (1984), *Boundary Modification for Kernel Regression*, Communications in Statistics. Theory and Methods, **13**, 893-900.

Rothe, C. (2009), *Nonparametric estimation of distributional policy effects*, forthcoming in Journal of Econometrics.

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley.

Tikhonov, A. and V. Arsenin (1977), *Solutions of Ill-posed Problems*, Winston & Sons, Washington D.C.

Van der Vaart, A.W. and J.A. Wellner (1996), *Weak Convergence and Empirical Processes, Springer*, New York.

Vapnik, A.C.M. (1998), *Statistical Learning Theory,* Wiley, New York.

Wahba, G. (1973), *Convergence Rates of Certain Approximate Solutions of Fredholm Integral Equations of the First Kind*, Journal of Approximation Theory, **7**, 167-185.

Walter, G. and J. Blum (1979), *Probability Density Estimation using Delta Sequences*, Annals of Statistics, **7**, 328-340.

Figure 1: Numerical Implementation