Econometrica, Vol. 79, No. 5 (September, 2011), 1541–1565

NONPARAMETRIC INSTRUMENTAL REGRESSION

BY S. DAROLLES, Y. FAN, J. P. FLORENS, AND E. RENAULT[1]

The focus of this paper is the nonparametric estimation of an instrumental regression function φ defined by conditional moment restrictions that stem from a structural econometric model $E[Y - \varphi(Z) \mid W] = 0$, and involve 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 analyzes 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.

1. INTRODUCTION

AN ECONOMIC RELATIONSHIP between a response variable Y and a vector Z of explanatory variables is often represented by the equation

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

where the function φ should define the relationship of interest while U is an error term.[2] 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 such as simultaneous equations, error-in-variables models, and 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

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

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

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

 DOI: 10.3982/ECTA6539

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 to need a regularization method. The connection between instrumental variables estimation and ill-posed inverse problems was pointed out by Florens (2003) and Newey and Powell (2003). Florens (2003) proposed to address this question using a Tikhonov regularization approach, which is also used in Carrasco and Florens (2000) to treat generalized method of moments (GMM) estimation with an infinite number of moment conditions. The Tikhonov approach was also adopted by Hall and Horowitz (2005). Newey and Powell (2003) resorted to a different analysis based on sieve estimation under regularization by compactness.

The literature on ill-posed inverse problems is huge, particularly in numerical analysis and image processing. The deconvolution problem is one of the main uses of inverse problems in statistics (see Carrasco and Florens (2011)). The main features of the instrumental variables estimation come from the necessity of the estimation of the equation itself (and not only the right hand side) and from the combination of parametric and nonparametric rates of convergence. The theory of inverse problems introduces to 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 though 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. Proofs and other details of mathematical rigor are provided in two appendices. Appendix A contains the proofs of the theorems while Appendix B, provided in the Supplemental Material (Darolles, Fan, Florens, and Renault (2011)) shows that the maintained assumptions can be derived from more primitive conditions on the Data Generating Process (DGP).

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

2. THE INSTRUMENTAL REGRESSION AND ITS IDENTIFICATION

2.1. *Definitions*

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 (c.d.f.) F. We assume that Y, the first coordinate of S, is square integrable. This condition is actually a condition on F; $\mathcal{F}$ denotes the set of all c.d.f.s that satisfy 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 $L^2_F(Y)$, $L^2_F(Z)$, and $L^2_F(W)$ the subspaces of L^2_F of real valued functions that depend 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 are only introduced when necessary.

DEFINITION 2.1: We call instrumental regression any function $\varphi \in L^2_F(Z)$ that satisfies the condition

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

Equivalently, φ corresponds to any solution of the 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 to identify φ uniquely 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 $\Sigma_{ZW} = \operatorname{cov}(Z, W)$ and $\Sigma_{YW} = \operatorname{cov}(Y, W)$. If Z and W have the same dimension and if Σ_{ZW} is nonsingular, then $A = \Sigma_{YW}\Sigma_{ZW}^{-1}$ and $b = \mu_Y - A\mu_Z$. We 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 the existence and uniqueness of φ.

It is useful to introduce the following two notations:

(i) $T : L^2_F(Z) \to L^2_F(W)$, $\varphi \to T\varphi = E[\varphi(Z) \mid W]$.

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

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; the reciprocal also holds. 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 that is dependent 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 c.d.f. F is characterized by its density $f(y, z, w)$ with respect to (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 $r(w) = \int y f_{Y,W}(y, w)\, dy / f_W(w)$.

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, one contribution of this paper is 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 c.d.f. 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 it corresponds to the notion of a complete statistic[3] (see Lehman and Scheffe (1950) and Basu (1955)). A systematic study is made in Florens and Mouchart (1986) and Florens, Mouchart, and Rolin (1990, Chap. 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 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 that deals 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 nonsingular, 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 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 (2010).

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

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 assuming that T and T^* are Hilbert–Schmidt operators, and is a sufficient condition of the compactness of T, T^*, TT^*, and T^*T, (see Lancaster (1958), and Darolles, Florens, and Renault (1998)). Therefore, there exists a singular values decomposition (see Kress (1999, Sec. 15.4)), that is, a sequence of nonnegative 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:

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) $\varphi_0 = 1$ and $\psi_0 = 1$.

(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

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

and

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

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 a corollary:

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 nonnull 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 nonsingular.[4] But if Σ_{WZ} is of full-column rank, 0 is not an eigenvalue of T^*T.

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

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 that of T^*T by inverting the λ_is.

Note that in these Fourier decompositions, the sequence of eigenvalues, albeit all positive, decreases fast to zero due to the Hilbert–Schmidt property. It should be stressed that the compactness and the Hilbert–Schmidt assumptions are *not* 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 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 instrumental variable (IV) estimation.

We conclude this section with a result that illustrates 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* $W = (W_1, W_2) \in R^{q_1} \times R^{q_2} (q_1 + q_2 = q)$. *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 by T_1^* *its dual. Then* T_1 *is still a 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 nondecreasing 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

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

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, that is, the $N(0, \frac{1-(\rho_1^4+\rho_2^4)}{1-(\rho_1^2+\rho_2^2)})$. 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 $L_F^2(Z)$. Then $r = T\varphi$ with*

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

Although Proposition 3.1 ensures the existence of the solution φ of the inverse problem (3.1), this problem is said to be ill-posed because a noisy measurement of r—$r + \delta\psi_i$, say (with δ arbitrarily small)—leads 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.

Although in the 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 noncontinuity.

A variety of regularization schemes are available in the literature[5] (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

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

or, equivalently,

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

[5]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 (2003), Carrasco, Florens, and Renault (2007), and references therein.

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 the (noncontinuous) generalized inverse. Then α is chosen to be 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

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

Note that the regularization bias is

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

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

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

It is then clear that the following statements hold:

(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 that implies $r \in \Psi_1^F$.
(iii) $r \in \Psi_\beta^F$, $\beta > 1 \Longrightarrow \varphi \in \Phi_{\beta-1}^F$.
(iv) $\Phi_\beta^F = \mathcal{R}[(T^*T)^{\beta/2}]$ and $\Psi_\beta^F = \mathcal{R}[(TT^*)^{\beta/2}]$.[6]

The condition $\varphi \in \Phi_\beta^F$ is called a *source condition* (see, e.g., Engl, Hanke, and Neubauer (2000)). It involves both the properties of the solution φ (through its Fourier coefficients $\langle\varphi,\varphi_i\rangle$) and the properties of the conditional expectation operator T (through its singular values λ_i). As an example, Hall and Horowitz (2005) assumed $\langle\varphi,\varphi_i\rangle \sim \frac{1}{i^a}$ and $\lambda_i \sim \frac{1}{i^b}$. Then $\varphi \in \Phi_\beta^F$ if $\beta < \frac{1}{b}(a - \frac{1}{2})$. 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

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

here we maintain 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 $\langle \varphi, \varphi_i \rangle$) as well as the degree of ill-posedness of the inverse problem (rate of decay of singular values λ_i).[7]

It is also worth stressing that the source condition is even more restrictive when the inverse problem is severely ill-posed, because the rate of decay of the singular values λ_i of the conditional expectation operator is exponential. This is the case, in particular, when the random vectors Z and W are jointly normally distributed (see Example 2.3). In this case, the source condition requires that the Fourier coefficients $\langle \varphi, \varphi_i \rangle$ (in the basis of Hermite polynomials φ_i) of the unknown function φ are themselves going to zero at an exponential rate. In other words, we need to assume that the solution φ is extremely well approximated by a polynomial expansion. If, on the contrary, these Fourier coefficients have a rate of decay that is only like $1/i^a$ for some $a > 0$, our maintained source condition is no longer fulfilled in the Gaussian case and we would no longer obtain the polynomial rates of convergence documented in Section 4.2. Instead, we would get rates of convergence that are polynomial in $\log n$ as in super-smooth deconvolution problems (see, e.g., Johannes, Van Bellegem, and Vanhems (2011)).

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, and 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 is the only scheme 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 us to increase its qualification. Let us consider the sequence of iterated regularization schemes

$$\varphi_{(1)}^\alpha = (\alpha I + T^*T)^{-1} T^*T\varphi,$$
$$\varphi_{(k)}^\alpha = (\alpha I + T^*T)^{-1}\big[T^*T\varphi + \alpha\varphi_{(k-1)}^\alpha\big],$$
$$\vdots$$

[7]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 (2011), and Johannes, Van Bellegem, and Vanhems (2011)).

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

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

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, and Van Bellegem (2011a)).

4. STATISTICAL INVERSE PROBLEM

4.1. *Estimation*

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 values in $[0,1]^p$ and $[0,1]^q$, respectively. 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[8] and let $K_h(\cdot,\cdot)$ denote a univariate generalized kernel function with the properties $K_h(u,t) = 0$ 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), where examples of $K_+(\cdot,\cdot)$ and $K_-(\cdot,\cdot)$ are provided.

EXAMPLE 4.1: Define

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

where $\mathrm{Lip}([a_1,a_2])$ denotes the space of Lipschitz continuous functions on $[a_1,a_2]$. Define $K_+(\cdot,\cdot)$ and $K_-(\cdot,\cdot)$ as follows:

[8]We use h and h_N interchangeably in the rest of this paper.

(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

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

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 a class 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 $f_Z(z)$[9]:

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

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

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

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

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

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

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

[9]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.

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 *not*, 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 $\alpha I + \hat{T}^*\hat{T}$. However, for given α and consistent estimators $\hat{T}$ and $\hat{T}^*$, $\alpha I + \hat{T}^*\hat{T}$ is invertible for N sufficiently large. The estimator of φ is then obtained by plugging consistent estimators of T^*, T, and r in the solution (3.2) of the minimization (3.3).

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

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 explained in Section 5 below.

4.2. *Consistency and Rate of Convergence*

Estimation of the instrumental regression as defined in Section 4.1 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: $\|\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi\|^2 = O_P(\frac{1}{N} + h_N^{2\rho})$.

Assumption A.4 is not about estimation of $r = E[Y \mid W]$, but rather 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* (i) $\alpha_N \to 0$ *with* $N\alpha_N^2 \to \infty$, (ii) $h_N \to 0$ *with* $Nh_N^{p+q} \to \infty$, $Nh_N^{2\rho} \to c < \infty$, *and* (iii) $\beta \geq 1$ *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 greater than $\frac{p+q}{2}$. This constant is minimally 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, (i.e., 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 compared with 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 us to choose

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

The two terms are then equivalent,

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

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 $h_N^{2\rho} = O(\frac{1}{N})$. 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 $\alpha_N \propto N^{-1/((\beta\wedge 2)+2)}$, it takes

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

which simply reinforces the constraint[10] $Nh_N^{p+q} \to \infty$. Since we maintain the assumption $h_N^{2\rho} = O(\frac{1}{N})$, it simply takes

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

To summarize, we have proved the following corollary.

COROLLARY 4.2: *Under Assumptions A.1–A.4, take one of the following conditions to be fulfilled*:

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

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

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

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

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

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

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[11]

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

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

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

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

and

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

with

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

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

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

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

Equation (5.1) is then equivalent to

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

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

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 $a_l(\varphi)$, $l=1,\dots,n$

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

or

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

with

$$\overline{a}=(a_l(\varphi))_l,\quad \overline{y}=(y_n)_n,$$
$$E=\big(b_m(A_n(w))\big)_{n,m},\quad F=\big(a_l(B_m(z))\big)_{l,m}.$$

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 approximations $a_l(\varphi)\simeq\varphi(z_l)$ and $b_l(\psi)\simeq\psi(w_l)$. Equation (5.2) is then a linear system where the unknowns are the $\varphi(z_n)$, $n=1,\dots,N$.

To illustrate the power of our approach and its simplicity, we present the following simulated example. The data generating process is

$$Y=\varphi(Z)+U,$$
$$Z=0.1W_1+0.1W_2+V,$$

where

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

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 0.45 (kernel on the Z variable) or 0.9 (kernel on the W variable) for $\varphi(Z)=Z^2$, and 0.45 and 0.25

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.[12] For $\varphi(Z) = Z^2$, we present in Figure 1, 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. 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 the same case.

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 that extends the discrepancy principle of Morozov (1993).

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

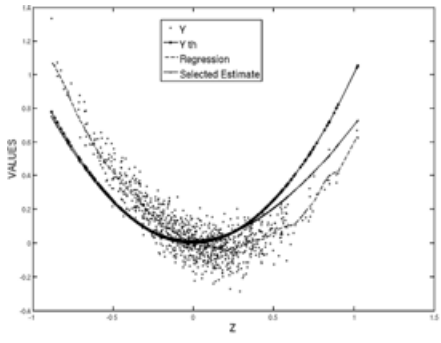
$$\varepsilon^{\alpha_N} = \hat{T}^*\hat{r} - \hat{T}^*\hat{T}\hat{\varphi}_{(2)}^{\alpha_N},$$

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

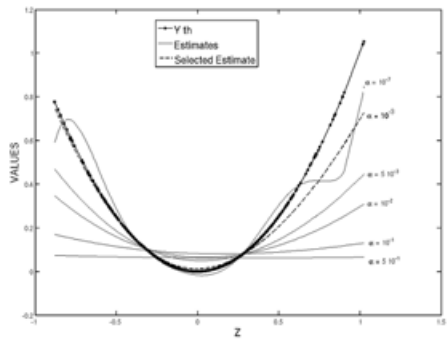
$$\|\varepsilon^{\alpha_N}\| \leq \|\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi\| + \left\|\hat{T}^*\hat{T}\varphi - \hat{T}^*\hat{T}\hat{\varphi}_{(2)}^{\alpha}\right\|.$$

To simplify the exposition, let us 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 $\|\hat{T}^*\hat{T}(\hat{\varphi}_{(2)}^{\alpha_N} - \varphi)\|^2 = \|(\hat{T}^*\hat{T}\varphi)_{(2)}^{\alpha_N} - \hat{T}^*\hat{T}\varphi\|^2 = O_P(\alpha_N)$. This last property requires a regularization method of qualification at least 4 to

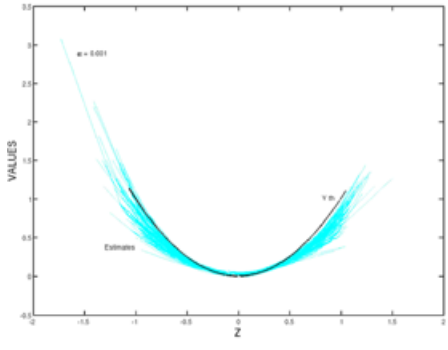
[12]Note that for simplicity we have not used generalized kernels in the simulation.



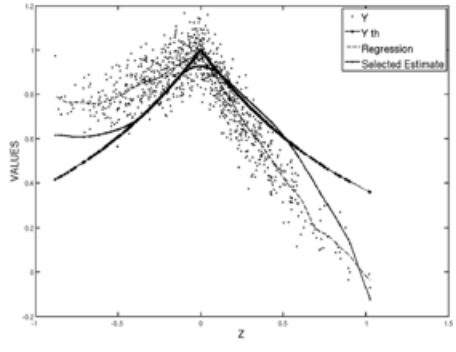
Graph 1



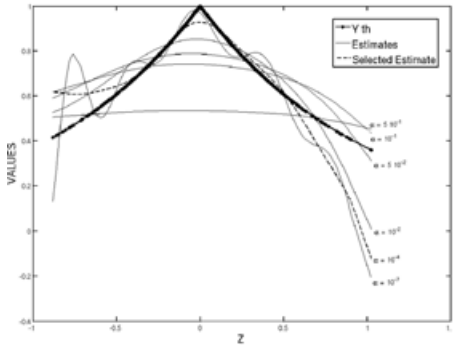
Graph 2



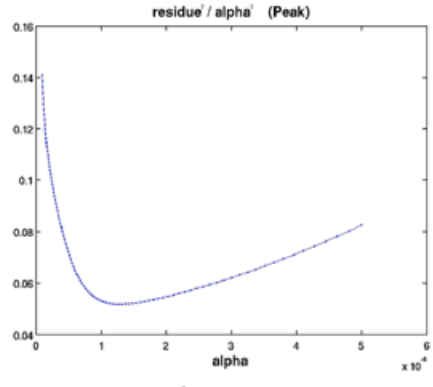
Graph 3



Graph 4



Graph 5



Graph 6

FIGURE 1.—Numerical implementation.

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

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

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

6. CONCLUSION

This paper considers the nonparametric estimation of a regression function in the presence of a simultaneity problem. We establish a set of general sufficient conditions to ensure consistency 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.

- A first extension is to analyze the case where the explanatory variables Z contain exogenous variables that are 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 semiparametric case, the procedure is described in Feve and Florens (2010).
- The treatment of semiparametric models (additive, partially linear, index models, ...) (see Florens, Johannes, and Van Bellegem (2011b) and Ai and Chen (2003)) or nonparametric models with constraints is helpful to reduce the curse of dimensionality.
- We need to improve and to study more deeply the adaptive selection of the bandwidths and of the regularization parameter.
- The structure L^2 of the spaces may be modified. In particular, Sobolev spaces can be used and the penalty norm can incorporate the derivatives (see Gagliardini and Scaillet (2006) and Blundell, Chen, and Christensen (2007)). This approach is naturally extended in terms of Hilbert scales (see Florens, Johannes, and Van Bellegem (2011a)).
- Separable models can be extended to nonseparable models or, more generally, to nonlinear problems (duration models, auctions, GMM, dynamic models) (see Ai and Chen (2003)).

- A Bayesian approach to the nonparametric instrumental variables estimation (Florens and Simoni (2011)) enhances the use of a Gaussian process prior similar to machine learning.
- In a preliminary version of this paper, we gave a proof of the asymptotic normality of $\langle \widehat{\varphi}^{\alpha_N} - \varphi, \delta \rangle$. This result is now presented in a separate paper.

APPENDIX A: PROOFS OF MAIN RESULTS

PROOF OF PROPOSITION 2.1: (i) $\Longleftrightarrow$ (ii) Part (ii) implies (i). Conversely, let us consider φ such that

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

Then

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

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, and 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 $\mathcal{N}(T) = \{0\}$. *Q.E.D.*

PROOF OF THEOREM 2.1: Let us first remark that

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

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

The eigenvalues may be compared pairwise. Using the Courant theorem (see Kress (1999, Chap. 15, Theorem 15.14, p. 276)), we get

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

PROOF OF THEOREM 4.1: The proof of Theorem 4.1 is based on 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,

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

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

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

with

$$B_1 = \alpha_N(\alpha_N I + \hat{T}^*\hat{T})^{-1}\hat{T}^*(\hat{T} - T)(\alpha_N I + T^*T)^{-1}\varphi,$$
$$B_2 = \alpha_N(\alpha_N I + \hat{T}^*\hat{T})^{-1}(\hat{T}^* - T^*)T(\alpha_N I + T^*T)^{-1}\varphi.$$

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,

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

while

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

Therefore

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

Q.E.D.

REFERENCES

AI, C., AND X. CHEN (2003): "Efficient Estimation of Conditional Moment Restrictions Models Containing Unknown Functions," *Econometrica*, 71, 1795–1843. [1553,1560]

BASU, D. (1955): "On Statistics Independent of a Sufficient Statistic," *Sankhya*, 15, 377–380. [1545]

BLUNDELL, R., X. CHEN, AND D. KRISTENSEN (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669. [1545,1560]

CARRASCO, M., AND J. P. FLORENS (2000): "Generalization of GMM to a Continuum of Moment Conditions," *Econometric Theory*, 16, 797–834. [1542,1548]

——— (2011): "A Spectral Method for Deconvolving a Density," *Econometric Theory*, 27, 546–581. [1542]

CARRASCO, M., J. P. FLORENS, AND E. RENAULT (2007): "Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, Vol. 6B, ed. by J. Heckman and E. Leamer. Elsevier/North Holland, 5633–5751. [1542,1548,1550]

CHEN, X., AND M. REISS (2011): "On Rate Optimality for Ill-Posed Inverse Problems in Econometrics," *Econometric Theory*, 27, 497–521. [1550]

DAROLLES, S., Y. FAN, J. P. FLORENS, AND E. RENAULT (2011): "Supplement to 'Nonparametric Instrumental Regression'," *Econometrica Supplemental Material*, 79, http://www.econometricsociety.org/ecta/Supmat/6539_proofs.pdf. [1542]

DAROLLES, S., J. P. FLORENS, AND E. RENAULT (1998): "Nonlinear Principal Components and Inference on a Conditional Expectation Operator," Unpublished Manuscript, CREST. [1546]

ENGL, H. W., M. HANKE, AND A. NEUBAUER (2000): *Regularization of Inverse Problems*. Netherlands: Kluwer Academic. [1548-1551]

FEVE, F., AND J. P. FLORENS (2010): "The Practice of Nonparametric Estimation by Solving Inverse Problem: The Example of Transformation Models," *The Econometric Journal*, 13, 1–27. [1545,1560]

FLORENS, J. P. (2003): "Inverse Problems and Structural Econometrics: The Example of Instrumental Variables," in *Advances in Economics and Econometrics: Theory and Applications*, ed. by M. Dewatripont, L. P. Hansen and S. J. Turnovsky. Cambridge University Press, 284–311. [1542,1548]

——— (2005): "Endogeneity in Non Separable Models. Application to Treatment Effect Models Where Outcomes Are Durations," Unpublished Manuscript, Toulouse School of Economics. [1541]

FLORENS, J. P., AND M. MOUCHART (1986): "Exhaustivité, Ancillarité et Identification en Statistique Bayesienne," *Annales d'Economie et de Statistique*, 4, 63–93. [1545]

FLORENS, J. P., AND A. SIMONI (2011): "Nonparametric Estimation of an Instrumental Variables Regression: A Quasi Bayesian Approach Based on a Regularized Posterior," *Journal of Econometrics* (forthcoming). [1561]

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. [1542]

FLORENS, J. P., J. JOHANNES, AND S. VAN BELLEGEM (2011a): "Identification and Estimation by Penalization in Nonparametric Instrumental Regression," *Econometric Theory*, 27, 472–496. [1551,1560]

——— (2011b): "Instrumental Regression in Partially Linear Models," *Econometric Journal* (forthcoming). [1560]

FLORENS, J. P., M. MOUCHART, AND J. M. ROLIN (1990): *Elements of Bayesian Statistics*. New York: Dekker. [1545,1561]

——— (1993): "Noncausality and Marginalization of Markov Process," *Econometric Theory*, 9, 241–262. [1545]

GAGLIARDINI, C., AND O. SCAILLET (2006): "Tikhonov regularisation for Functional Minimum Distance Estimators," Unpublished Manuscript, Swiss Finance Institute. [1560]

GROETSCH, C. (1984): *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. London: Pitman. [1548]

HALL, P., AND J. HOROWITZ (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *The Annals of Statistics*, 33, 2904–2929. [1542,1549,1560]

HARDLE, W., AND O. LINTON (1994): "Applied Nonparametric Methods," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden. Elsevier/North Holland, 2295–2339. [1544]

HOROWITZ, J., AND S. LEE (2007): "Non parametric Instrumental Variables Estimation of a Quantile Regression Model," *Econometrica*, 75, 1191–1208. [1541]

IMBENS, G., AND W. NEWEY (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. [1541]
JOHANNES, J., S. VAN BELLEGEM, AND A. VANHEMS (2011): "Convergence Rates for Ill-Posed Inverse Problems With an Unknown Operator," *Econometric Theory*, 27, 522–545. [1550]
KRESS, R. (1999): *Linear Integral Equations*. New York: Springer. [1546,1548,1549,1562]
LANCASTER, H. (1958): "The Structure of Bivariate Distributions," *Annals of Mathematical Statistics*, 29, 719–736. [1546]
LEHMANN, E. L., AND H. SCHEFFE (1950): "Completeness Similar Regions and Unbiased Tests Part I," *Sankhya*, 10, 305–340. [1545]
LUENBERGER, D. G. (1969): *Optimization By Vector Space Methods*. Wiley. [1561]
MOROZOV, V. A. (1993): *Regularization Methods for Ill-Posed Problems*. FL: CRC Press. [1558]
MULLER, H.-G. (1991): "Smooth Optimum Kernel Estimators Near Endpoints," *Biometrika*, 78, 521–530. [1551]
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*, 5, 974–987. [1548]
NEWEY, W., AND J. POWELL (2003): "Instrumental Variables Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578. [1542,1545]
NEWEY, W., J. L. POWELL, AND F. VELLA (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603. [1542]
TIKHONOV, A., AND V. ARSENIN (1977): *Solutions of Ill-Posed Problems*. Washington, DC: Winston & Sons. [1548]
VAPNIK, A. C. M. (1998): *Statistical Learning Theory*. New York: Wiley. [1544]
WAHBA, G. (1973): "Convergence Rates of Certain Approximate Solutions of Fredholm Integral Equations of the First Kind," *Journal of Approximation Theory*, 7, 167–185. [1548]

DRM—Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France and Lyxor Asset Management; serge.darolles@dauphine.fr,

Dept. of Economics, Vanderbilt University, VU Station B 351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, U.S.A.; yanqin.fan@vanderbilt.edu,

Toulouse School of Economics, Université Toulouse 1 Capitole, Manufacture des Tabacs, 21 Allée de Brienne, 31000 Toulouse, France; florens@cict.fr,

and

Dept. of Economics, P.O. Box B, Brown University, Providence, RI 02912, U.S.A. and CENTER, Tilburg; Eric_Renault@brown.edu.

Manuscript received June, 2002; final revision received December, 2010.