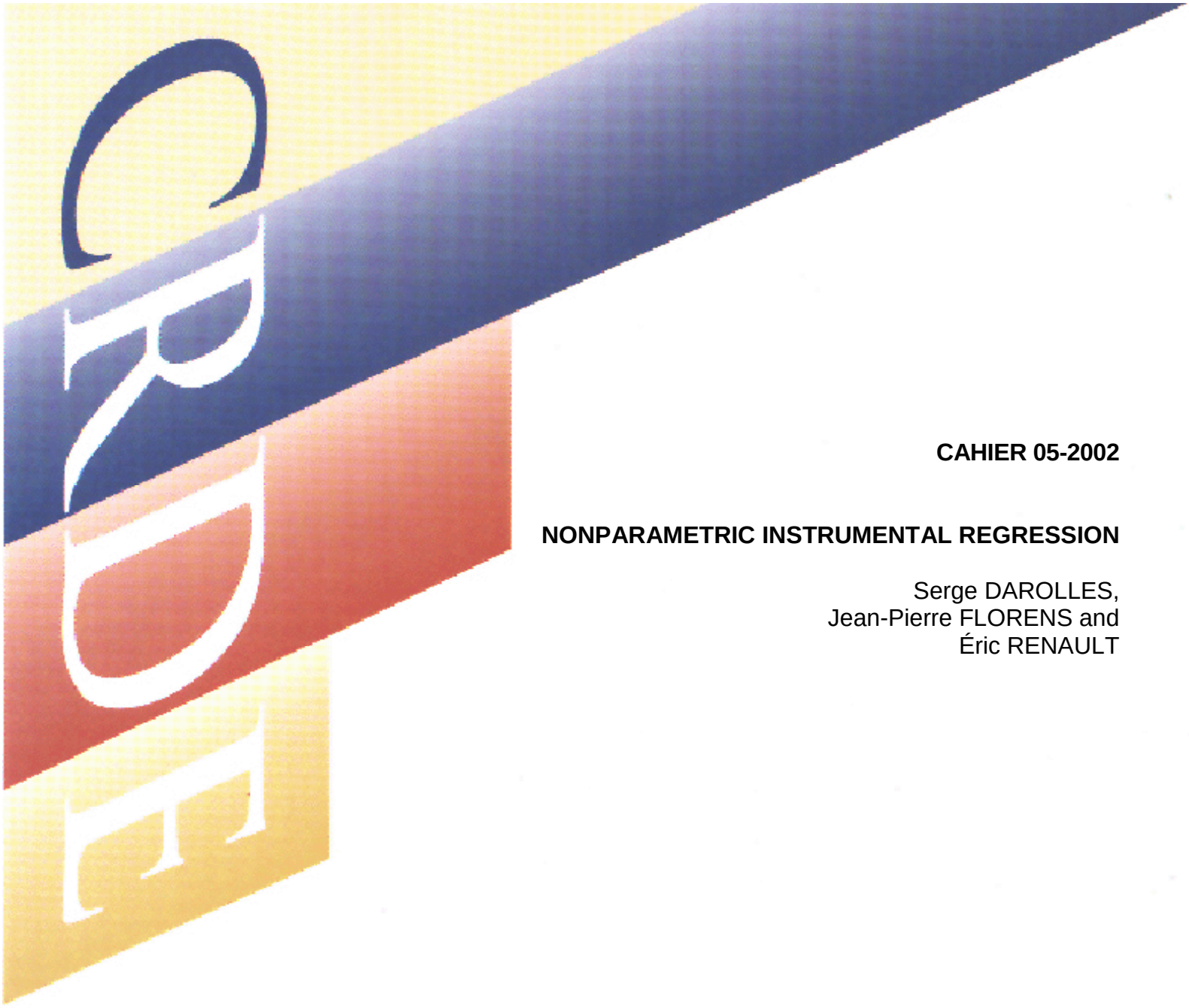

CAHIER 05-2002

NONPARAMETRIC INSTRUMENTAL REGRESSION

Serge DAROLLES,
Jean-Pierre FLORENS and
Éric RENAULT

**Centre de recherche
et développement en économique**

C.P. 6128, succursale Centre-ville
Montréal QC H3C 3J7

Téléphone : (514) 343-6557
Télécopieur : (514) 343-5831
crde@crde.umontreal.ca
http://www.crde.umontreal.ca/

Université de Montréal

CAHIER 05-2002

NONPARAMETRIC INSTRUMENTAL REGRESSION

Serge DAROLLES[1], Jean-Pierre FLORENS[2] and Éric RENAULT[3]

[1] Société Générale Asset Management, Hedge Funds Quantitative Research and CREST

[3] GREMAQ and IDEI, Université de Toulouse

[3] Centre de recherche et développement en économique (C.R.D.E.), CIRANO and Département de sciences économiques, Université de Montréal, and CREST-INSEE

May 2002

The authors would first like to thank their coauthors on papers strongly related with this one : M. Carrasco, C. Gouriéroux, J. Heckman, C. Meghir and E. Vytlacil. They also acknowledge helpful comments from the editor, R. Blundell, four referees and D. Bosq, X. Chen, L. Hansen, P. Lavergne, J.M. Loubes, W. Newey, J.M. Rolin and A. Vanhems. The authors 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.

RÉSUMÉ

Nous nous intéressons à l'estimation non paramétrique 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 $E\left[Y-\varphi(Z)|W\right]=0,$ 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, équation intégrale, problème mal posé, régularisation de Tikhonov, lissage par noyau

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 : $E\left[Y-\varphi(Z)|W\right]=0,$ 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.

Key words : 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 an equation:

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

where the function $\varphi(.)$ 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 if 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 the case in various situations: simultaneous equations, error-in-variables models, treatment model with endogenous selection.

The objective of this paper is to analyze the endogeneity problem of Z in a more general way than in these specific models and to avoid any parametric restriction on the φ function.

The first question is to add assumptions to equation (1.1) in order to characterize φ. Two general strategies exist in the literature, at least for linear models. The first one consists to introduce some hypothesis on the joint distribution of U and Z (for example on the variance matrix). The second one increases the vector of observables from (Y, Z) to (Y, Z, W), where W designates instrumental variables. The first approach was essentially followed in the error-in-variables models and some similarities exist with the instrumental model analysis (see e.g. Malinvaud (1970) , Florens, Mouchart, Richard (1974 and 1987) for the linear case). Instrumental variable analysis was proposed by Reiersol (1941), Reiersoll (1945) and extended by Theil (1953), Basmann (1957) and Sargan (1958).

This paper considers an instrumental variables treatment of the endogeneity. However, even in the instrumental variables framework, definition of functional parameter of interest remains ambiguous in the general nonlinear case. Three possible definitions of φ have been proposed[2]:

i) The first one replaces $E[U \mid Z] = 0$ by $E[U \mid W] = 0$, or equivalently it defines φ as solution of:

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

This definition was the foundation of the analysis of simultaneity in linear models or parametric nonlinear models (see Amemiya (1974)), but

[1] We remain true to the tradition in Econometrics of additive error terms. See e.g. Imbens and Newey (2001) for alternative structural approaches.

[2] A general comparison between these three concepts and their extensions to more general treatment models is done in (Florens, Heckman, Meghir, Vytlacil (2001)).

its extension to the nonparametric case comes up against difficulties. This paper treats this problem in the framework of ill-posed inverse problems (see for previous tentative Newey, Powell (2000), quoted in Pagan, Ullah (1999));

ii) A second approach is now called *control function approach* and was systematized by Newey, Powell et Vella (1999)). This technic was previously developed in specific models (e.g. Mills ratio correction in some selection models for example). The starting point is to compute $E[Y \mid Z, W]$ which satisfies:

$$E[Y \mid Z, W] = \varphi(Z) + h(Z, W), \tag{1.3}$$

where $h(Z, W) = E[U \mid Z, W]$. Equation (1.3) does not characterize φ. However we can assume that there exist a function V (the *control function*) of (Z, W) (typically $Z - E[Z \mid W]$) which captures all the endogeneity of Z in the sense: $E[U \mid W, V] = E[U \mid V]$. This implies that (1.3) may be rewritten in:

$$E[Y \mid Z, W] = \varphi(Z) + h(V), \tag{1.4}$$

and, under some conditions, φ may be identified from (1.4), up to an additive constant term.

iii) A third definition follows from the literature on treatment model (see e.g. Imbens, Angrist (1994), Heckman, Ichimura, Smith, Todd (1998) and Heckman, Vytlacil (1999)). We simplify extremely this analysis by considering Z and W as scalar. *Local instrument* is defined by $\frac{\partial E[Y|W]}{\partial W} / \frac{\partial E[Z|W]}{\partial W}$, and the function of interest φ is assumed to be characterized by the relation:

$$\frac{\frac{\partial E[Y|W]}{\partial W}}{\frac{\partial E[Z|W]}{\partial W}} = E\left[\frac{\partial \varphi}{\partial Z} \mid W\right]. \tag{1.5}$$

These three concepts are identical in the linear normal case but differ in general, as it is shown in the two following examples.

Example 1.1: Let us consider a trivariate zero mean normal distribution (Y, Z, W). The linear function βZ where $\beta = E[YW] / E[ZW]$ satisfies the three conditions (1.2), (1.4) and (1.5), with $V(Z, W) = Z - E[Z \mid W]$. More generally, any function φ such that $E[Y - \varphi(Z)] = 0$ and W is independent of $(V, Y - \varphi(Z))$, satisfies the two conditions (1.2) and (1.4) . Then, we get:

$$\frac{\partial E[Y \mid W]}{\partial W} = \frac{\partial E[\varphi(Z) \mid W]}{\partial W} = \frac{\partial}{\partial W} \int \varphi(Z)\, p(Z - E(Z|W))\, dZ,$$

where p is the density of V. Under boundary conditions (1.5) follows by integration by part.

Example 1.2:

If a function φ^* fulfills (1.4), we get:

$$E[Y|W] = E[\varphi^*(Z)|W] + E[h(V)|W],$$

whereas, for a function φ conformable to (1.2):

$$E[Y|W] = E[\varphi(Z)|W].$$

Therefore:

$$E[\varphi(Z) - \varphi^*(Z)|W] = E[h(V)|W],$$

is not constant in general, even if $V = Z - E[Z|W]$. The difficulty comes from the fact that, besides the normal case, V is not independent of W in general and nonlinear functions $h(V)$ of V may be correlated with W. An explicit counterexample with conditional heteroscedasticity is provided in Appendix E. It is also shown that such nonlinearities will imply that neither φ nor $\varphi^*$are solution of (1.5).

The paper analyses the definition of the structural parameter implicitly derived from the functional equation (1.2). This is actually an equation of the type $A(\varphi, F) = 0$, where F is the probability distribution of (Y, Z, W). We point out the condition on F which determines uniquely the solution. Estimation of φ is obtained by solving $A(\varphi, \hat{F}_N) = 0$, where $\hat{F}_N$ is a smooth estimator of F. However this equation has no solution which depends continuously on F (ill-posed inverse problem) and it must be transformed into a regularized inverse problem. The asymptotic properties of the solution are finally given. Contrarily to most of the nonparametric asymptotic theories, we do not obtain a speed of convergence just depending on the sample size and on the bandwidth. It also depends on the distribution of the variables (through the dependance scheme between the instruments and the endogenous variables) and on the behavior of a Tikhonov regularization parameter. However we can compute lower bound of the speed of convergence and discuss optimal choices of the regularization parameters. A general concept of poor instruments and more precisely a measure of the information about the instrumental regression function provided by a given set W of instruments is proposed through the asymptotic behavior of Tikhonov regularized solutions.

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. The subvectors Z and W

may have some elements in common. 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* 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. Typically F is the true distribution function from which are generated the observations 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 $\varphi \in L^2_F(Z)$ 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 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, $A = \Sigma_{YW}\Sigma^{-1}_{ZW}$ 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.

Example 2.2: We assume that Z and W have both a discrete support $\{1, 2, ..., K\}$. In this case, conditions (2.1) amount to a system of K equations about the K possible unknown values of φ. It is a Cramer system if and only if the $K \times K$ matrix giving the conditional probability distribution of Z given W is non singular.

It will be useful to introduce the two following notations:

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

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

These two linear operators satisfy:

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

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

$$A(\varphi, F) = T_F(\varphi) - r_F = 0, \tag{2.3}$$

where $r_F(W) = E[Y \mid W]$. 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.

Remark 2.1: The spaces $L_F^2(Z)$ and $L_F^2(W)$ are defined for given probability distributions of Z and W. We may be led to change the reference probability measures to restrict φ to belong to a subset of $L_F^2(Z)$ and to allow r_F to be in a space larger than $L_F^2(W)$. In particular, this modification will be necessary in order to consider non compact supports and distributions with density not bounded from below by a strictly positive number. The main complexity introduced by this change is the modification of the dual operator T_F^*. For this reason, this extension is not considered in the paper and is just sketched in Appendix C.

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*:

$$\int \varphi(z) \frac{f(.,z,w)}{f(.,.,w)} dz = r_F(w), \tag{2.4}$$

where $r_F(w) = \int y \frac{f(y,.,w)}{f(.,.,w)} dy$.

The estimation of a function by solving an integral equation is a usual problem in nonparametric statistic. Indeed the estimation of the density function g itself of a random variable Y can be seen as the resolution of:

$$\int g(u) \mathbf{I}_{]-\infty, y[} du = G(y), \tag{2.5}$$

where the cumulative function G is replaced by its empirical counterpart. However the estimation issue of φ from (2.4) is even more difficult than the estimation of g defined by (2.5) since:

i) on the one hand, Hardle, Linton (1994) explain that (2.5) is an ill-posed inverse problem whose necessary regularization leads to a nonparametric speed of convergence of the estimator of g deduced by (2.5) from the empirical cumulative function which is a root-N consistent estimator of G.

ii) on the other hand, the inverse problem (2.4) is not only ill-posed (see Section 3 below) but its inputs for statistical estimation of φ are nonparametric estimators of the functions f and r_F, 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_F 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_F is one to one (or equivalently the null space $\mathcal{N}(T_F)$ of T_F is reduced to zero). This abstract condition on F can be related to a probabilistic point of view using the fact that T_F is a conditional expectation operator. We introduce the following definition.

Definition 2.2 *: A random vector U is strongly identifiable by a random vector V if we have $E[\psi(U) \mid V] = 0$ a.s. $\Rightarrow \psi = 0$ a.s..*

This concept is well-known in statistics and corresponds to the notion of complete statistic[3] (see Lehman, Scheffe (1950), Basu (1955). A systematic study is made in Florens, Mouchart, (1986), and Florens, Mouchart, Rolin (1990), Chapter 5 under the name of strong identification (in a L^2 sense) of the σ-field generated by the random vector U by the σ-field generated by the random vector V. Definition 2.2 implies the following obvious result:

Proposition 2.1 *: φ is identifiable if and only if Z is strongly identifiable by W.*

The characterization of identification in terms of "*completeness of the conditional distribution function of Z given W*" was already provided by Newey, Powell (2000). They also discussed the two particular cases detailed in examples 2.3 and 2.4 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 2.3: 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, Rolin (1993)):

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

i) $\mathcal{N}(\Sigma_{ZZ}) = \mathcal{N}(\Sigma_{WZ})$;
ii) $\mathcal{N}(\Sigma_{WZ}) \subset \mathcal{N}(\Sigma_{ZZ} - \Sigma_{ZW}\Sigma_{WW}^{+}\Sigma_{WZ})$;
iii) $\text{Rank}(\Sigma_{ZZ}) = \text{Rank}(\Sigma_{WZ})$.

In particular, if Σ_{ZZ} is regular, the dimension of W must be greater 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.

Example 2.4: If $Z \in \{a_1, ..., a_k\}$ and $W \in \{b_1, ..., b_l\}$ are discrete, and if P is the $l \times k$ matrix of conditional probabilities of Z given W, then strong identification is equivalent to $\text{Rank}(P) = k$.

Despite the abstract character of Proposition 2.1, this identification condition can be checked in specific models (see e.g. Ai, Blundell, Chen (2001)). It can also be interpreted in terms of operators related to T_F as shown by the following corollary[4].

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

i) φ *is identifiable;*

ii) $T_F^* T_F$ *is one to one;*

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

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. This latter case will be considered in Paragraph 2.3.

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_F and T_F^* are Hilbert Schmidt operators, and is a sufficient condition of compactness of T_F, T_F^*, $T_F T_F^*$ and $T_F^* T_F$ (see Lancaster (1968), Darolles, Florens, Renault (1998)). Therefore there is a sequence of non negative real numbers $\lambda_0 = 1 \geq \lambda_1 \geq \lambda_2$ and two sequences of functions φ_i, $i \geq 0$, and ψ_j, $j \geq 0$ such that (see Kress (1998), 15.4):

i) φ_i, $i \geq 0$, is an orthonormal sequence of $L_F^2(Z)$ (i.e. $\langle \varphi_i(Z), \varphi_j(Z) \rangle = \delta_{ij}$, $i, j \geq 0$, where δ_{ij} is the Kronecker symbol) and ψ_j, $j \geq 0$, is an orthonormal sequence of $L_F^2(W)$.

[4] All the proofs are given in Appendix A.

ii) $T_F^* T_F [\varphi_i (Z)] = \lambda_i^2 \varphi_i (Z)$, $i \geq 0$;

iii) $T_F T_F^* [\psi_i (W)] = \lambda_i^2 \psi_i (W)$, $i \geq 0$;

iv) $\varphi_0 (Z) = 1$, $\psi_0 (W) = 1$;

v) $\langle \varphi_i (Z), \psi_j (W) \rangle = \lambda_i \delta_{ij}$, $i, j \geq 0$;

vi) $\forall g \in L_F^2(Z)$, $g(z) = \sum_{i=0}^{\infty} \langle g (Z), \varphi_i (Z) \rangle \varphi_i (z) + \bar{g}$, where $\bar{g} \in \mathcal{N}(T_F)$;

vii) $\forall h \in L_F^2(W)$, $h(w) = \sum_{i=0}^{\infty} \langle h (W), \psi_i (W) \rangle \psi_i (w) + \bar{h}$, where $\bar{h} \in \mathcal{N}(T_F^*)$

Similarly we obtain the decomposition of the joint density $f(., z, w)$ of random variables Z and W from the eigenfunctions and eigenvalues:

$$f(., z, w) = f(., z, .) f(., ., w) [1 + \sum_{i=1}^{\infty} \lambda_i \varphi_i (z) \psi_i (w)]. \tag{2.6}$$

Actually, it can even be shown (see Kress (1998), 15.4) that for all i:

$$\begin{aligned} T_F \varphi_i &= \lambda_i \psi_i, \\ T_T^* \psi_i &= \lambda_i \varphi_i, \end{aligned}$$

and then:

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

and:

$$T_F^* [h (W)] (z) = E [h (W) | Z = z] = \sum_{t=0}^{\infty} \lambda_i < h (W), \varphi_i (W) > \varphi_i (z).$$

The statistical interpretation of these expansions is the following. If one considers an ordered sequence of eigenvalues $\lambda_0 \geq \lambda_1 \geq ... \geq \lambda_N$, the truncated sum $\sum_{i=0}^{N} \lambda_i \langle g (Z), \varphi_i (Z) \rangle \psi_i (w)$ is the best L^2-approximation of $E[g(Z) \mid W]$ by an affine fonction on the nonlinear functions $\psi_i (W)$ of W. In other words we are looking for the best nonlinear instruments (see the Best Nonlinear Two Stage Least Squares by Amemiya (1975). The ordering of the eigenelements is not needed for the asymptotic theory we propose in this paper but it is clearly useful for small sample performance (see Darolles, Florens and Renault (1998)).

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

Corollary 2.2 *: Under assumption A.1, φ is identifiable if and only if* 0 *is not an eigenvalue of* $T_F^* T_F$.

Note that the two operators $T_F^* T_F$ and $T_F T_F^*$ have the same non null eigenvalues. But, for example, if W and Z are jointly normal, 0 is an eigenvalue of $T_F T_F^*$ as soon as $\dim W > \dim Z$ and Σ is non singular[5]. But if Σ_{WZ} is full-column rank, 0 is not an eigenvalue of $T_F^* T_F$.

The strong identification assumption corresponds to $\lambda_i > 0$ for any i and it characterizes a strong dependence between the two random variables. In particular we can directly deduce the Fourier decomposition of the inverse of $T_F^* T_F$ from the one of $T_F^* T_F$.

2.3 The variables in common case

We now assume that Z and W become (Z, X) and (W, X) respectively, where Z, X and W have no element in common. The condition (2.1) becomes:

$$Y = \varphi(Z, X) + U, \qquad E[U \mid X, W] = 0. \tag{2.7}$$

The last condition could be extended to $E[U \mid X, W] = E[U \mid X]$ (see Florens, Heckman, Meghir, Vytlacil (2001)), but this case will not be analyzed here.

The general identification condition given in Proposition 2.1 remains true if it is stated conditionally to the exogenous variables X (see Florens, Mouchart, Rolin (1990), 5):

Proposition 2.2 *: φ is identifiable if and only if Z is conditionally strongly identifiable by W, given X, that is if:*

$$E[\psi(Z, X) \mid X, W] = 0 \ a.s. \ \Rightarrow \psi = 0 \ a.s.$$

Unfortunately, the methodology put forward in this paper cannot be fully extended to this general case since the variables in common X prevent the conditional expectation operator of (Z, X) given (X, W) to be an Hilbert Schmidt operator (Assumption A.1 is no more fulfilled).

However, we are going to be able to extend our methodology to some separable cases of the following form:

$$\varphi(Z, X) = \alpha(X) + \sum_{\ell=1}^{L} \beta_\ell(X) \gamma_\ell(Z_\ell), \tag{2.8}$$

where Z_ℓ, $\ell = 1, ..., L$, are subvectors of Z with no elements in common. Note that (2.8) is quite general since it encompasses in particular the class of additive models ($L = 1$, $\beta_1(X) = 1$) which are widely used for dealing

[5] In this case $a'\Sigma_{WZ} = 0 \Longrightarrow T_F^*(a'W) = 0$.

with the curse of dimensionality. The case of discrete explanatory variables Z (treatment models, see e.g. Abadie (2001), Das (2001)) is also nested in (2.8), with Z_ℓ, $\ell = 1, ..., L$ being a collection of binary variables.

The basic idea of the extension of our inference methodology to the general setting (2.8) is the cancellation of the role of the variables in common X by the following regression equation:

$$Y - E[Y \mid X] = \sum_{\ell=1}^{L} \beta_\ell (X) (\gamma_\ell(Z_\ell) - E[\gamma_\ell(Z_\ell) \mid X]) + U. \tag{2.9}$$

Therefore, the estimation of such additive multiplicative instrumental regression model will combine inversions of regularized conditional expectation operator and backfitting. The technical details of this extension are beyond the scope of this paper and only the no variables in common case will be explicitly considered in the following sections. It is nevertheless worthwhile to notice that the identification issues are easy to address in the general setting (2.8).

First, it allows us to weaken the identification assumption of proposition 2.2.

Definition 2.3 *Z is conditionally linearly identifiable by W given X if for any family $(\lambda_i (X), \rho_i (Z))$ $i = 1, ..., n$ of square integrable functions:*

$$\sum_{i=1}^{n} \lambda_i (X) E[\rho_i (Z) - E(\rho_i (Z) | X) | X, W] = 0 \ a.s.$$

$$\Longrightarrow \sum_{i=1}^{n} \lambda_i (X) [\rho_i (Z) - E(\rho_i (Z) | X)] = 0 \ a.s.$$

Of course, if the random variable X is degenerate, conditional linear identification is askin to strong identification of Z by W. But in general, conditional linear identification of Z by W, given X, is a weaker assumption than conditional strong identification. To see that this condition is well-suited to the setting (1.8), let us consider without loss of generality the particular case $L = 1$:

$$\varphi (Z, X) = \alpha (X) + \beta (X) \gamma (Z), \tag{2.10}$$

where the function β is allowed to belong to a subset $\mathcal{B}$ of $L^2_F (X)$. Typically, $\mathcal{B}$ contains only the constant functions in the particular case of additive models. To deal with non constant functions β we need to extend a concept of measurable separability:

Definition 2.4

(i) *Two random vectors X and Z are measurably separable (or variation free) if any function of X a.s. equal to a function of Z is a.s. constant.*

(ii) *Two random vectors X and Z are $\mathcal{B}$ stongly measurably separable if for any b,c,d in$L_F^2(X), L_F^2(Z)$ and $L_F^2(Z)$ respectively and for any $\beta \in \mathcal{B}$, the almost sure equality:*

$$\beta(X)\, c(Z) = b(X) - d(Z),$$

implies that

- Either β and b are constant

- or c and d are constant

The first concept of measurable separability has been introduced by Florens, Mouchart and Rolin (1990). The second concept is stronger in general insofar as $\mathcal{B}$ contains some constant functions. A counterexample is provided in Appendix E to show that the two concepts are not equivalent. However, $\mathcal{B}$-strong measurable separability amounts to measurable separability when $\mathcal{B}$ contains only constant functions. Moreover, if it were possible to differentiate the identity $\beta(x)c(z) = b(x) - d(z)$ with respect to real variables x and z, it could easily be shown that measurable separability implies $\mathcal{B}$-strong measurable separability.

Conditional linear identifiability and $\mathcal{B}$-strong measurable separability are sufficient to identify the instrumental regression function (2.10) up to some normalisation conditions:

Proposition 2.3 *Let us consider the instrumental regression function defined by:*

$$\begin{aligned} \varphi(Z,X) &= \alpha(X) + \beta(X)\,\gamma(Z), \\ (\alpha,\beta,\gamma) &\in \mathcal{A}\times\mathcal{B}\times\mathcal{C}, \end{aligned}$$

if $\mathcal{A},\mathcal{B},\mathcal{C}$ satisfy:

(i) *For any β, $\beta^* \in \mathcal{B}$,*

$P[\beta(X) = 0] = 0$,

$\beta^*/\beta \in \mathcal{B}$,

$\exists a \in \mathbb{R} \quad \beta^* = c\beta \Longrightarrow c = 1.$

(ii) *For any $\gamma, \gamma^* \in \mathcal{C}$,*

$0 \notin \mathcal{C}$,

$\exists a \in \mathbb{R} \quad \gamma^* = \gamma + a \Longrightarrow a = 0.$

(iii) *When $\mathcal{C}$ contains some constant functions, $\mathcal{A}$ contains only the null function*

(iv) *X and Z are $\mathcal{B}$-strongly measurably separable,*

Then, when Z is conditionally linearly identifiable by W, given X, the function α, β and γ are identified.

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

A linear inverse problem is defined by two linear spaces G, H, and by an equation:

$$Lg = h, \tag{3.1}$$

where $g \in G$, $h \in H$, and L a linear operator from G to H. This equation must be solved in g. If there is a continuous inverse operator L^{-1}, the problem is said a *well-posed inverse problem*. If L^{-1} does not exist or is not continuous, the inverse problem is said an *ill-posed* one. Non existence of inverse means that no solution exists and non continuity implies that small perturbations on h may be transformed in large perturbations of the solution.

Ill-posed inverse problems receive a great attention in the literature (see e.g. Wahba (1973), Nashed, Wahba (1974), Tikhonov, Arsenin (1977), Groetsch (1984), Kress (1998). Recent surveys of applications of inverse problems in statistics are Van Rooij, Ruymgaart (1999) or Vapnik (1998). For econometric applications see e.g. Carrasco, Florens (2000a) and Florens (2000). In finite dimension, linear operators are continuous, but this property disappears in infinite dimension. Moreover, if L is a continuous one to one compact operator from G to G, the range of L can be equal to G only if G is finite dimensional. We will see that in general the solution of problem (2.3) does not exist (overidentification problem). The inversion problem is extended to generalized inverses which are not continuous. Then this solution is transformed into a regularized solution.

3.1 Overidentification

Equation (2.3) admits a solution if and only if the regression function r_F belongs to the range of T_F. Basically this is a property of the *cdf* F. So we introduce the subset of *cdf* satisfying it:

$$\mathcal{F}^0 = \{F \in \mathcal{F} : r_F \in \mathcal{R}(T_F) \text{ and } \mathcal{N}(T_F) = \{0\}\}.$$

If $F \in \mathcal{F}^0$ the equation (2.3) has an unique solution:

$$\varphi = T_F^{-1} r_F. \tag{3.2}$$

As already mentioned, under Assumption $A.1.$, the function φ can be computed using Fourier decomposition of any function belonging to $L_F^2(Z)$. We obtain:

$$\varphi(z) = \sum_{i=0}^{\infty} \frac{1}{\lambda_i} \langle r_F, \psi_i \rangle \varphi_i(z), \tag{3.3}$$

where $\langle r_F, \psi_i \rangle = E[r_F(W)\psi_i(W)] = E[Y\psi_i(W)]$. The assumption $F \in \mathcal{F}^0$ implies that the series (3.3) converges in L^2 sense. We introduce a well specification hypothesis.

Assumption A.2: The data generating distribution F is an element of $\mathcal{F}^0$.

However, with usual estimators $\hat{F}_N$ of F which are of finite rank, for any N, $\hat{F}_N$ does not belong to $\mathcal{F}^0$ because the null set of $T_{\hat{F}_N}$ is not reduced to zero.

3.2 Generalized inverse

We replace equation (2.3) by:

$$\varphi = \arg \min_{\lambda \in L_F^2(Z)} \|A(\lambda, F)\|^2. \tag{3.4}$$

This approach is quite usual and transforms an inversion problem in a generalized inverse problem. It is also standard in overidentified models to replace an exact condition by a minimization problem: this is the case in the *GMM* analysis. We do not discuss here the optimality of the transformation of an exact relation to a minimization problem. This question becomes difficult in the infinite dimensional case (see e.g. Carrasco, Florens (2000a)).

To ensure the existence of a solution for (3.4), we introduce the following set of *cdf*:

$$\mathcal{F}^* = \{F \in \mathcal{F} : r_F \in \mathcal{R}(T_F) + \mathcal{N}(T_F^*)\}.$$

For any F, $r_F \in \overline{\mathcal{R}(T_F)} + \mathcal{N}(T_F^*) = L_F^2(W)$ and then $\mathcal{F}^*$ may not contain distribution such that $\mathcal{R}(T_F)$ is not closed. However by definition, $\mathcal{F}^0 \subset \mathcal{F}^*$ and then the true *cdf* is in $\mathcal{F}^*$. Usual estimators of F determine operators $T_{\hat{F}_N}$ with finite dimensional range (and then close) which are also elements of $\mathcal{F}^*$. To ensure uniqueness of the solution of (3.4), we consider the Moore-Penrose generalized inverse:

Proposition 3.1 *: For any F in $\mathcal{F}^*$, there is a unique function φ (still called the instrumental function) of minimal norm, solution of the optimization problem* (3.4)*. This solution may be decomposed in:*

$$\varphi(z) = \sum_{i/\lambda_i \neq 0} \frac{1}{\lambda_i} \langle r_F, \psi_i \rangle \varphi_i(z). \tag{3.5}$$

A proof can be founded e.g. in Luenberger (1969) [6]. Actually it is easy to check that, under our identification condition, the Moore-Penrose generalized inverse amounts to solve the following equation which is implied by (2.3):

$$T_F^* T_F \varphi = T_F^* r_F. \tag{3.6}$$

Example 3.1: Let us continue Example 2.3. in the normal case with non singular variance matrix. If $\dim W > \dim Z$, a solution of $T_F\varphi = r_F$ exists only under a particular assumption on the variance matrix. If this assumption is not satisfied we can solve the minimization problem (3.4). We first look for a solution of this problem in the class of affine functions $\varphi(Z) = AZ + b$. In this class a unique solution to (3.4) is given by: $A = (\Sigma_{ZW}\Sigma_{WW}^{-1}\Sigma_{WZ})^{-1}\Sigma_{ZW}\Sigma_{WW}^{-1}\Sigma_{WY}$, and $b = \mu_Y - A\mu_Z$. Actually, we get with this function the only solution of (3.4) for the following two reasons:

i) $Az + b$ satisfies the condition (3.6);

ii) $T_F^* T_F$ is one to one since $rank\ \Sigma_{ZW}$ is equal to $\dim Z$ (this follows from Corollary 2.2).

Example 3.2: Let us consider a binary endogenous variable $Z \in \{0, 1\}$. The instrumental regression must satisfy: $\varphi(0)(1 - p(w)) + \varphi(1)p(w) = E[Y \mid W = w]$, where $p(W) = P[Z = 1 \mid W]$. The model is identified if $p(W)$ is not constant and the F such that φ exists is characterized by the property: $E[Y \mid W = w]$ is an affine function of $p(w)$. Thus, the solution of (3.6) is obviously $\varphi(z) = \varphi(0) + (\varphi(1) - \varphi(0))z$, with $\varphi(0)$ and $\varphi(1)$ characterized by the two following equations[7]:

$$\varphi(0) + (\varphi(1) - \varphi(0))E[p(W) \mid Z = z] = E[E[Y \mid W] \mid Z = z],$$

for $z = 0, 1$.

[6] An additional extension could be obtained using Picard's theorem (see e.g. Kress (1998), p. 279). If r_F is not in $\mathcal{R}(T_F) + \mathcal{N}(T_F^*)$, but in $\overline{\mathcal{R}(T_F)}$, the solution φ given in (3.5) may still be used if the serie converges in L^2 (i. e. $\sum_i \lambda_i^{-2} \langle r_F, \psi_i \rangle^2 < \infty$). This extension does not seem relevant for our analysis because the F we consider is assumed to be in $\mathcal{F}^0$ (the true distribution) or with a finite range (the estimator).

[7] This method can easily be extended to the case with exogenous variables X in common. In this case, the two unknown $\varphi(0)$ and $\varphi(1)$ are function of X. With obvious notations, the two equations becomes: $\varphi_0(x) + (\varphi_1(x) - \varphi_0(x))E[p(W) \mid Z = z, X = x] = E[E[Y \mid W = w, X = x] \mid Z = z, X = x]$, $z = 0, 1$.

3.3 Ill-posed problem regularization

Except in the particular cases where we can restrict φ to belong to a finite dimensional space (see Examples 3.1 and 3.2), the initial problem (2.3) is an ill-posed problem for a general F because T_F is not invertible. If $F \in \mathcal{F}^*$ we have defined a solution by (3.5) but the problem remains ill-posed because the solution is not continuous in r_F. For example if r_F is perturbed in $r_F + \delta\psi_i$ (with δ arbitrarily small), the perturbed φ is equal to $\varphi + \frac{\delta}{\lambda_i}\psi_i$ which can be very large because $\lambda_i \to 0$. (for details see Tikhonov, Arsenin (1977) or Kress (1998)). We need to define a regularized solution to our problem which satisfies a continuity condition[8].

A first way to regularize the solution is to truncate the sum in (3.5). As the eigenvalues are ranked in a decreasing order, we can keep only the first $k+1$ eigenvalues:

$$\varphi^k(z) = \sum_{i=0}^{k} \frac{1}{\lambda_i} \langle r_F, \psi_i \rangle \varphi_i(z). \tag{3.7}$$

We can also eliminate the eigenvalues that are smaller than a given threshold (spectral cut-off or thresholding regularization):

$$\varphi^{\lambda_s}(z) = \sum_{i/\lambda_i > \lambda_s} \frac{1}{\lambda_i} \langle r_F, \psi_i \rangle \varphi_i(z). \tag{3.8}$$

It is worth noticing that the way chosen by Newey, Powell (2000) to circumvent the problem "*by restricting the set* Θ *over which estimation is carried out to be a compact subset of a normed set of functions*" (when Θ denotes the set of possible solutions φ) might be interpreted as a type of regularization (for regularization by compactification see Tykhonov, Arsenin (1977)). In this paper we use a different regularization, called *Tikhonov regularization*. The initial problem $T_F\varphi = r_F$ is transformed in:

$$(\alpha I + T_F^* T_F)\varphi^\alpha = r_F^*, \tag{3.9}$$

where $\alpha > 0$ is a given number, and $r_F^* = T_F^* r_F$. This equation is an *integral Fredholm type II* equation which can be written (in the case of a dominated probability) as:

$$\alpha\varphi^\alpha(z) + \int \varphi^\alpha(u)\, a(u,z)\, du = \int y b(y,z) dy, \tag{3.10}$$

where:

$$a(u,z) = \int \frac{f(.,u,w)}{f(.,.,w)} \frac{f(.,z,w)}{f(.,z,.)} dw, \tag{3.11}$$

[8] This continuity condition is necessary to deduce a consistent estimator of φ from a consistent estimator of r_F.

and

$$b(y,z) = \int \frac{f(y,.,w)}{f(.,.,w)} \frac{f(.,z,w)}{f(.,z,.)} dw. \tag{3.12}$$

Under Assumption *A*.1. the solution of (3.9) can be computed using Fourier decomposition. We obtain:

$$\varphi^{\alpha}(z) = \sum_{i=0}^{\infty} \frac{\lambda_i}{\alpha + \lambda_i^2} \langle r_F, \psi_i \rangle \varphi_i(z). \tag{3.13}$$

For a fixed α, the problem (3.9) is well-posed. Indeed $(\alpha I + T_F^* T_F)^{-1}$ is bounded since $\|(\alpha I + T_F^* T_F)^{-1}\| < \frac{1}{\alpha}$, and then continuous. Moreover, when α goes to 0, φ^{α} converges in L^2 to φ (see Kress (1998), 15.5).

We can interpret the Tikhonov regularization as a penalized version of the optimization problem (3.4), i.e:

$$\varphi^{\alpha} = \arg \min_{\lambda \in L_F^2(Z)} \|A(\lambda, F)\|^2 + \alpha \|\lambda\|^2. \tag{3.14}$$

4 Statistical Inverse Problem

4.1 Estimation

The joint distribution of (Y, Z, W) is not known and is estimated from the observations of a sample of this random vector.

Assumption A.3: *The data* (y_n, z_n, w_n), $n = 1, ..., N$, *define an i.i.d sample of* (Y, Z, W).

This independence is a simplifying assumption and could be extended to weakly dependent (stationary mixing) observations.

We estimate F using a kernel smoothing of the empirical distribution. The estimator $\hat{F}_N$ is defined through its density w.r.t. the Lebesgue measure:

$$\hat{f}_N(y,z,w) = \frac{1}{N} \sum_{n=1}^{N} K_{y,h_{yN}}(y - y_n) K_{z,h_{zN}}(z - z_n) K_{w,h_{wN}}(w - w_n),$$

where K_y, K_z, K_w are respectively 1, p, and q dimensional kernels, h_{yN}, h_{zN}, h_{wN} are three bandwidths, and for example $K_{z,h_{zN}}(z - z_n) = h_{zN}^{-p} K_z((z - z_n)/h_{zN})$. In the applications, the bandwidths differ, but they are all the same speed represented in the following by the notation h_N. We associate to $\hat{F}_N$ estimated operators $T_{\hat{F}_N}$ and $T_{\hat{F}_N}^*$. These operators are not one to one and have a finite dimensional range.

In the same way F can be replaced by $\hat{F}_N$ in all the Fourier decompositions presented previously to obtain an indirect estimator of φ. In most

usual inverse problems, the right hand side of the equation $Lg = h$ is observed with errors or estimated but the operator L is perfectly known. In our problem both L and h are unknown and estimated. In other words, we are faced with the stochastic ill-posed problem as in Vapnik (1998), 7. The implications of the unknown character of L may be seen in particular in the discussion of Assumption A.5 (see Appendix B).

Definition 4.1 *: If α_N is a positive N-dependent number, we call estimated instrumental regression function the (uniquely defined) function:*

$$\hat{\varphi}_N^{\alpha_N}(z) = (\alpha_N I + T_{\hat{F}_N}^* T_{\hat{F}_N})^{-1} r_{\hat{F}_N}^*, \tag{4.1}$$

with $r_{\hat{F}_N}^* = T_{\hat{F}_N}^* r_{\hat{F}_N}$.

Equivalently the estimated instrumental regression function $\hat{\varphi}_N^{\alpha_N}$ satisfies the integral equation:

$$\alpha_N \hat{\varphi}_N^{\alpha_N}(z) + \int \hat{\varphi}_N^{\alpha_N}(u)\, \widehat{a}_N(u,z)\, du = \int y \widehat{b}_N(y,z) dy, \tag{4.2}$$

where $\widehat{a}_N(u,z)$ and $\widehat{b}_N(y,z)$ are the kernel estimators of $a(u,z)$ and $b(y,z)$ introduced in Subsection 3.3. This estimator can be computed directly as a solution of (4.2) and it reduces to a finite dimensional inverse problem. The practical implementation of this computation is detailed in Appendix D. Note that the computation of estimators of λ_i, φ_i, ψ_i, are not required and the asymptotic properties of $\hat{\varphi}_N^{\alpha_N}$ do not rest upon the asymptotic properties of the estimators of eigenvalues and eigenvectors (see Darolles, Florens, Gouriéroux (1998) and Chen, Hansen and Scheinkman (2000) for a statement of these later properties).

Estimation of the instrumental regression function requires consistent estimations of $T_F^* T_F$ and r_F^*. 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 $T_F^* T_F$ and r_F^*. We use kernel smoothing techniques to make the paper more user-friendly, but we can generalize the approach and use any other nonparametric techniques (for a sieve approach, see Chen, Shen (1998)). The main point is the speed of convergence of the norms given for kernel smoothing by Assumptions A.4 and A.5 below.

4.2 Consistency and speed of convergence

Usual nonparametric estimation is essentially focused on the estimation of a function at a particular value of the variables. In our case, the nonparametric estimates are used as elements of a functional equation which must be solved, in order to estimate the functional parameter of interest.

Consistent estimation of this function then requires that $T^*_{\hat{F}_N} T_{\hat{F}_N}\varphi$ and $r_{\hat{F}_N}$ converge globally to their limit (see e.g. Kress (1995), 15). A natural type of convergence is in quadratic mean, that is, in $L^2_F(Z)$.

Assumption A.4: *There exists* $\rho \geq 2$ *such that*[9] $\forall \lambda \in L^2_F(Z)$*:*

$$\|(T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F)\lambda\|^2 = O\left(\left(\frac{1}{N h_N^p} + h_N^{2\rho}\right)\|\lambda\|^2\right).$$

Assumption A.5: *There exists* $\rho \geq 2$ *such that:*

$$\|r^*_{\hat{F}_N} - T^*_{\hat{F}_N} T_{\hat{F}_N}\varphi\|^2 = O\left(\frac{1}{N} + h_N^{2\rho}\right).$$

We show in Appendix B that standard regularity conditions on the true F and on φ imply that Assumptions A.4 and A.5 are satisfied. Typically, in the case of kernel estimation, ρ will be the minimum between the order of the kernel and the order of differentiability of f. In other words, the rate of convergence appearing in Assumption A.5 in the same than the rate of convergence of the kernel estimations of $T^*_{\hat{F}_N}\varphi$. For simplicity, we consider the same ρ in Assumptions A.4 and A.5. This condition is satisfied if we take the minimum in case of different values of ρ.

As already announced, the curse of dimensionality is binding only with respect to the dimension ρ of Z and not with respect to the dimension q of W. Actually, we will see now that for the purpose of estimation of φ, larger is the range of T^*_F, that is richer is the set W of instruments, better it is.

Convergence property of $\hat{\varphi}^{\alpha_N}_N$ is deduced from the following decomposition:

$$\begin{aligned}\hat{\varphi}^{\alpha_N}_N - \varphi &= (\alpha_N I + T^*_{\hat{F}_N} T_{\hat{F}_N})^{-1}[r^*_{\hat{F}_N} - T^*_{\hat{F}_N} T_{\hat{F}_N}\varphi] \\ &+ [(\alpha_N I + T^*_{\hat{F}_N} T_{\hat{F}_N})^{-1} T^*_{\hat{F}_N} T_{\hat{F}_N} - (\alpha_N I + T^*_F T_F)^{-1} T^*_F T_F]\varphi \\ &+ \varphi^{\alpha_N} - \varphi,\end{aligned}$$

where φ^{α_N} is defined in (3.9). Then:

$$\begin{aligned}\hat{\varphi}^{\alpha_N}_N - \varphi &= (\alpha_N I + T^*_{\hat{F}_N} T_{\hat{F}_N})^{-1}[r^*_{\hat{F}_N} - T^*_{\hat{F}_N} T_{\hat{F}_N}\varphi] \\ &- (\alpha_N I + T^*_{\hat{F}_N} T_{\hat{F}_N})^{-1}[T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F](\varphi^{\alpha_N} - \varphi) \\ &+ \varphi^{\alpha_N} - \varphi,\end{aligned}$$

using

$$(\alpha_N I + T^*_F T_F)^{-1} T^*_F T_F = I - \alpha_N (\alpha_N I + T^*_F T_F)^{-1},$$

[9] All the $O()$ are actually relative to the true data probability distribution.

and, as a consequence:

$$\varphi^{\alpha_N} - \varphi = \alpha_N(\alpha_N I + T_F^* T_F)^{-1}\varphi. \tag{4.3}$$

The above decomposition emphasizes the three elements of the difference between $\hat{\varphi}_N^{\alpha_N}$ and φ. The first term is due to the estimation of the right hand side r_F of the equation $T_F\varphi = r_F$. The second one is due to the estimation of the operator T_F, and the last one come from the regularization.

Let us remark that $\|(\alpha I + T^*_{\hat{F}_N} T_{\hat{F}_N})^{-1}\|^2 \leq 1/\alpha_N^2$ (see e.g. Groetsch (1984)) and recall that $\alpha_N \to 0$ implies $\|\varphi^{\alpha_N} - \varphi\| \to 0$ by virtue of the identification assumption (see e.g. Kress (1998)). Finally, we obtain the following theorem.

Theorem 4.1 *: Under Assumptions A.1-A.5,*

i) $\|\hat{\varphi}_N^{\alpha_N} - \varphi\|^2$
$= O\left(\frac{1}{\alpha_N^2}\left(\frac{1}{N} + h_N^{2\rho}\right) + \frac{1}{\alpha_N^2}\left(\frac{1}{Nh_N^p} + h_N^{2\rho}\right)\|\varphi^{\alpha_N} - \varphi\|^2 + \|\varphi^{\alpha_N} - \varphi\|^2\right)$

ii) if $\alpha_N \to 0$, $h_N^{2\rho}/\alpha_N^2 \to 0, \frac{1}{\alpha_N^2 N h_N^p} \sim O(1)$, *then* $\|\hat{\varphi}_N^{\alpha_N} - \varphi\| \to 0$ *in probability as* $N \to \infty$.

A natural question concerns now the selection rule of the two regularization parameters α_N and h_N in order to optimize the speed of convergence of $\hat{\varphi}_N^{\alpha_N}$ to φ.

As shown by A.4 and A.5, h_N has to be seen as a smoothing parameter for the nonparametric regression on Z of the variables r_F and $T_F\varphi$. The convergence to zero of the regularization parameter α_N will ensure the convergence of φ_{α_N} towards the true unknown φ at a rate depending upon the richness of the set W of instruments. We choose to measure this richness directly through the speed β of convergence of φ^{α_N} towards φ:

Definition 4.2 *: For* $0 < \beta \leq 2$, Φ_β *is the set of functions* φ *of* $L_F^2(Z)$ *such that:*

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

We will say that Φ_β is the set of functions β- instrumentalizable by W. Note that is φ is β-instrumentalizable by W, it is a fortiori β'-instrumentalizable for $\beta' < \beta$.

To understand Definition 4.2, it is worthwhile to refer to the following decomposition resulting from (4.3):

$$\|\varphi^{\alpha_N} - \varphi\|^2 = \sum_{j=0}^{\infty} \frac{\alpha_N^2}{\alpha_N^2 + 2\alpha_N\lambda_j^2 + \lambda_j^4}\langle\varphi, \varphi_j\rangle^2, \tag{4.4}$$

which implies that $\|\varphi^{\alpha_N} - \varphi\|^2$ is not greater than the sum of anyone of the following two series:

$$\alpha_N^2 \sum_{j=0}^{\infty} \frac{\langle \varphi, \varphi_j \rangle^2}{\lambda_j^4}, \tag{4.5}$$

and

$$\frac{\alpha_N}{2} \sum_{j=0}^{\infty} \frac{\langle \varphi, \varphi_j \rangle^2}{\lambda_j^2}. \tag{4.6}$$

Since λ_j^2 define the j-th squared nonlinear canonical correlation between W and Z (see e.g. Lancaster (1968) and Darolles, Florens, Renault (1998)), the convergence of these series means that these correlations are sufficiently large, that is that the instruments are sufficiently rich. In the best favorable case, we have:

$$\sum_{j=0}^{\infty} \frac{\langle \varphi, \varphi_j \rangle^2}{\lambda_j^4} < +\infty,$$

and $\beta = 2$. This is of course the case in particular if the function φ is spanned by a finite set of canonical variables φ_j.This implies that any Φ_β is dense in $L_F^2(z)$. But a more general case[10] is:

$$\sum_{j=0}^{\infty} \frac{\langle \varphi, \varphi_j \rangle^2}{\lambda_j^2} < +\infty,$$

which allows us to choose $\beta = 1$. Note that this last condition is in particular fulfilled if φ belongs to the range $\mathcal{R}(T_F^*)$ of T_F^* since in this case:

$$\langle \varphi, \varphi_j \rangle^2 = \langle T_F^* \, \mu, \varphi_j \rangle^2 = \lambda_j^2 \langle \mu, \psi_j \rangle^2,$$

and

$$\sum_{j=0}^{\infty} \frac{\langle \varphi, \varphi_j \rangle^2}{\lambda_j^2} = \sum_{j=0}^{\infty} \langle \mu, \psi_j \rangle^2 = \|\mu\|^2.$$

To conclude, instrumental regression needs a sufficiently rich set W of instruments at two stages:

i) on the one hand, as shown by Corollary 2.1, the range $\mathcal{R}(T_F^*)$ that is the space of functions of the form $E[\psi(W) \mid Z]$, has to be dense in $L_F^2(Z)$ to identify the instrumental regression φ.

[10] We thank W. Newey for drawing our attention on this general case.

ii) on the other hand, if this range is equal to $L_F^2(Z)$ (as in the finite dimensional case) or at least contains the true unknown φ, then the rate β of convergence towards φ of the sequence of its Tikhonov regularizations is not smaller than one.

This last remark extends the notion of weak instruments as proposed by Nelson, Startz, Zivot (1998), Straiger and Stock (1997) and Wang and Zivot (1998). With β as measure of the richness of the instruments, we will say that weak instruments correspond to β smaller than one. Of course, for a given function φ,one will try to define a set W of instruments such that φ is β-instrumentalizable for some $\beta \geq 1$.

Generally speaking, the existence of β guarantees a minimum rate of convergence of our estimator $\hat{\varphi}_N^{\alpha_N}$ towards φ for a convenient choice of the regularization parameter h_N and α_N.

Theorem 4.2 *: Under Assumptions A1-A5, we get:*

$$N^{\frac{\beta}{2+\beta}} \|\hat{\varphi}_N^{\alpha_N} - \varphi\|^2 = O(1),$$

if $\varphi \in \phi_\beta$, $\frac{p}{2\rho} \leq \frac{\beta}{2+\beta}$ *and*

$$\alpha_N = k_1 N^{-\frac{1}{2+\beta}},$$

$$h_N = k_2 N^{-\frac{1}{2\rho}},$$

where k_1 *and* k_2 *are constant terms.*

Note that Theorem 4.2, which is a direct consequence of Theorem 4.1 *i)*, provides only a lower bound to the rate of convergence while the actual one should be greater than $\frac{\beta}{2+\beta}$. Several remarks about the underlying trade off are worth noticing:

i) The rate $\frac{\beta}{2+\beta}$ is optimal with respect to the bound[11] provided by Theorem 4.1 *i)*. The leading term of the convergence is then provided by the estimations of r_F^* and by the regularization. This speed of convergence is not influenced by the statistical uncertainty about the operator $T_F^* T_F$ that one has to invert.

ii) To get this "*optimal*" rate of convergence, we have undersmoothed the estimation of $T_F^* T_F$ by choosing $h_N = N^{-\frac{1}{2\rho}}$ instead of the larger one $h_N =$

[11] The true optimal choice of α_N and h_N, which would have considered directly $\|\hat{\varphi}_N^{\alpha_N} - \varphi\|$ and not its upper bound given by Theorem 4.1 would have depended on the behaviour of the sequences λ_j and $\langle \varphi, \varphi_j \rangle$. This is apparent from the asymptotic probability distributions derived in the next subsection.

$N^{-\frac{1}{p+2\rho}}$ which would have been optimal for the non parametric estimation of $T_F^* T_F$. The resulting higher variance of $T^*_{\hat{F}_N} T_{\hat{F}_N}$ has a negligible cost for the purpose of the estimation of φ with respect to the two other costs: statistical uncertainty about r_F^* and regularization bias. Actually, a larger h_N would have increased the bias of the estimator of r_F^* and deteriorated[12] the global rate of convergence of $\hat{\varphi}_{\alpha_N}$ towards φ.

iii) Theorem 4.2 requires a relation between the degree of smoothness ρ of the density and β which can be interpreted as the identifying power of the instruments about φ. This relation is in particular satisfied in the case $p = 1, \rho = 2$ and $\beta = 1$. If $p = 2$ and $\beta = 1$ ρ must be greater or equal to 3.

iv) Greater β is, faster is the obtained convergence of $\hat{\varphi}_N^{\alpha_N}$ towards φ. In the most favorable case (i.e. $\beta = 2$), one gets $\|\hat{\varphi}_N^{\alpha_N} - \varphi\| = O(N^{-\frac{1}{2}})$ that is a convergence twice slower than in the parametric case.

Actually, as expected, the *"optimal"* rate $\frac{\beta}{2+\beta}$ is in general smaller than the rate of convergence $\frac{2\rho}{p+2\rho}$ that one would have obtained through a standard nonparametric regression of Y on the p variables Z. The inequality $\frac{2\rho}{p+2\rho} > \frac{\beta}{2+\beta}$ is in particular guaranteed if $p < 2\rho$ (since $\beta \leq 2$).

4.3 Asymptotic probability distributions

For economic applications, one may be interested either by the unknown function $\varphi(Z)$ itself, or only by its moments, including covariances with some known functions. These moments may for instance be useful for testing economic statements about scale economies, elasticities of substitutions, and so on.

For such tests, one will only need the empirical counterparts of these moments and their asymptotic probability distribution. An important advantage of the instrumental variable approach is to allow to estimate the covariance between $\varphi(Z)$ and $\delta(Z)$ for a large class of functions. Actually our identification assumption amounts to ensure that the range $\mathcal{R}(T_F^*)$ is dense in $L_F^2(Z)$ (see Corollary 2.2) and for any δ in this range:

$$\exists \psi \in L_F^2(W),\ \delta(Z) = E[\psi(W) \mid Z],$$

and then $Cov[\varphi(Z), \delta(Z)] = Cov[\varphi(Z), E[\psi(W) \mid Z]] = Cov[\varphi(Z), \psi(W)] = Cov[E[\varphi(Z) \mid W], \psi(W)] = Cov[Y, \psi(W)]$, can be estimated with standard parametric techniques. For instance, if $E[\delta(Z)] = 0$, the empirical counter-

[12] Of course, one could also choose for the estimation of $T_F^* T_F$ a bandwidth h_N different from the one used for estimating r_F^*. But, this would not improve the global speed of convergence of φ^{α_N} to φ.

part of $Cov[Y, \psi(W)]$, i.e.:

$$\frac{1}{N}\sum_{n=1}^{N} y_n \psi(w_n),$$

is a root-N consistent estimator of $Cov[\varphi(Z), \delta(Z)]$, and:

$$\sqrt{N}\left[\frac{1}{N}\sum_{n=1}^{N} y_n \psi(w_n) - Cov[\varphi(Z), \delta(Z)]\right] \xrightarrow{d} \mathcal{N}(0, Var[Y\psi(W)]),$$

where $Var[Y\psi(W)]$ will also be estimated by its sample counterpart[13]. However in practice this analysis has very limited interest because even if δ is given, ψ is not known and must be estimated by solving the integral equation $\delta(Z) = E[\psi(W) \mid Z]$, where the conditional distribution of W given Z is also estimated.

Then the real problem of interest is to estimate $Cov[\varphi(Z), \delta(Z)]$, or $\langle \varphi, \delta \rangle$ by replacing φ by $\hat{\varphi}_N^{\alpha_N}$ to derive the asymptotic distribution of this estimation procedure.

We slightly simplify our analysis if we introduce an homoskedasticity assumption.

Assumption A.6: *The error term is homoskedastic*: $Var[U \mid W] = \sigma^2$.

Moreover, since our estimator $\hat{\varphi}_N^{\alpha_N}$ is defined as the solution of the equation:

$$(\alpha_N I + T^*_{\hat{F}_N} T_{\hat{F}_N})\varphi = T^*_{\hat{F}_N} r_{\hat{F}_N},$$

its asymptotic behavior depends upon the right hand side $T^*_{\hat{F}_N} r_{\hat{F}_N}$ of this equation. This right hand side involves two nested standard functional estimations whose study is not the focus of interest of this paper. This is the reason why we will only maintain a natural assumption about it:

Assumption A.7: *For a suitable choice of* h_N,

$$\sqrt{N}(T^*_{\hat{F}_N} r_{\hat{F}_N} - T^*_{\hat{F}_N} T_{\hat{F}_N} \varphi) \Longrightarrow \mathcal{N}(0, \sigma^2 T^*_F T_F).$$

[13] For the purpose of interpretation, it is worthwhile to relate this total variance to the underlying regression equation:

$$Var[Y\psi(W)] = Var[\varphi(Z)\psi(W)] + E[(U^2 + 2U\varphi(Z))\psi^2(W)],$$

that is

$$Var[Y\psi(W)] = Var[\varphi(Z)\psi(W)] + E[\psi^2(W)\, Var[U|W]] + 2E[\psi^2(W)\, Cov[U, \varphi(Z)|W]].$$

The previous convergence is a functional convergence in distribution in the Hilbert space $L^2_F(Z)$ (see e.g. Van der Vaart, Wellner (1996)). Appendix B shows that this assumption is satisfied under regularity conditions on the data density.

Our proof requires a lower bound condition on this density. This condition can be avoided under some technicalities which modify in particular the asymptotic variance operator of the normal distribution. This extension is considered in Appendix C.

We have simplified the asymptotic distribution by assuming a zero mean which is obtained by choosing h_N decreasing faster than its optimal value ($h_N = O(N^{-(\frac{1}{2\rho}+\varepsilon)})$, $\varepsilon > 0$).

Let us first consider the case where the regularization parameter α is kept constant. In that case the linear operators $(\alpha I + T^*_{\hat{F}_N} T_{\hat{F}_N})^{-1}$ and $(\alpha I + T^*_F T_F)^{-1}$ are bounded and, using a functional version of the Slutsky theorem (see Chen, White (1992)), it is immediately checked that:

$$\sqrt{N}(\hat{\varphi}^\alpha_N - \varphi^\alpha - b^\alpha_N) \Longrightarrow \mathcal{N}(0, \Omega), \tag{4.7}$$

where

$$b^\alpha_N = \alpha \left[(\alpha I + T^*_{\hat{F}_N} T_{\hat{F}_N})^{-1} - (\alpha I + T^*_F T_F)^{-1} \right] \varphi,$$

and

$$\Omega = \sigma^2 (\alpha I + T^*_F T_F)^{-1} T^*_F T_F (\alpha I + T^*_F T_F)^{-1}.$$

Some comments may illustrate this first result:

i) The convergence obtained in (4.7) is still a functional distributional convergence in the Hilbert space $L^2_F(Z)$, which in particular implies the convergence of scalar product $\sqrt{N}\langle \hat{\varphi}^\alpha_N - \varphi^\alpha - b^\alpha_N, \delta \rangle$ to scalar normal distribution $\mathcal{N}(0, \langle \delta, \Omega \delta \rangle)$.

ii) The convergence of $\hat{\varphi}^\alpha_N$ involves two bias terms. The first bias is $\varphi^\alpha - \varphi$. This term is due to the regularization and does not decrease if α is constant. The second one, b_N follows from the estimation error of T_F. This bias decreases to zero when N increases, but at a lower speed than $\sqrt{N}$.

iii) The asymptotic variance in (4.7) can be seen as generalization of the two stage least squares asymptotic variance. An intuitive (but not correct) interpretation of this result could be the following: if α is small, the asymptotic variance is approximately $\sigma^2 (T^*_F T_F)^{-1}$, which is the functional extension of $\sigma^2 (E(ZW') E(WW')^{-1} E(WZ'))^{-1}$.

We now analyze the case when α_N goes to zero. The functional convergence result is not preserved. But asymptotic normality of scalar product is in general still valid as stated by the following theorem.

Theorem 4.3 *: Under assumptions A1 - A7, if* $\alpha_N^2 N h_N^p \to \infty$ *and* $\alpha_N^2/h_N^{2\rho} \to \infty$*, let* $\delta \in L_F^2(Z)$ *and:*

$$s_N(\delta) = \left\| T_F \left(\alpha_N I + T_F^* T_F\right)^{-1} \delta \right\|^2$$

If:

$$\frac{\|\delta_{\alpha_N} - \delta\|}{\alpha_N \sqrt{s_N(\delta)}} = O(1) \tag{4.8}$$

and in particular if $\delta \in \Phi_1$*, then :*

$$\sqrt{\frac{N}{s_N(\delta)}} \left\langle \hat{\varphi}_N^{\alpha_N} - \varphi^{\alpha_N} - b_N^{\alpha_N}, \delta \right\rangle \Longrightarrow \mathcal{N}\left(0, \delta^2\right) \tag{4.9}$$

If $\delta \in \phi_\beta$ *with* $\beta \geq 1$*:*

$$\upsilon_N(\delta) = \frac{N}{s_N(\delta)} \geq O\left(\alpha_N^{2-\beta}.N\right) \propto N^{\frac{2\beta}{2+\beta}} \tag{4.10}$$

for $\alpha_N \propto N^{-\frac{1}{2+\beta}}$*.*

Let us first stress that $\upsilon_N(\delta)$ is a well-defined rate of convergence going to infinity since:

$$\begin{aligned} \|T_F(\alpha_N I + T_F^* T_F)^{-1}\delta\|^2 &= \sum_{j=0}^{\infty} \frac{\lambda_j^2}{(\alpha_N + \lambda_j^2)^2} \langle \varphi_j, \delta \rangle^2 \\ &\leq \sum_{j=0}^{\infty} \frac{\lambda_j^2}{2\alpha_N \lambda_j^2} \langle \varphi_j, \delta \rangle^2 = \frac{\|\delta\|^2}{2\alpha_N}, \end{aligned}$$

and therefore:

$$\upsilon_N(\delta) \geq \frac{2\alpha_N N}{\|\delta\|^2} \to \infty.$$

On the other hand, $\upsilon_N(\delta)$ cannot of course going to infinity faster than N since, for N sufficiently large, we have:

$$\begin{aligned} \|T_F(\alpha_N I + T_F^* T_F)^{-1}\delta\|^2 &\geq \sum_{j=0}^{\infty} \frac{\lambda_j^2}{(\alpha_N + 1)^2} \langle \varphi_j, \delta \rangle^2 \\ &\geq \sum_{j=0}^{\infty} \frac{\lambda_j^2}{2} \langle \varphi_j, \delta \rangle^2 = \frac{1}{2}\|T_F \delta\|^2 > 0, \end{aligned}$$

since our maintained identification assumption precisely means that $T_F\delta$ cannot be zero for a non zero function δ.

Theorem 4.3 implies that root-N consistency is obtained when $\delta \in \Phi_2$. As already stressed in the comments about Definition 4.2, Φ_2 contains any finite dimensional vectorial space spanned by the canonical variates φ_j. On the other hand, if we only know that $\delta \in \Phi_\beta$, $0 < \beta < 2$, we have:

$$\begin{aligned} s_N(\delta) &= \sum_{j=0}^{\infty} \frac{\lambda_j^2}{\left(\alpha_N + \lambda_j^2\right)^2} \left\langle \varphi_j, \delta \right\rangle^2 \\ &\leq \frac{1}{\alpha_N^2} \sum_{j=0}^{\infty} \frac{\alpha_N^2}{\left(\alpha_N + \lambda_j^2\right)^2} \left\langle \varphi_j, \delta \right\rangle^2 \\ &= \frac{\left\| \delta_{\alpha_N} - \delta \right\|^2}{\alpha_N^2} = O\left(\alpha_N^{\beta-2}\right) \end{aligned} \tag{4.11}$$

In other words, larger values of β guarantee larger rates of convergence $\nu_N(\delta) = \frac{N}{s_N(\delta)}$. The minimal condition (4.8) is implied by $\delta \in \Phi_1$ since it is the case where $\left\| \delta_{\alpha_N} - \delta \right\|^2$goes to zero at least as fast as α_N.

Let us recall that we were able to propose a root-N consistent estimator of $Cov\left[\varphi(Z), \delta(Z)\right]$ for any $\delta \in \mathcal{R}\left(T_F^*\right)$. As shown in the comments following Definition 4.2, we know in this case that $\delta \in \Phi_1$ but there is no reason to claim that $\delta \in \Phi_\beta$ for some $\beta > 1$. The root-N consistency is obtained by an estimator $\frac{1}{N}\sum_{n=1}^{N} y_n \psi(w_n)$ of $Cov\left(\varphi(Z), \delta(Z)\right)$ which considers the function ψ such that

$$\delta(Z) = \left[T_p^* \psi\right](Z) = E\left[\psi(W) | Z\right],$$

as known. The slower speed of convergence provided by Theorem 4.3 is the price to pay for Tikhonov regularization.

Moreover, all the above discussions on rates of convergence actually describe some bounds for these rates. This is the reason why the dimensions of Z and W do not explicitly appear. However, the inequality (4.10) shows that the difference between $s_N(\delta)$ and its upper bound $\frac{\left\| \delta_{\alpha_N} - \delta \right\|^2}{\alpha_N^2}$ is tightly related to the respective behavior of α_N, to the shape of the eigenvalues λ_j's and on the specific choice of δ.

The role of the dimension of the Z variables appears in several places (choice of the bandwidth, hypothesis of Theorem 4.2). The role of the dimension of W is less explicit even it is an important element of the asymptotic behavior of our estimator. Actually the values and the rate of decline of the λ_j 's depend on the dimension of W, as shown by the following result:

Proposition 4.1 *Let us assume that* $W = (W_1, W_2) \in \mathbb{R}^{q_1} \times \mathbb{R}^{q_2}$ $(q_1 + q_2 = q)$ *and denotes by* T_{F_1} *the operator*

$$\varphi \in L_Z^2 \to E(\varphi | W_1) \in L_{W_1}^2$$

and T_{F1}^* its dual. Then T_{F1} is still an Hilbert Schmidt operator and the eigenvalues of T_{F1}^* T_{F1} $\lambda_{j_1}^2$ satisfies

$$\lambda_{j_1} \leq \lambda_j$$

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

Example 4.1 *Consider the case $(Z, W_1, W_2) \in \mathbb{R}^3$ 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 $T_F^* T_F$ is characterized by*

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

and its eigenvalues λ_j^2 are $\left(\rho_1^2 + \rho_2^2\right)^j$. The eigenvalues of $T_{F1}^* T_{F1}$ are $\lambda_{j_1}^2 = \left(\rho_1^2\right)^j$ and the eigenvectors are the Hermite polynomials of the $N(0,1)$ distribution.

The eigenvectors of $T_F^* T_F$ 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 last question of interest we consider in this section concerns the behavior of the bias term in Theorem 4.3 for particular choices of α_N and h_N. We consider first the square of the bias generated by the estimation of T_F, i.e.:

$$\begin{aligned}
&\upsilon_N(\delta)\langle b_N^{\alpha_N}, \delta\rangle^2 \\
&= \upsilon_N(\delta)\langle \alpha_N(\alpha_N I + T_F^* T_F)^{-1}[T_F^* T_F - T_{\hat{F}_N}^* T_{\hat{F}_N}](\alpha_N I + T_F^* T_F)\varphi, \delta\rangle^2 \\
&\leq \alpha_N^2 N \frac{\|(\alpha_N I + T_F^* T_F)^{-1}\delta\|^2}{\|T_F(\alpha_N I + T_F^* T_F)^{-1}\delta\|^2} \|(T_F^* T_F - T_{\hat{F}_N}^* T_{\hat{F}_N})(\alpha_N I + T_{\hat{F}_N}^* T_{\hat{F}_N})\varphi\|^2
\end{aligned}$$

Using the same methodology than in the analysis of the consistency, we obtain:

$$\begin{aligned}
\upsilon_N(\delta)\langle b_N^{\alpha_N}, \delta\rangle^2 &= O\left(N\left(\frac{1}{N h_N^p} + h_N^{2\rho}\right)\|\varphi^{\alpha_N} - \varphi\|^2\right) \\
&= O\left(N\left(\frac{1}{N h_N^p} + h_N^{2\rho}\right)\alpha_N^{\beta}\right).
\end{aligned}$$

under Assumption A.6. Consider for simplicity the case $p = 1$, $\rho = 2$, and $\beta = 1$. Under the choice $h_N = k_2 N^{-(\frac{1}{4}+\varepsilon)}$ and $\alpha_N = k_2 N^{-\frac{1}{3}}$, this expression converges to zeros when $N \to \infty$.

However the regularization bias does not converges to zero. For example if $\zeta = \varphi_0 (= 1)$, $\upsilon_N(\delta)\langle b_N^{\alpha_N}, \delta\rangle^2 = O(N\alpha_N^2)$ which converges to infinity. This divergence rate is an upper bound. Indeed, we have:

$$\begin{aligned}
\upsilon_N(\delta)\langle \varphi_N^{\alpha_N} - \varphi, \delta\rangle^2 &= \frac{N\langle \alpha_N(\alpha_N I + T_F^* T_F)^{-1}\varphi, \delta\rangle^2}{\|T_F(\alpha_N I + T_F^* T_F)^{-1}\delta\|^2} \\
&= \alpha_N^2 N \frac{\langle \varphi, (\alpha_N I + T_F^* T_F)^{-1}\delta\rangle^2}{\|T_F(\alpha_N I + T_F^* T_F)^{-1}\delta\|^2} \\
&\leq \alpha_N^2 N \|\varphi\|^2 \frac{\|(\alpha_N I + T_F^* T_F)^{-1}\delta\|^2}{\|T_F(\alpha_N I + T_F^* T_F)^{-1}\delta\|^2}.
\end{aligned}$$

5 Concluding Remarks

This paper presents an efficient way to estimate nonparametrically a relation between endogenous variables using an instrumental variables definition. We also consider asymptotic properties of this estimator and the main results concern lower bounds of the speed of convergence and the asymptotic normality of the regularized solution of an ill-posed inverse problem. The resolution of this problem raises numerous questions:

i) The choice of the regularization parameter must be discussed. This choice is similar to the choice of the perturbation parameter in a ridge regression function; (see Carrasco, Florens (2000b) and references there in).

ii) We could adopt others types of regularization of the ill-posed inverse problem. In particular we regularize the problem if we choose the instrumental regression function in the set of monotonous functions. Of course the economic theory must valid this option.

iii) The treatment of several Z variables rises the usual curse of dimensionality problem. Usual technics of dimensionality reduction in nonparametric regression, such as additive models or index models, may be applied in our framework (see for a control function approach Blundell, Powell (1999));

iv) An estimation of particular functional associated to φ may be performed. A particular example is given by average derivative estimation which can be extended from the regression case to the instrumental variables case (see Florens, Heckman, Meghir, Vytlacil (2001) or Florens, Larribeau (1995)). An other example is to deduce from φ an other function by solving a differential equation like $\lambda'(z_1) = \varphi(z_1, \lambda(z_1))$ - see Loubes and Vanhems (2001)). Application could be the non parametric estimation of a surplus function in presence of endogeneous prices.

v) We may extend our result to weakly dependent dynamic data or to heteroskedastic models;

vi) The concept of regularity of functions relatively to an operator introduced through the definition of the Φ_β sets must be compared to other regularity notions (see for some preliminary results Carrasco, Florens (2001)).

vii) Finally a particularly interesting point could be to construct a fully nonparametric endogeneity test. A first idea would be to compare the estimated instrumental regression function $\hat{\varphi}_N^{\alpha_N}$ to a nonparametric estimator m_N of the conditional expectation function $E[Y \mid Z]$ by computing $\int(\hat{\varphi}_N^{\alpha_N} - m_N)^2\pi(z)dz$ (where π is a suitable weigthing function). A better approach could be to transform the equality $\varphi(z) = E[Y \mid Z]$ into $E[E[Y \mid W] \mid Z] = E[E[E[Y \mid Z] \mid W] \mid Z]$. All the conditional expectations should be estimated and the test of the equality may be performed.

APPENDIX

A Proofs

A.1 Proof of Corollary 2.1

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

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

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 $\mathcal{R}(T_F^*) = \mathcal{N}(T_F)^{\perp}, \overline{\mathcal{R}(T_F^*)} = L_F^2(Z)$ is tantamount to $\mathcal{N}(T_F) = \{0\}$.

A.2 Proof of Proposition 2.3

We have:

$$E[Y|X,W] = \alpha(X) + \beta(X) \in [\gamma(Z)|X,W].$$

If $\psi^*(X,Z) = \alpha^*(X) + \beta^*(X)\gamma^*(Z)$ is another instrumental regression function, we also have:

$$E[Y|X,W] = \alpha^*(X) + \beta^*(X) E[\gamma^*(Z)|X,W].$$

By taking the conditional expectation given X and making the differene we get :

$$\begin{aligned} & \beta(X)[E[\gamma(Z)|X,W] - E[\gamma(Z)|X]] \\ = \; & \beta^*(X)[E[\gamma^*(Z)|X,W] - E[\gamma^*(Z)|X]]. \end{aligned}$$

From the assumption of conditional linear identification, we deduce:

$$\begin{aligned} & \beta(X)[\gamma(Z) - E(\gamma(Z)|X)] \\ \overset{a.s}{=} \; & \beta^*(X)[\gamma^*(Z) - E(\gamma^*(Z)|X)]. \end{aligned}$$

Since $P[\beta(X)=0]=0$, we can write:

$$\begin{aligned}&\gamma(Z)-\beta^{**}(X)\gamma^{*}(Z)\\&\overset{a.s}{=}E[\gamma(Z)-\beta^{**}(X)\gamma^{*}(Z)|X],\end{aligned}$$

where β^{**} is defined almost surely by:

$$\beta^{**}(X)\overset{a.s}{=}\frac{\beta(X)}{\beta^{*}(X)}\in\mathcal{B}.$$

Then, since X and Z are $\mathcal{B}$-strongly measurably separable, there are only two possible cases:

1st case : The functions $\beta^{**}(X)$ and $E[\gamma(Z)-\beta^{**}(X)\gamma^{*}(Z)|X]$ are constant:

Then, if $\beta^{**}(X)=c$, we have $c=1$ by the normalization condition (i) and then:

$\gamma(Z)-\gamma^{*}(Z)$ is constant and thus equal to zero by the normalization condition (ii). Thus:

$$\beta=\beta^{*}\text{and }\gamma=\gamma^{*}.$$

Therefore:

$$\alpha(X)=E[Y|X,W]-\beta(X)E[\gamma(X)|X,W]=\alpha^{*}(X).$$

The functions α,β and γ are identified.

2nd case: The functions $\gamma(Z)$ are $\gamma^{*}(Z)$ are constant.

By the normalization condition (iii): $\alpha=\alpha^{*}=0$ while by the normalization condition (ii)

$$\gamma=\gamma^{*}=c\neq 0.$$

Therefore $\beta=\beta^{*}$.

A.3 Proof of Theorem 4.3

Let us denote by ξ_N the random variable $\sqrt{N}(T^{*}_{\hat{F}_N}r_{\hat{F}_N}-T^{*}_{\hat{F}_N}T_{\hat{F}_N}\varphi)$ and by ξ its limit distribution:

$$\sqrt{N}(\hat{\varphi}_N^{\alpha_N}-\varphi^{\alpha_N}-b_N^{\alpha_N})=(\alpha_N I+T^{*}_{\hat{F}_N}T_{\hat{F}_N})^{-1}\xi_N,$$

where, from Assumption A.7,

$$\xi_N \Longrightarrow \xi = N(0, \sigma^2 T_F^* T_F).$$

We introduce $\hat{M}_N = (\alpha_N I + T^*_{\hat{F}_N} T_{\hat{F}_N})^{-1}$ and $M_N = (\alpha_N I + T_F^* T_F)^{-1}$. For any $\delta \in L_Z^2$, we have:

$$\sqrt{\upsilon_N(\delta)}\langle \hat{\varphi}_N^{\alpha_N} - \varphi^{\alpha_N} - b_N^{\alpha_N}, \zeta \rangle = A + A_1 + A_2 + A_3,$$

where

$$A = \frac{\langle M_N \xi, \delta \rangle}{\langle M_N T_F^* T_F M_N \xi, \delta \rangle^{\frac{1}{2}}},$$

$$A_1 = \frac{\langle M_N (\xi_N - \xi), \delta \rangle}{\langle M_N T_F^* T_F M_N \xi, \delta \rangle^{\frac{1}{2}}},$$

$$A_2 = \frac{\langle (\hat{M}_N - M_N) \xi, \delta \rangle}{\langle M_N T_F^* T_F M_N \xi, \delta \rangle^{\frac{1}{2}}},$$

$$A_3 = \frac{\langle (\hat{M}_N - M_N)(\xi_N - \xi), \delta \rangle}{\langle M_N T_F^* T_F M_N \xi, \delta \rangle^{\frac{1}{2}}}.$$

The term A follows a $N(0, \sigma^2)$ and we must check that A_1, A_2 and A_2 tend to zero in probability. First, we get:

$$A_1^2 \leq \|\xi_N - \xi\|^2 \frac{\|M_N \delta\|^2}{\|T M_N \delta\|^2} \to 0.$$

Then we have:

$$A_2^2 \leq \|\xi\| \|M_N\| \|T^*_{\hat{F}_N} T_{\hat{F}_N} - T_F^* T_F\|^2 \frac{\|M_N \delta\|^2}{\|T M_N \delta\|^2},$$

which goes to zero if $\frac{1}{\alpha_N^2}\left(\frac{1}{N h_N^p} + h_N^{2\rho}\right) \to 0$. Finally, the term A_3 goes to zero faster than the terms A_1 and A_2.

A.4 Proof of proposition 4.1

Let us first remark that

$$\begin{aligned}
&\int \frac{f^2(.,z,w_1,.)}{f^2(.,z,.)\,f^2(.,.,w_{1,.})} f(.,z,.)\,f(.,.,w_{1,.})\,dzdw_1 \\
&= \int \left\{ \int \frac{f(.,z,w_1,w_2)}{f(.,z,.)\,f(.,.,w_1,w_2)} f(w_2|w_{1,.})\,dw_2 \right\}^2 f(.,z,.)\,f(.,.,w_1,.)\,dzdw \\
&\leq \int \frac{f^2(.,z,w_1,w_2)}{f^2(.,z,.)\,f^2(.,.,w_{1,}w_2)} f(.,z,.)\,f(.,.,w_1,w_2)\,dzdw_1dw_2.
\end{aligned}$$

by Jensen's inequality for conditional expectations. The first term is the H.S. norm of $T_F^{1*}T_F^1$ and one is the H.S. norm of $T_F^*\ T_F$. Then $T_F^{1*}T_F^1$ is an Hilbert Schmidt operator and $\sum_j \lambda_{j_1}^2 \leq \sum_j \lambda_j^2$.

The eigen values may be compared pairwise: using the Courant theorem (see Kress (1998), 15) we have

$$\begin{aligned}
\lambda_j^2 &= \min_{\rho_0,\rho_1,\ldots,\rho_{j-1}\in L_z^2} \max_{\substack{\|\varphi\|=1 \\ \varphi\perp(\rho_0,\ldots,\rho_{j-1})}} \langle T_F^* T_F \varphi, \varphi \rangle \\
&= \max_{\substack{\|\varphi\|=1 \\ \varphi\perp(\rho_0,\ldots,\rho_{j-1})}} \|E(\varphi|w)\|^2 \\
&\geq \max_{\substack{\|\varphi\|=1 \\ \varphi\perp(\rho_0,\ldots,\rho_{j-1})}} \|E(\varphi|w_1)\|^2 \\
&\geq \min_{\rho_0,\rho_1,\ldots,\rho_{j-1}\in L_z^2} \max_{\substack{\|\varphi\|=1 \\ \varphi\perp(\rho_0,\ldots,\rho_{j-1})}} \left\langle T_F^{1*} T^1 \varphi, \varphi \right\rangle \\
&= \lambda_{j_1}^2.
\end{aligned}$$

B A first discussion of A.4, A.5 and A.7

The objective of this appendix is to give a set of *natural* conditions which implies the main assumptions A.4, A.5 and A.7. These conditions are extremely common in nonparametric analysis, but the more questionable hypothesis assumed that the data density, defined on a compact support, is bounded from below by a strictly positive number. Even if this hypothesis is also retained by numerous papers (see e.g. Salinelli (1998), Imbens and Newey (2001)), we introduce in appendix C an extension of our approach which covers the general case. In this section, we focus on the bounded case.

Assumption B.1: *The variables Y, Z and W take values in a compact set $\mathcal{X} \subset \mathbf{R} \times \mathbf{R}^p \times \mathbf{R}^q$.*

Assumption B.2: *The probability density function f is d-continuously differentiable on $\mathcal{X}$.*

Assumption B.3: *The probability density function f is bounded from below by $\varepsilon > 0$.*

Assumption B.4: *The kernels K_y, K_z, K_w are bounded, symmetric, of order*[14] *r.*

Lemma B.1 *: Under Assumptions B.1-B.4, Assumption A.4 is satisfied with $\rho = \min(r, d)$.*

Proof. First let us remark that, if $\lambda \in L_F(Z)$,

$$\|(T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F)\lambda\|^2 \leq \|T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F\|^2_\infty \|\lambda\|^2,$$

where

$$\|T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F\|_\infty = \sup_{\|g\| \leq 1} \|(T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F)g\|.$$

This norm is majored by the Hilbert-Schmidt *(HS)* norm of this operator which satisfies (see Dunford and Schwartz (1963), XI.6):

$$\begin{aligned} A^2_N &= \|T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F\|^2_{HS} \\ &= \int \left(\frac{\int [\hat{e}_N(z,u,w) - e(z,u,w)]\, dw}{f(.,u,.)} \right)^2 f(.,z,.)\, f(.,u,.)\, dudz, \end{aligned}$$

[14] The kernel K is of order r if:

$$\forall \alpha \in N^d,\ \alpha_1 + ... + \alpha_d \in \{1, ..., r-1\},\ \int x_1^{\alpha_1} ... x_d^{\alpha_d} K(x)\, dx = 0;$$

$$\exists \alpha \in N^d,\ \alpha_1 + ... + \alpha_d = r,\ \int\int x_1^{\alpha_1} ... x_d^{\alpha_d} K(x)\, dx \neq 0.$$

Note that, if K_1 and K_2 are of order r, then $K_1 K_2$ is also of order r.

where

$$e(z,u,w) = \frac{f(.,u,w)\, f(.,z,w)}{f(.,z,.)\, f(.,.,w)},$$

and $\hat{e}_N(z,u,w)$ is the kernel estimator of $e(z,u,w)$. Under Assumptions B.1-B.3, Assumption A.1 (Existence of the Hilbert Schmidt norm) is obviously satisfied.

We can restrict the integration domain to the interior of the compact support of e since it boundaries has a zero measure[15]. We linearize the term $\hat{e}_N(z,u,w) - e(z,u,w)$ to get:

$$\begin{aligned}
&\hat{e}_N(z,u,w) - e(z,u,w) \\
&\approx \frac{f(.,z,w)}{f(.,z,.)\, f(.,.,w)}[\hat{f}_N(.,u,w) - f(.,u,w)] \\
&+ \frac{f(.,u,w)}{f(.,z,.)\, f(.,.,w)}[\hat{f}_N(.,z,w) - f(.,z,w)] \\
&- \frac{f(.,u,w)\, f(.,z,w)}{[f(.,z,.)]^2 f(.,.,w)}[\hat{f}_N(.,z,.) - f(.,z,.)] \\
&- \frac{f(.,u,w)\, f(.,z,w)}{f(.,z,.)\,[f(.,.,w)]^2}[\hat{f}_N(.,.,w) - f(.,.,w)] = \sum_{j=1}^{4} B_j,
\end{aligned}$$

and

$$A_N = \| \frac{\sum_{j=1}^{4} \int B_j dw}{[f(.,u,.)]^2} \| + R_N \leq \sum_{j=1}^{4} \| \frac{\int B_j dw}{[f(.,u,.)]^2} \| + R_N.$$

In this expression the norm are the usual norm in the Hilbert space of functions and R_N is a remainder term resulting for the linearization. Let us consider for example the first term of the sum:

$$\| \frac{\int B_1 dw}{[f(.,u,.)]^2} \|^2 = \int \frac{f(.,z,.)}{f(.,u,.)} \left(\int b_1(z,w) (\hat{f}_N(.,u,w) - f(.,u,w)) dw \right)^2 dz du,$$

with:

$$b_1(z,w) = \frac{f(.,z,w)}{f(.,z,.)\, f(.,.,w)}.$$

[15] The behavior of the integral of $\left(\int [\hat{e}_N(z,u,w) - e(z,u,w)]\, dw \right)^2$ could be studied at the boundaries of the support. This behavior would be analogous the behavior at a interior, with a higher bias term (see e.g. Wand, Jones (1995)). We tank D. Bosq for this remark.

The random variable $\|\frac{\int B_1 dw}{[f(.,u,.)]^2}\|^2$ is positive, so its rate of convergence is the same[16] than the rate of $E\left[\|\frac{\int B_1 dw}{[f(.,u,.)]^2}\|^2\right]$, i.e.:

$$E\left[\int \frac{f(.,z,.)}{f(.,u,.)}\left(\int b_1(z,w)\left(\hat{f}_N(.,u,w)-f(.,u,w)\right)dw\right)^2 dzdu\right].$$

The usual argument on the behavior of $E[(\hat{f}_N - f)^2]$ may be trivially extended to the integral of the density and we get:

$$E\left[\left(\int b_1(z,w)\left(\hat{f}_N(.,u,w)-f(.,u,w)\right)dw\right)^2\right] = O_p\left(\frac{1}{Nh_N^p}+h_N^{2\min(r,d)}\right).$$

Note that the integration w.r.t. w implies that only the dimension of Z appears in the exponent of h_N. We mutliply this quantity by $\frac{f(.,z,.)}{f(.,u,.)}$ and we integrate out z and u. The convergence rate is not modified and we get

$$\|\frac{\int B_1 dw}{[f(.,u,.)]^2}\|^2 = O_p\left(\frac{1}{Nh_N^p}++h_N^{2\min(r,d)}\right).$$

The same argument applies for the four terms. The remainder term in an negligible using an elementary extension of usual argument on the behavior the MSE of the kernel regression estimation (see e.g. Bosq (1998), theorem 3.1 p. 68). ■

The following lemma is essential to check assumption A.5 and A.7 from B.1-B.4. Actually this result will be used both as a verification of assumption A.5 and for assumption A.7. For simplicity we only consider the homoscedastic case.

Lemma B.2 *: Under Assumptions A.1, A.2, A.3, A.6 and technical Assumptions B1-B.3, we get:*

$$\frac{\sqrt{N}}{N}\sum_{i=1}^{N}\frac{f(.,z,w_i)}{f(.,z,.)\,f(.,.,w_i)}\left[y_i-\varphi(z_i)\right] \Longrightarrow N(0,\sigma^2 T_F^* T_F).$$

Proof. We denote by:

$$V_N(z) = \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\frac{f(.,z,w_i)}{f(.,z,.)\,f(.,.,w_i)}\left[y_i-\varphi(z_i)\right] = \sqrt{N}\times\frac{1}{N}\sum_{i=1}^{N}\beta_i(z),$$

[16] If $X_N \geq 0$ is such that $E(X_N) \sim O(1)$, then $X_N \sim O(1)$. This is an application of Bienaymé-Tchebychev inequality: if $E(X_N) < M$ as $N \to \infty$, we get $P[X_N > \frac{M}{\varepsilon}] \leq E(X_N)\times\frac{\varepsilon}{M} \leq \varepsilon, \forall\varepsilon$ and $X_N \sim O(1)$. In the same way $E(X_N) \sim O(\alpha_N) \Longrightarrow X_N \sim O(\alpha_N)$.

where $\beta_i(z)$ is a sequence of i.i.d. random variables satisfying:

$$E\left[\beta_i(z)\right] = 0,$$

and

$$E\left[\beta_i^2(z)\right] = \sigma^2 \int \frac{f(.,z,w_i)^2}{f(.,z,.)^2 f(.,.,w_i)^2} f(.,z,.)\, dz < \infty.$$

Moreover, we have:

$$E\left[\|\beta_i\|^2\right] = \sigma^2 \int \frac{f(.,z,w)^2}{f(.,z,.)^2 f(.,.,w)^2} f(.,z,.)\, f(.,.,w)\, dzdw < \infty.$$

This Hilbert Schmidt assumption is implied by Assumption B.1-B.3. So V_N converges to a gaussian process (see Van der Vaart, Wellner, theorem 1.8.4 p. 50). The variance is given by an operator K such that, for any $\varphi, \psi \in L_Z^2$, we have:

$$\begin{aligned}
\langle K\psi, \varphi\rangle &= E[\langle V_N, \varphi\rangle\langle \psi, V_N\rangle] \\
&= E[\int\int V_N(z)V_N(u)\varphi(z)\psi(u) f(.,z,.)\, f(.,u,.)\, dzdu] \\
&= \int\int E[V_N(z)V_N(u)]\varphi(z)\psi(u) f(.,z,.)\, f(.,u,.)\, dzdu \\
&= \sigma^2 \int\int E[\frac{f(.,z,w_i)\, f(.,u,w_i)}{f(.,z,.)\, f(.,.,w_i)}]\varphi(z)\psi(u) f(.,z,.)\, f(.,u,.)\, dzdu.
\end{aligned}$$

We obtain:

$$K\psi = \sigma^2 \int\int \frac{f(.,z,w)\, f(.,u,w)}{f(.,z,.)\, f(.,.,w)} \psi(u)\, dzdu = \sigma^2 T_F^* T_F \psi,$$

and finally we get :

$$V_N \Longrightarrow N(0, \sigma^2 T_F^* T_F). \blacksquare$$

Lemma B.3 *: Under Assumptions A.1, A.2, A.3, A.6 and technical Assumptions B.1-B.4, Assumption A.5 is satisfied with* $\rho = \min(r, d)$.

Proof. Using standard linearization, we first replace $A_N = r_{\hat{F}_N}^* - T_{\hat{F}_N}^* T_{\hat{F}_N} \varphi$ by:

$$\tilde{A}_N = \int \frac{f(.,z,w)}{f(.,z,.)\, f(.,.,w)} (y - \varphi(u)) \hat{f}_N(y,u,w) dydudw. \tag{B.1}$$

Actually, from the above expansion of the integrand of A_N as the product of four functionals of f, the derivative is computed as the sum of four

corresponding terms. But three among these four terms are nil thanks to the assumption $T_F\varphi = r_F$.

Moreover, we can decompose $\tilde{A}_N$ as follow:

$$\tilde{A}_N = \frac{1}{N}\sum_{i=1}^{N} \frac{f(.,z,w_i)}{f(.,z,.)\, f(.,.,w_i)}(y_i - \varphi(z_i)) + R.$$

Then we get:

$$\|\tilde{A}_N\| \leq \|\frac{1}{N}\sum_{i=1}^{N} \frac{f(.,z,w_i)}{f(.,z,.)\, f(.,.,w_i)}(y_i - \varphi(z_i))\| + \|R\|.$$

The remainding term R is of the form:

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\int a(s)K_{h_N}(s-s_i)ds - a(s_i)\right\},$$

where $s = (y,u,w)$ and $K_{h_N}(s-s_i) = K_{y,h_{yN}}(y-y_n)K_{z,h_{zN}}(u-z_n)\,K_{w,h_{wN}}(w-w_n)$

$$a(s) = \frac{f(.,z,w)}{f(.,z,.)\, f(.,.,w)}(y - \varphi(u))$$

and the usual analysis of bias term in kernel smoothing inference shows that the norm of this term in an $O\left(h_N^{\min(r,d)}\right)$. The result follows from the application of Lemma B.2 ■

Assumption B.5: The smoothing parameter h_N satisfies $Nh_N^{2\min(r,d)} \to 0$ as $N \to \infty$.

Lemma B.4 *: Under Assumptions A.2, A.3, A.6 and technical Assumptions B.1-B.5, Assumption A.7 is satisfied.*

Proof. The linearized form of $\sqrt{N}(r^*_{\hat{F}_N} - T^*_{\hat{F}_N}T_{\hat{F}_N}\varphi)$ can be written as:

$$\frac{\sqrt{N}}{N}\sum_{i=1}^{N} \frac{f(.,z,w_i)}{f(.,z,.)\, f(.,.,w_i)}(y_i - \varphi(z_i)) - \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\left\{\int a(s)K_{h_N}(s-s_i)ds - a(s_i)\right\},$$

where $a(s)$ is defined in the proof of Lemma 3. Under Assumption B.5, the second term of the previous decomposition goes to zero. We use Lemma 2 to get:

$$\sqrt{N}(r^*_{\hat{F}_N} - T^*_{\hat{F}_N}T_{\hat{F}_N}\varphi) \Longrightarrow N(0, \sigma^2 T_F^* T_F). \blacksquare$$

C Generalization allowing to consider non bounded densities

C.1 Definitions

Let us consider the random vector $S = (Y, Z, W) \in \mathbf{R} \times \mathbf{R}^p \times \mathbf{R}^q$, with cumulative distribution function F, and the two cumulative distribution functions G and H defined on $\mathbf{R}^p$ and $\mathbf{R}^q$ respectively. We assume that:

$$L^2_G(Z) \subset L^2_F(Z) \text{ and } L^2_F(W) \subset L^2_H(W), \tag{C.1}$$

where $L^2_G(Z)$ (resp. $L^2_H(W)$) denotes the space of squared integrable functions with respect to G (resp. H).

Remark C.1 *: If we only consider distributions characterized by their densities with respect to the Lebesgue measure f, g, h, we easily check that the two previous conditions are satisfied if it exists two strictly positive numbers c_1 and c_2 satisfying:*

$$f(., z, .) \leq c_1 g(z) \ F\text{-a.s. and } h(w) \leq c_2 f(., ., w) \ F\text{-a.s.}. \tag{C.2}$$

The two constants are then necessarily greater than one.

Remark C.2 *: The relation between F and G, H can be interpreted in two ways:*

Remark 1 *i) we can fix G and H and consider the class $\mathcal{F}$ of F satisfying the two constrains*

ii) we can choose G and H depending on F (for example, we choose G as the marginal distribution of F, conditionally to a subset on which $f(z)$ is bounded).

In ii), G and H must be estimated. To simplify the presentation, we adopt i) in which G and H are fixed, but the approach can be generalized to ii).

For any F belonging to $\mathcal{F}$, the conditional expectation operator T_F is now considered as an operator from $L^2_G(Z)$ to $L^2_H(W)$. We always assume that T_F is an Hilbert Schmidt operator relatively to these spaces. This is equivalent to assume that:

$$\int \frac{f^2(., z, w)}{g^2(z) f^2(., ., w)} g(z) h(w) dz dw < \infty. \tag{C.3}$$

Remark C.3 *: If the conditional expectation operator from $L^2_F(Z)$ to $L^2_F(W)$ satisfies an Hilbert Schmidt condition, we obtain the property (C.3) from the conditions (C.2).*

Definition C.1 *: The function φ belonging to $L^2_G(Z)$ is an instrumental regression if $T_F\varphi = r_F$, with $r_F = E[Y \mid W]$.*

Remark C.4 *: Since the function φ is now defined in an restricted space, the identification condition becomes: the function φ is identifiable if we have $E[\lambda(Z) \mid W] = 0$ a.s.and $\lambda \in L^2_G(Z) \Rightarrow \lambda = 0$ a.s..*

C.2 Dual, spectral decomposition and regularization

Let us first denote that the dual of T_F as an operator from $L^2_G(Z)$ to $L^2_H(Z)$ is not the conditional expectation of the functions W given Z. In the dominated case, T^*_F satisfies:

$$T^*_F\psi(z) = \int \frac{f(.,z,w)h(w)}{g(z)f(.,.,w)}\psi(w)dw, \tag{C.4}$$

because $\langle T_F\varphi, \psi\rangle_H = \langle \varphi, T^*_F\psi\rangle_G$ ($\langle .,.\rangle_H$ denotes the inner product in $L^2_H(Z)$ and $\langle .,.\rangle_G$ denotes the inner product in $L^2_G(Z)$). We have:

$$T^*_F T_F\varphi(z) = \int \left\{ \frac{1}{g(z)} \int \frac{f(.,z,w)f(.,u,w)h(w)}{f^2(.,.,w)}dw \right\} \varphi(u)du. \tag{C.5}$$

The Hilbert Schmidt assumption always implies the compactness of T_F, T^*_F, $T^*_F T_F$, $T_F T^*_F$, the existence of vectors $\varphi_j \in L^2_G(Z)$, $\psi_j \in L^2_H(Z)$, and λ^2_j satisfying the properties *i)* to *viii)* of Subsection 2.2. The general theory of regularization applies for this choice of T^*_F. We define φ^α by:

$$\varphi^\alpha = (\alpha I + T^*_F T_F)^{-1} T^*_F r_F, \tag{C.6}$$

and we get $\|\varphi^\alpha - \varphi\|_G \to 0$.

C.3 Estimation

The estimation of φ is obtained by replacing the density f and its margins by their estimators $\widehat{f}_N$ (see Section 4). We assume that:

$$\frac{K_{w,h_{wN}}\left(\frac{w-w_n}{h_N}\right)}{\sum_n K_{w,h_{wN}}\left(\frac{w-w_n}{h_N}\right)} \in L^2_H(W) \text{ and } \frac{K_{z,h_{zN}}\left(\frac{z-z_n}{h_N}\right)}{\sum_n K_{z,h_{zN}}\left(\frac{z-z_n}{h_N}\right)} \in L^2_G(Z).$$

We do not detail the computation but we just underline that $(\alpha_N I + T^*_{\hat{F}_N} T_{\hat{F}_N})$ is a finite rank operator, and the solution of the equation:

$$(\alpha_N I + T^*_{\hat{F}_N} T_{\hat{F}_N})\varphi = T^*_{\hat{F}_N} r_{\hat{F}_N},$$

is obtained as in Annex D.

C.4 Asymptotic properties

We must check whether the assumptions A.4 and A.5 are satisfied. This imposes some additional regularity assumptions. For simplicity we restrict our attention to second order kernels and second order differentiable functions. The first property to check concerns the expectation of the Hilbert Schmidt norm of $T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F$, i.e.:

$$\int \frac{1}{g(u)g(z)} \left\{ \int \left[\frac{\widehat{f}_N(.,z,w)\widehat{f}_N(.,u,w)}{\widehat{f}^2_N(.,.,w)} - \frac{f(.,z,w)f(.,u,w)}{f^2(.,.,w)} \right] h(w)dw \right\}^2 dudz. \tag{C.7}$$

The computation is done by linearization of the terms in the bracket. Let us consider for example the first term of this linearization, , i.e.

$$\frac{f(.,u,w)}{f^2(.,.,w)}(\widehat{f}_N(.,z,w) - f(.,z,w)).$$

The integral with respect to w is approximated by:

$$\frac{1}{Nh^p_N} \sum_n K_{z,h_{zN}} \left(\frac{z - z_n}{h_N} \right) \frac{f(.,u,w_i)h(w_i)}{f^2(.,.,w_n)} - \int \frac{f(.,z,w)f(.,u,w)}{f^2(.,.,w)} h(w)dw, \tag{C.8}$$

up to an h^2_N term which we integrate in the bias term. This term contribute to the norm by a bias term and a variance term. The variance term is:

$$\frac{1}{Nh^p_N} \int K^2(u)du \int \frac{1}{g(z)g(u)} \frac{f^2(.,u,w)h^2(w)}{f^2(.,.,w)} f(.,z,w)dwdudz. \tag{C.9}$$

This integral must be convergent. This can be obtained by replacing the conditions $(B.2)$ by:

$$f(z) \leq d_1 g^2(z) \ F\text{-a.s. and } h(w) \leq d_2 f^2(w) \ F\text{-a.s..} \tag{C.10}$$

and by assuming that $\int f(w \mid z)g(z)dz \leq m$. We obtain:

$$\begin{aligned} &\int \frac{1}{g(z)g(u)} \frac{f^2(.,u,w)h^2(w)}{f^2(.,.,w)} f(z,w)dwdudz \\ &\leq d_1 d_2 m \int \frac{f^2(.,u,w)}{g^2(u)f^2(.,.,w)} g(u)h(w)dudw, \end{aligned} \tag{C.11}$$

and the integral converges with the Hilbert Schmidt assumption. The convergence to the squared bias term imposes some additional assumptions on the second derivatives of f. We must for example assume that:

$$\int \frac{\left(\partial^2 f\right)^2}{g^2(u)f^2(.,.,w)} g(u)h(w)dudw < \infty,$$

where $\partial^2 f$ denotes the sum of the second derivatives of $f(z,w)$ with respect to z and w.

The second element to establish the asymptotic properties is the asymptotic behavior of $T^*_{\hat{F}_N} r_{\hat{F}_N} - T^*_{\hat{F}_N} T_{\hat{F}_N} \varphi$. This term can be decomposed as $T^*_F(r_{\hat{F}_N} - T_{\hat{F}_N}\varphi) + (T^*_{\hat{F}_N} - T_F)(r_{\hat{F}_N} - T_{\hat{F}_N}\varphi)$. Since $r_{\hat{F}_N} - T_{\hat{F}_N}\varphi \to r_F - T_F\varphi = 0$, we check that under very general assumptions, the second term of this decomposition is negligible with respect to the first term. We have then by linearization of the conditional expectation:

$$\begin{aligned} T^*_F(r_{\hat{F}_N} - T_{\hat{F}_N}\varphi) &= \int \frac{f(.,z,w)h(w)}{g(z)f^2(.,.,w)}(y - \varphi(u))\widehat{f}_N(y,u,w)dydudw \\ &= \frac{1}{N}\sum \frac{f(.,z,w_i)h(w_i)}{g(z)f^2(.,.,w_i)}(y_i - \varphi(z_i)) + R. \end{aligned}$$

The first term, when multiplying by $\sqrt{N}$ converges in $L^2_G(Z)$ to a gaussian law, with zero mean and with variance $\sigma^2\Omega$ characterized by:

$$\langle \Omega\varphi_1, \varphi_2 \rangle_G = \int w(z,w)\varphi_1(z)\varphi_2(u)g(u)dudz,$$

with

$$w(z,w) = \frac{1}{g(z)} \int \frac{f(.,z,w)f(.,u,w)h^2(w)}{f^3(.,.,w)}dw.$$

The Hilbert space convergence comes from the condition:

$$\int \frac{f^2(.,z,w)h^2(w)}{g(z)f^4(.,.,w)} g(z)f(.,.,w)dzdw < \infty,$$

obtained with $(C.2)$ and the Hilbert Schmidt condition. Finally, we check with the usual approach that the remainder term is proportional to h_N^2.

D Numerical implementation

We will show in this appendix how our estimation procedure reduces to finite dimensional matrix computation. To simplify the notations introduced in section 4, we drop out indexes ($\alpha_N, \hat{\varphi}_N^{\alpha_N}, K_{z,h_{z^N}}$..becomes $\alpha, \hat{\varphi}, K$...). Definition (4.1) becomes:

$$\begin{aligned} \alpha\hat{\varphi}(z) &+ \sum_n \sum_i \int \varphi(z) K(z - z_i)\, dz = \int \frac{K(w-w_i)K(w-w_n)}{\sum_\ell K(w-w_\ell)} dw \frac{K(z-z_n)}{\sum_\ell K(z-z_\ell)} \\ &= \sum_n \sum_i y_i \int \frac{K(w-w_i)K(w-w_n)}{\sum_\ell K(w-w_\ell)} dw \frac{K(z-z_n)}{\sum_\ell K(z-z_\ell)} \end{aligned}, \tag{D.1}$$

or equivalently

$$\alpha\hat{\varphi}(z) + \sum_{n,i} \xi_i a_{in} \beta_n(z) = \sum_{n,i} y_i a_{in} \beta_n(z) \tag{D.2}$$

where

$$\begin{aligned} a_{in} &= \int \frac{K(w - w_i) K(w - wn)}{\sum_\ell K(w - w_\ell)} dw \\ \xi_i &= \int \hat{\varphi}(z) K(z - z_i)\, dz \\ \beta_n(z) &= \frac{K(z-z_n)}{\sum_\ell K(z-z_\ell)} \end{aligned}$$

If we multiply this equality by $K(z - z_j)$ we get, after integrated out z:

$$\alpha\xi_j + \sum_{n,i} \xi_i a_{in}\, b_{nj} = \sum_{n,i} y_i a_{in}\, b_{nj} \tag{D.3}$$

where

$$b_{nj} = \int \frac{K(z - z_n) K(z - z_j)}{\sum K(z - z_\ell)} dz.$$

Using the matrix notation:

$$\xi = (\xi_i)_{i=1,\ldots,N} \qquad A = (a_{in})_{\substack{i=1,\ldots N \\ n=1,\ldots,N}} \qquad B = (b_{in})_{\substack{i=1,\ldots,N \\ n=1,\ldots,N}}$$
$$y = (y_i)_{i=1,\ldots,N} \qquad \beta(z) = (\beta_n(z))_{n=1,\ldots,N}$$

We have to solve:

$$\left(\alpha I + AB'\right)\xi = AB'y \tag{D.4}$$

which implies, using D.2

$$\begin{array}{rcl} \hat{\varphi}(z) & = & \frac{1}{\alpha}(y-\xi)' A\beta(z) \\ & = & \frac{1}{\alpha} y' \left[I - \left[\alpha I + AB'\right]^{-1} AB'\right]' A\beta(z) \end{array} \tag{D.5}$$

To a pratical point of view we need first to compute by integration the elements of A and B and to inverse an $N \times N$ matrix.

This computation can be simplified if we approximate, for example, ξ_i by $\varphi(z_i)$ and by simplification of the elements of the A matrix. Using this approximation the estimator $\hat{\varphi}$ is a solution of:

$$\begin{array}{rcl} \alpha\hat{\varphi}(z) & + & \sum_n \dfrac{\sum_i \hat{\varphi}(z_i) K(w_n - w_i)}{\sum_\ell K(w_n - w_\ell)} \dfrac{K(z - z_n)}{\sum_\ell K(z - z_\ell)} \\ & = & \sum_n \dfrac{\sum_i y_i K(w_n - w_i)}{\sum_\ell K(w_n - w_\ell)} \dfrac{K(z - z_n)}{\sum_\ell K(z - z_\ell)} \end{array} \tag{D.6}$$

If $z = z_j$ $j = 1, ..., N$, we get an $N \times N$ linear system which can be solved in the $\hat{\varphi}(z_j)$ and $\hat{\varphi}(z)$ can be compute using (D.6) for any z. Properties of this approximation had to be studied. Roughly speaking the approximation magnitude is on the same nature of the bias term in kernel smoothing $\left(h_N^{2p}\right)$ and the asymptotic behavior of this estimator is identical to the estimator studied in the paper.

E Counterexamples

E.1 About Example 1.2

Let us define :

$V = Z - E(Z|W)$ and $h(v) = V^2$.

Then, if φ and $\varphi^*$are well-defined as solutions of (1.2) and (1.4) respectively:

$$E\left[\varphi(Z) - \varphi^*(Z)|W\right] = E\left[v^2|W\right] = Var\left[Z|W\right] = \sigma^2(W).$$

For an explicit example, let us consider the case:

$$E\left[Y|Z,W\right]=aZ+bZ^{2}+V^{2},$$

that is: $\varphi^{*}\left(Z\right)=aZ+bZ^{2}$. Then:

$$\begin{aligned} E\left[\varphi\left(Z\right)|W\right] &= E\left[\varphi^{*}\left(Z\right)|W\right]+\sigma^{2}\left(W\right) \\ &= aE\left(Z|W\right)+bE\left(Z^{2}|W\right)+\sigma^{2}\left(W\right). \end{aligned}$$

If, for instance:

$$E\left(Z|W\right)=\sigma\left(W\right),$$

$$E\left[\varphi\left(Z\right)|W\right]=a\sigma\left(W\right)+2b\sigma^{2}\left(W\right)+\sigma^{2}\left(W\right).$$

This is consistent with:

$$\varphi\left(Z\right)=aZ+\left(b+\frac{1}{2}\right)Z^{2}=\varphi^{*}\left(Z\right)+\frac{Z^{2}}{2}.$$

In this case:

$$\frac{\partial E\left(Y|W\right)}{\partial W}=a\frac{\partial E\left(Z|W\right)}{\partial W}+2\left(2b+1\right)E\left(Z|W\right)\frac{\partial E\left(Z|W\right)}{\partial W},$$

and therefore a sotution $\tilde{\varphi}^{*}$ to (1.5) is given by:

$$\tilde{\varphi}^{*}\left(Z\right)=aZ+\left(2b+1\right)Z^{2}=\varphi^{*}\left(Z\right)+\left(b+1\right)Z^{2}.$$

E.2 About measurable separability

Let X and Z be two binary variables:

$$\begin{aligned} P\left[X=0\right]+P\left[X=1\right]&=1, \\ P\left[Z=0\right]+P\left[Z=1\right]&=1. \end{aligned}$$

As soon as:

$$0<P\left[X=Z\right]<1,$$

X and Z are measurable separable. To see this, let us consider for example them are:

$$P\left[X=0,Z=1\right]>0.$$

Then two functions $\alpha(X)$ and $\gamma(Z)$ such that:

$$\alpha(X) \overset{a.s.}{=} \gamma(Z),$$

should fulfill:

$$\gamma(0) = \gamma(1),$$

and

$$(\alpha(0) = \gamma(0)) \text{ or } (\alpha(1) = \gamma(1)).$$

In this case, we conclude that α and γ are constant functions. However, one may find in general two non-constant functions $a(X)$ and $b(Z)$ such that:

$$a(X) + b(Z) \overset{a.s.}{=} \beta(X) + \gamma(Z),$$

with

$$\beta(0) \neq 0 \text{ and } \beta(1) \neq 0.$$

To see this, let us assume:

$$P[X = 1, Z = 0] = 0,$$

and define $a(0), a(1), b(0), b(1)$ arbitrarily but conformable with:

$$\begin{cases} a(0) \neq a(1), b(0) \neq b(1), \\ a(0) + b(0) \neq 0, a(1), b(1) \neq 0, a(0) + b(1) \neq 0. \end{cases}$$

Then, if we define:

$$\begin{cases} \gamma(0) = \frac{a(0)+b(0)}{\beta(0)}, \\ \beta(1) = \frac{a(0)+b(1)}{\gamma(0)}, \\ \gamma(1) = \frac{a(1)+b(1)}{\beta(1)}, \end{cases}$$

we do ensure:

$$a(X) + b(Z) \overset{a.s.}{=} \beta(X)\gamma(Z).$$

■

REFERENCES

Abadie A. (2001), *Semiparametric Instrumental Variable Estimation of Treatment ResponseModels*, Discussion Paper, Harvard University.

Ai, C. , R. Blundell and X. Chen (2001), *Engle Curves with Endogenous Expenditures*, Discussion Paper.

Amemiya, T. (1974), *The Non Linear Two-Stage Least Squares Estimator,* Journal of Econometrics, **2**, 105-110.

Amemiya, T. (1975), *The Non Linear Limited-Information Maximum-Likelihood Estimator and the Modified Non Linear Two-Stage Least Squares Estimator,* Journal of Econometrics, **3**, 375-386.

Basmann, R.L. (1957), *A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equations*, Econometrica, **25**, 77-83.

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

Blundell, R., and J., Powell (1999), , *Endogeneity in Single Index Models*, Manuscript, UCL.

Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*, Lecture Notes in Statistics, Springer-Verlag, New York, 2^{nd} edition.

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

Carrasco, M., and J.P., Florens (2000b), *Efficient GMM Estimation Using the Empirical Characteristic Function,* Discussion Paper, GREMAQ, University of Toulouse.

Carrasco, M., and J.P., Florens (2001), *Spectral Method for Deconvolving a Density,* Discussion Paper, GREMAQ.

Chen X., Hansen, L.P. and J. Scheinkman (2000), Principal Components and the Long Run, Working Paper, University of Chicago.

Chen, X., and X., Shen (1998), *Sieve Extremum Estimates for Weakly Dependent Data,* Econometrica, **66**, 2.

Chen, X., and H. White (1992), *Central Limit and Functional Central Limit Theorems for Hilbert Space-Valued Dependent Processes*, Working Paper, University of San Diego.

Darolles, S., Florens, J.P., and C., Gouriéroux (1998), *Kernel Based Nonlinear Canonical Analysis*, Discussion Paper CREST 9858, forthcoming in Journal of Econometrics.

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

Das, M. (2001), *Instrumental Variables Estimation of Nonparametric Models with Discrete Endogenous Regressors*, Discussion Paper, Columbia University.

Dunford, N., and J., Schwartz (1963), *Linear Operators 2,* Wiley, New York.

Florens, J. P. (2000), *Inverse Problems and Structural Econometrics: The Example of Instrumental Variables*, Invited Lecture at the 8^{th} World Congress of the Econometric Society.

Florens, J.P., Heckman, J., Meghir, C. and E. Vytlacil (2001), *Instrumental Variables, Local Instrumental Variables and Control Functions,* Manuscript, University of Toulouse.

Florens, J.P., and S., Larribeau (1995), *Derivative Consistent Estimation of Misspecified Models,* Manuscript, University of Toulouse.

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., Mouchart, M., and J.F. Richard (1974), *Bayesian Inference in Error-in-variables Models*, Journal of Multivariate Analysis, **4**, 419-432.

Florens, J.P., Mouchart, M., and J.F. Richard (1987), *Dynamic Error-in-variables Models and Limited Information Analysis*, Annales d'Economie et Statistiques, **6/7**, 289-310.

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

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

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

Hansen, L.P. (1982), *Large Sample Properties of Generalized Method of Moments Estimators,* Econometrica, **50**, 1029-1054.

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

Heckman, J., Ichimura, H., Smith, J., and P., Todd (1998), *Characterizing Selection Bias Using Experimental Data*, Econometrica, **66**, 1017-1098.

Heckman, J., and V., Vytlacil (1999), *Local Instrumental Variables*, Working Paper, University of Chicago.

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

Imbens, G., and W., Newey (2001), *Identification and Inference in Triangular Simulteneous Equations Models without Additivity*, Discussion Paper.

Kress, R. (1998), *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.

Loubes, J.M. and A. Vanhems (2001), *Differential Equation and Endogeneity*, Discussion Paper, GREMAQ, University of Toulouse.

Luenberger O. (1969), *Optimization by Vector Space Methods*, Wiley, New York.

Malinvaud E. (1970), *Methodes Statistiques de l'Econométrie*, Dunod, Paris.

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** n°6, 974-987.

Nelson C.R., R. Startz and F. Zivot (1998), *Valid Confidence Intervals and Inference in the Presence of Weak Instruments*, International Economic Review, 39, 1119-1144.

Newey, W., and J., Powell (2000), *Instrumental Variables for Nonparametric Models,* MIT Discussion Paper.

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

Pagan A.R. (1986), *Two Stage and Related Estimators and Their Applications*, Review of Economic Studies, **53**, 513-538.

Pagan A., and A., Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press.

Reiersol, O. (1941), *Confluence Analysis of Lag Moments and other Methods of Confluence Analysis*, Econometrica, **9**, 1-24.

Reiersol, O. (1945), *Confluence Analysis by Means of Instrumental Sets of Variables*, Arkiv for Mathematik, Astronomie och Fysik, 32.

Sargan, J.D. (1958), *The Estimation of Economic Relationship using Instrumental Variables,* Econometrica, **26**, 393-415.

Salinelli, E (1998), *Non Linear Principal Component i: Absolutely Continuous Variables*, Annals of Statistics, **86**, 596-616.

Staiger D. and J.H. Stock (1997), *Instrumental Variables and Week Instruments*, Econometrica, **65**, 557-586.

Theil, H.(1953), *Repeated Least Squares Applied to complete Equations System*, The Hague: Central Planning Bureau (mimeo).

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.

Van Rooij 0. and C.H. Ruymgaart (1999), *On Inverse Estimation* in Asymptotic, Non Parametrics and Time Series, 579-613, Dekker, 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.

Wand, M.P. and M.C.Jones (1995), *Kernel Smoothing*, Chapman and Hull, London.

Wang J. and F. Zivot (1998), *Inference on a Structural Parameter in Instrumental Variables Regressions with Weak Instruments*, Econometrica, **66**, 1389-1404.