Nonparametric Instrumental Regression[1]

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

April 26, 2001

[1]We first want to thank our coauthors on papers strongly related with this one: M. Carrasco, C. Gouriéroux, J. Heckman, C. Meghir and E. Vytlacil. We also acknowledge helpful comments from the editor R. Blundell, the four referees and X. Chen, L. Hansen, P. Lavergne, W. Newey, J.M. Rolin. We thank the participants to conferences and seminars in Toulouse, Santiago, London, Chicago, Montreal, Stony Brook and Louvain-la-Neuve, Seattle, Standford, Paris.

[2]Société Générale Asset Management, Hedge Funds Quantitative Research and CREST, Paris.

[3]IDEI and GREMAQ-Université des Sciences Sociales, Toulouse.

[4]Université de Montréal.

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

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

Classification JEL: C14, C30.

Résumé

Nous nous interessons à l'estimation nonparamétrique d'une fonction de régression instrumentale φ. Cette fonction est définie à l'aide de conditions de moment provenant d'un modèle économétrique structurel de la forme $E\left[Y-\varphi\left(Z\right)\mid W\right]=0$, 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, Equation intégrale, problème mal posé, Régularisation de Tikhonov, Lissage par noyau.

Classification JEL : C14, C30.

1 Introduction

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

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

where the function $\varphi(.)$ should define the relationship of interest while U is an error term. 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. [34, Malinvaud (1970), ch. 9], [20, Florens, Mouchart, Richard (1974)] or [21, Florens, Mouchart, Richard (1987)] for the linear case). Instrumental variable analysis was proposed by [40, Reiersol (1941)], [41, Reiersol (1945)] and extended by [43, Theil (1953)], [4, Basmann (1957)] and [42, 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[1]:

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 [2, Amemiya (1974)]), but its

[1] A general comparison between these three concepts and their extensions to more general treatment models is done in [18, Florens, Heckman, Meghir, Vytlacil (2001)].

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 [36, Newey, Powell (2000)], quoted in [39, Pagan, Ullah (1999)]);

ii) A second approach is now called *control function approach* and was systematized by [37, Newey, Powell, 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. [29, Imbens, Angrist (1994)], [27, Heckman, Ichimura, Smith, Todd (1998)] and [28, 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 $V(Z, W) = Z - E[Z \mid W]$.

Example 1.2: We consider the case of a trivariate random vector (Y, Z, W) where: $W \sim N(0, 1)$, $Z \mid W \sim N(W, W^2)$ and $Y \mid Z, W \sim N(Z + (Z - W)^2, 1)$. Using the first definition, we get $\varphi(Z) = Z + \frac{1}{2}Z^2$, while the second definition with the control function $V(Z, W) = Z - W$ gives $\varphi(Z) = Z + Cst$. The third definition implies $\varphi(Z) = Z + Z^2 + Cst$.

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.

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 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 supplementary restriction is imposed 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 supplementary 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 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 $\Sigma_{YW} = cov(Y,W)$. The variance Σ_{WW} of W is assumed to be a regular matrix. If Z and W have the same dimension and if Σ_{ZW} is regular, $A = \Sigma_{YW}\Sigma_{ZW}^{-1}$ and $b = \mu_Y - A\mu_Z$. We will see later that this linear solution is the unique solution of (2.2) in the normal case. If Z and W do not have the same dimension, supplementary conditions are needed for existence and uniqueness of φ.

It would be useful to use the two following notations:

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

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

These two linear operators satisfy:

$$\langle \varphi(Z), \psi(W) \rangle = E[\varphi(Z)\psi(W)] = \langle T_F\varphi(Z), \psi(W) \rangle = \langle \varphi(Z), T_F^*\psi(W) \rangle,$$

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

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

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: It should be useful to modify the spaces $L_F^2(Z)$ and $L_F^2(W)$. The function φ may be constrained to be in a subspace of $L_F^2(Z)$ and r_F may be considered as an element of a larger space 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 checked in Appendix B.

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

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

where $r_F(w) = \int y f(y \mid w) dy$.

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 estimation. 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, [26, 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 impact on the resulting speed of convergence of the estimator of φ. Only matters the dimension of the vector Z of variables which enter into φ.

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 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 for the statisticians and it corresponds to the notion of complete statistic[2] (see [32, Lehmann, Scheffe (1950)], [5, Basu (1955)]). A systematic study is made in [22, 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 [36, Newey, Powell (2000)]. They also discussed the two particular cases detailed in examples 2.2 and 2.3 below. Actually the strong identification

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

assumption can be interpreted as a nonparametric rank condition as it is shown in the following example dealing with the normal case.

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

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

(see [23, Florens, Mouchart, Rolin (1993)]). 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, then it is the unique instrumental regression.

Example 2.3: 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 this identification condition, it can be checked in specific models (see e.g. [6, Blundell, Chen, Powell (2001)]). This identification condition can also be interpreted in terms of operators related to T_F as shown by the following corollary[3].

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 ensures 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 [31, Lancaster (1968)], [15, Darolles, Florens, Renault (1998)]). Therefore there is a sequence of 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:

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

i) φ_i, $i \geq 0$, is an orthonormal sequence of $L^2_F(Z)$ (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 $L^2_F(W)$.

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) $\forall g \in L^2_F(Z)$, $g(z) = \sum_{i=0}^{\infty} \langle g(Z), \varphi_i(Z) \rangle \varphi_i(z) + \bar{g}$, where $\bar{g} \in \mathcal{N}(T_F)$ and $T_F[g(Z)](w) = \sum_{i=0}^{\infty} \lambda_i \langle g(Z), \varphi_i(Z) \rangle \psi_i(w)$,

vii) $\forall g \in L^2_F(W)$, $g(w) = \sum_{i=0}^{\infty} \langle g(W), \psi_i(W) \rangle \psi_i(w) + \bar{g}$, where $\bar{g} \in \mathcal{N}(T^*_F)$ and $T^*_F[g(W)](z) = \sum_{i=0}^{\infty} \lambda_i \langle g(W), \psi_i(W) \rangle \varphi_i(z)$

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

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

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 regression of $g(Z)$ on the nonlinear functions of W. In other words we are looking for the best nonlinear instruments (see the Best Nonlinear Two Stage Least Squares by [3, 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.

The strong identification assumption of Z by W can be characterized in terms of the singular values decomposition of T_F.

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 in the regular normal case, it could be checked that 0 is an eigenvalue of $T_F T^*_F$ if $\dim W > \dim Z$. The strong identification assumption corresponds to $\lambda_i > 0$ for any i and it characterizes a strong dependance 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 A 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 [18, Florens, Heckman, Meghir, Vytlacil (2001)]), but this case will not be analyzed here. The general identification condition given in Proposition 2.1 and Corollary 2.1 remains true, but the characterization in terms of spectral decomposition cannot be used directly in the subsection since Assumption $A.1$ is clearly not fulfilled.

Nethertheless, in order to be able to apply Assumption $A.1$, we consider only the particular additive case[4] characterized by: $\exists \nu \in L_F^2(Z)$ and $\mu \in L_F^2(X)$ such that $\varphi(Z, X) = \nu(Z) + \mu(X)$. Under this assumption model (2.7) is clearly not identifiable: an arbitrary constant may be added to ν and substracted to μ. In order to eliminate this problem we assume that μ is subject to a location constraint, e.g. $\mu(0) = 0$ or $E(\mu) = 0$.

An obvious consequence of the additive structure of our problem is that the parameters ν and μ are identifiable if and only if the equation $\nu(z) + \mu(x) = 0$ implies $\nu = 0$ and $\mu = 0$. A very general identification condition follows from the following concept (see [22, Florens, Mouchart, Rolin (1990), chapter 5]).

Definition 2.3 *: Two random vectors U and V are measurably separable (or variation free) if any function of U a.s. equal to a function of V is a.s. constant.*

We can now state the result concerning the uniqueness of the previous decomposition.

Proposition 2.2 *: ν and μ are identifiable if and only if:*

i) $E[\nu(Z) - E[\nu(Z) \mid X] \mid X, W] = 0 \Longrightarrow \nu = 0$ *a.s.*

ii) Z and X are measurably separable.

Up to now we simplify our presentation by assuming that Z and W have no element in common and that Assumption $A.1.$ is satisfied. The estimation of this additive model with endogenous variables combines inversion of regularized integral operator and backfitting. The extension of the methodology proposed in this paper is work in progress.

[4] Such an additive structure is often put forward to deal with the curse of dimensionality.

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. [46, Wahba (1973)], [35, Nashed, Wahba (1974)], [44, Tikhonov, Arsenin (1977)], [24, Groetsch (1984)], [30, Kress (1998)] or for econometric applications [9, Carrasco, Florens (2000)] and [17, Florens (2000)]). In finite dimension linear operators are continuous, but this property disappears in infinite dimension. In particular, if L is a continuous one to one compact operator from G to G, L is onto only if G has a finite dimension. 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:

$$\mathcal{F}^0 = \{F \in \mathcal{F} : E[Y \mid W = w] \in \mathcal{R}(T_{_F}) \text{ and } \mathcal{N}(T_{_F}) = \{0\}\}.$$

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

$$\varphi = T_{_F}^{-1} r_{_F}. \tag{3.2}$$

Under Assumption *A*.1. the function φ can be computed using Fourier decomposition (**??**) of any function belonging to $L^2_{_F}(Z)$. 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 $\langle r_{_F}, \psi_i \rangle = E[r_{_F} \psi_i] = E[Y \psi_i(W)]$. 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 usual *cdf*, and in particular usual estimators of F, are not in $\mathcal{F}^0$ because one or the two properties characterizing $\mathcal{F}^0$ are not satisfied. If $\hat{F}_N$ is an estimator of F, $\hat{F}_N$ will not be in $\mathcal{F}^0$ for any N even if the true F is in $\mathcal{F}^0$. Indeed $\mathcal{F}^0$ has an empty interior for the topologies insuring the convergence of $\hat{F}_N$ to F.

3.2 Generalized inverse

We replace equation (2.3) by:

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

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. [9, Carrasco, Florens (2000)]).

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

$$\mathcal{F}^* = \{F \in \mathcal{F} : E[Y \mid W = w] \in \mathcal{R}(T_F) + \mathcal{N}(T_F^*)\}.$$

For any F, $E[Y \mid W = w] \in \overline{\mathcal{R}(T_F)} + \mathcal{N}(T_F^*) = L^2_F(W)$ 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 [33, Luenberger (1969)][5]. Actually it is easy to check that the Moore-Penrose generalized inverse leads to solve the following equation which is implied by (2.3):

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

[5] An additional extension could be obtained using Picard's theorem (see e.g. [30, 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).

Example 3.1: Let us continue Example 2.2. 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 $b = \mu_Y - A\mu_Z$. This solution is actually unique for two reasons:

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

ii) $T_F^*T_F$ is one to one if *rank* Σ_{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: $\varphi(0)(1 - p(W)) + \varphi(1)p(W) = E[Y \mid W]$, 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: $E[Y \mid W]$ is an affine function of $p(W)$. Else the solution of (3.4) is obviously the solution of:

$$\begin{bmatrix} E(1-p(W))^2 & Ep(W)(1-p(W)) \\ Ep(W)(1-p(W)) & E(p(W))^2 \end{bmatrix} \begin{bmatrix} \varphi(0) \\ \varphi(1) \end{bmatrix} = \begin{bmatrix} E((1-p(W))Y) \\ E(p(W)Y) \end{bmatrix}.$$

3.3 Ill-posed problem regularization

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 [44, Tikhonov, Arsenin (1977)] or [30, Kress (1998), p. 279]). We need to define a regularized solution to our problem which satisfies a continuity condition[6].

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:

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

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

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 [36, 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 [44, Tikhonov, Arsenin (1977), chapter I]). 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:

$$\alpha\varphi^\alpha(z) + \int \varphi^\alpha(u)\, c(u,z)\, dz = \int y d(y,z) dy, \tag{3.10}$$

where:

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

and

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

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 [24, Groetsch (1984)] and the proof of Lemma A.3 in Appendix A).

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

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

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), $n = 1, ..., N$, *are i.i.d. samples of* (Y, Z, W).

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

Assumption A.4: *The error term is homoskedastic:* $Var[U \mid W] = \sigma^2$.

This assumption makes more user-friendly the asymptotic distribution of the estimators and is also not essential. 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 this paper, we concentrate our presentation to a Tikhonov regularization. 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. The inverse L^{-1} must be continuous in order to transform small perturbations of h into small perturbations of g. In our problem both L and h are unknown and estimated. 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:*

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

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

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

where $\hat{c}_N(z,w)$ and $\hat{d}_N(y,z)$ are the kernel estimators of $c(u,z)$ and $d(y,z)$ introduced in Subsection 3.3. This estimator can be computed in one step and it reduces to a finite dimensional inverse problem. The practical implementation of this computation is detailed in Appendix B. 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 follow from the asymptotic properties of the estimators of eigenvalues and eigenvectors (see [14, Darolles, Florens, Gouriéroux (1998)] for a statement of these 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 friendly-user, but we can generalize the approach and use any other nonparametric techniques (for a sieve approach, see [12, Chen, Shen (1998)]). The main point is the speed of convergence of the norms given for kernel smoothing by Assumptions A.5 and A.6.

4.2 Consistency and speed of convergence

Usual nonparametric estimation is essentially concentrated 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. [30, Kress (1995), ch. 15]). A natural type of convergence is square norm, i.e. in $L_F^2(Z)$.

Assumption A.5: $\forall \lambda \in L_F^2(Z)$, *it exists* $\rho \geq 2$ *such that:*

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

Assumption A.6: *It exists* $\rho \geq 2$ *such that:*

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

We show in Appendix B that standard regularity conditions on the true F and on φ imply that Assumptions A.6 and A.7 are satisfied. We concentrate our presentation on the implication of these assumptions on the properties of the estimator of the parameter of interest φ.

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

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

where φ^{α_N} is defined in (4.1). Then:

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

using:

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

and:

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

The above decomposition underlines 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. [24, Groetsch (1984)]) and recall that $\alpha_N \to 0$ implies $\|\varphi^{\alpha_N} - \varphi\| \to 0$ (see e.g. [30, Kress (1995), ch. 15]). Finally, we obtain the following theorem.

Theorem 4.1 *: Under Assumptions A.1-A.6,*

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

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* $\|\hat{\varphi}_N^{\alpha_N} - \varphi\| \underset{N\to\infty}{\to} 0$ *a.s.*.

A natural question concerns now the selection rule of α_N and h_N in order to optimize the speed of convergence of $\hat{\varphi}_N^{\alpha_N}$ to φ. The study requires some restrictions on the class of φ functions, which are necessary to control the speed of convergence to zero of the term $\|\varphi^{\alpha_N} - \varphi\|$.

Definition 4.2 *: Let* Φ_β, $\beta > 0$, *the subset of* $L^2_F(Z)$ *of functions* φ *such that:*

$$\|\varphi^{\alpha_N} - \varphi\|^2 = \alpha_N^2 \sum_{j=0}^{\infty} \frac{1}{(\alpha_N + \lambda_j^2)^2}\langle\varphi, \varphi_j\rangle^2 = O(\alpha_N^\beta). \tag{4.3}$$

The maximum value is $\beta = 2$, and is obtained if $\sum_{j=1}^{\infty} \frac{1}{(\alpha_N+\lambda_j^2)^2}\langle\varphi,\varphi_j\rangle^2$ converges. For example, any φ which is in a finite dimensional subspace generated by a finite subfamily of the φ_j's, is an element of Φ_2. This proves in particular that all the Φ_β sets are dense in $L_F^2(Z)$ under the identification condition.

A more interesting case correspons to $\beta = 1$. The previous property is then obtained if[7]:

$$\langle\varphi,\varphi_j\rangle^2 = \lambda_j^2 c_j, \tag{4.4}$$

where $\sum_{j=1}^{\infty} c_j < \infty$. Indeed, in that case, we get:

$$\begin{aligned}\alpha_N^2 \sum_{j=0}^{\infty} \frac{1}{(\alpha_N+\lambda_j^2)^2}\langle\varphi,\varphi_j\rangle^2 &= \alpha_N^2 \sum_{j=0}^{\infty} \frac{\lambda_j^2}{(\alpha_N+\lambda_j^2)^2} c_j \\ &\leq \frac{\alpha_N}{4}\sum_{j=0}^{\infty} c_j = O(\alpha_N),\end{aligned}$$

because $\left(\alpha_N+\lambda_j^2\right)^2 \geq 4\alpha_N\lambda_j^2$. Assumption (4.4) is in particular realized if:

$$\varphi \in \left\{\lambda \in L_F^2(Z) \text{ such that it exists } \mu \in L_F^2(Z) \text{ and } \lambda = T_F^*\mu\right\},$$

because $\langle\varphi,\varphi_j\rangle^2 = \langle T_F^*\mu,\varphi_j\rangle^2 = \lambda_j^2\langle\varphi,\psi_j\rangle^2$ and $\|\mu\|^2 = \sum_{j=0}^{\infty}\langle\mu,\psi_j\rangle^2$ is finite. This assumption is equivalent to $\varphi \in \mathcal{R}(T_F^*)$.

Theorem 4.2 *: Under Assumptions A.1-A.6, the condition $\varphi \in \Phi_\beta$ implies:*

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

and this speed of convergence is realized if:

$$\begin{aligned}\alpha_N &= k_1 N^{-\frac{1}{2+\beta}}, \\ h_N &= k_2 N^{-\frac{1}{2\rho}},\end{aligned}$$

where k_1 and k_2 are constant terms.

The proof of this theorem follows directly from Theorem 4.1 *i*). This result requires some comments.

i) The actual speed of convergence of $\|\hat{\varphi}_N^{\alpha_N} - \varphi\|^2$ is usually better than $N^{\frac{\beta}{2+\beta}}$. But this speed of convergence depends on the specific behavior of the Fourier coefficients $\langle\varphi,\varphi_j\rangle$ of φ in the basis of the φ_j and on the rate of convergence of the λ_j's. This rate characterizes the dependance between the instrumental and the endogenous variables. Moreover the proposed choice of α_N and h_N optimizes the speed of convergence of the right hand side

[7] A useful discussion with W. Newey helps us to clarify this argument.

majorization given in *i*) of Theorem 4.1. An optimal, but not very easy implementable, choice α_N and h_N should depends on φ, φ_j and λ_j. Some directions on this kind of selection of regularization parameters are proposed in [10, Carrasco, Florens (2000)] and [11, Carrasco, Florens (2001)]

ii) The obtained speed of convergence does not depend on the dimensions p and q, and is in the best case $N^{\frac{1}{2}}$, i.e. equal to the square root of the speed of convergence in the parametric case. In the general case $\beta = 1$, we obtain a speed of convergence of $N^{\frac{1}{3}}$. In any case, the speed of convergence is polynomial. The assumption $\varphi \in \Phi_\beta$ completes in our framework the usual regularity conditions on the functional parameters necessary to get Assumption A.6.

iii) The choice of the bandwidth h_N is not optimal for the nonparametric estimation of r_F and T_F. An optimal choice, in this sense, would give an h_N proportional to $N^{-\frac{1}{p+2\rho}}$, which is smaller than our bandwidth selection. Theorem 4.1 suggests an oversmoothing estimation of T_F and r_F before the resolution of the functional equation. However, if h_N is chosen proportional to $N^{\frac{1}{p+2\rho}}$, an optimal choice of α_N is $N^{-\frac{2\rho}{(p+2\rho)(2+\beta)}}$ and provides a slower speed of convergence (e.g. $N^{\frac{4}{15}}$ for $p = 1$, $\rho = 2$, $\beta = 1$ or $N^{\frac{2}{9}}$ for $p = 2$, $\rho = 2$, $\beta = 1$ smaller than $N^{\frac{1}{3}}$).

iv) The optimal speed of convergence provided by Theorem 4.1 may be seen as deduced as a trade off between variance and bias in the terms $\frac{1}{\alpha_N^2}\left(\frac{1}{Nh_N^p} + h_N^{2\rho}\right)\alpha_N^\beta$ and α_N^β (proportional to $\|\varphi^{\alpha_N} - \varphi\|^2$). The leading term of the convergence in then provided by the estimation of r_F and by the regularization. The speed of convergence is then identical if the operator T_F is known or if it is estimated. On the contrary, if h_N is chosen proportional to $N^{-\frac{1}{p+2\rho}}$ the leading variance term of the asymptotic behavior is the variance term of the estimation of the operator ($\frac{1}{\alpha_N^2 N h_N^p}$) and not of r_F.

4.3 Asymptotic distributions

The asymptotic distribution of our estimator, or of some transformations of this estimator, follows from the asymptotic distribution of the estimation of the right hand side of the equation $T_F\varphi = r_F$. More precisely the asymptotic behavior of $T^*_{\hat{F}_N} r_{\hat{F}_N} - T^*_{\hat{F}_N} T_{\hat{F}_N}\varphi$ will determine the distribution of $\hat{\varphi}_N^{\alpha_N}$ and the estimation error on T_F only introduce a bias term which may be cancelled by a suitable selection of the regularization parameters.

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

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

The previous convergence is a functional convergence in distribution in the Hilbert space $L^2_F(Z)$ (see e.g. [45, Van de Vaart, Wellner (1996)]). Appendix ... 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 ...

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

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 [13, Chen, White (1992)]), it is immediately checked that:

$$\sqrt{N}(\hat{\varphi}^\alpha_N - \varphi^\alpha - b^\alpha_N) \Longrightarrow \mathcal{N}(0, \Omega), \tag{4.5}$$

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 (4.5) is still a functional distributional convergence in the Hilbert space $L^2_F(Z)$.

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 (4.5) 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 approximatavely $\sigma^2 (T^*_F T_F)^{-1}$, which is the functional extension of $\sigma^2 (E(ZW')E(WW')^{-1}E(WZ'))^{-1}$. In the linear parametric case, T_F is replaced by $W'E(W'W)^{-1}E(W'Z)$ and T^*_F by its transposed matrix.

Let us now consider the case where α is a function of N such that $\alpha_N \to 0$ when N increases. As usual in nonparametric estimation, the functional convergence will be replaced by a punctual convergence and we consider inner products of $\hat{\varphi}^{\alpha_N}_N$ with given functions of $L^2_F(Z)$.

Theorem 4.3 *: Let $\zeta \in L_F^2(Z)$ such that:*

$$\frac{\|(\alpha_N I + T_F^* T_F)^{-1}\zeta\|}{\|T_F(\alpha I + T_F^* T_F)^{-1}\zeta\|} \sim O(1). \tag{4.6}$$

Under Assumptions A.1-A.6, if:

$$\frac{1}{\alpha_N^2}\left(\frac{1}{Nh_N^p} + h_N^{2\rho}\right) \to 0, \tag{4.7}$$

then

$$\sqrt{\upsilon_N(\zeta)}\langle \hat{\varphi}_N^{\alpha_N} - \varphi^{\alpha_N} - b_N^{\alpha_N}, \zeta\rangle \Longrightarrow \mathcal{N}\left(0, \sigma^2\right), \tag{4.8}$$

where

$$\upsilon_N(\zeta) = \frac{N}{\|T_F(\alpha I + T_F^* T_F)^{-1}\zeta\|^2} = \frac{N}{\sum\limits_{j=0}^{\infty} \frac{\lambda_j^2}{(\alpha_N + \lambda_j^2)^2}\langle \varphi_j, \zeta\rangle^2},$$

Assumption (4.6) is in particular true for any ζ element of a finite dimensional linear space generated by the φ_j's Assumption (4.7) is a little stronger than the assumption of theorem 4.2. The speed of convergence satisfies:

$$\upsilon_N(\zeta) \geq 4\alpha_N N\|\zeta\|,$$

thanks to $(\alpha_N + \lambda_j^2)^2 \geq 4\alpha_N\lambda_j^2$ and $\upsilon_N(\zeta) \to \infty$ because $\alpha_N^2 N \to \infty$. This speed of convergence depends on the behavior of α_N, on the shape of the eigenvalues λ_j's and on the specific choice of ζ. The dimensions of Z and W do not explicitly appears. However the dimension p of Z appears in condition (4.7) which determines the rate of decrease of α_N and the rate of decrease of the λ_j's which measure the dependance between Z and W depends and the two dimensions. For example it could be checked that the Hilbert Schmidt norm of T_F or T_F^*:

$$\sum_{j=0}^{\infty}\lambda_j^2 = \int \frac{f^2(z,w)}{f^2(z)f^2(w)} f(z)f(w)dzdw, \tag{4.9}$$

decreases if the dimension of Z increases or if the dimension of W decreases.

The last question of interest we consider in this section concerns the behavior of the bias term in (4.8) 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.:

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

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

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

under Assumption A.6. Consider for simplicity the case $p=1$, $\rho=2$, and $\beta=1$. Under the choice $h_N=k_2N^{-(\frac{1}{4}+\varepsilon)}$ and $\alpha_N=k_2N^{-\frac{1}{3}}$, this expression converges to zeros when $N\to\infty$.

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

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

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 one of the main result concerns the asymptotic normality of the regularized solution of an ill-posed inverse problem. We obtain this normality but with a speed of convergence depending on the behavior of a parameter and on the decreasing rate of eigenvalues of the operator $T_F^*T_F$ which captures the dependance between explanatory variables Z and instruments W. 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;

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 couse 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 [7, Blundell, Powell (1999)]);

iv) An estimation of particular functionals 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 [18, Florens, Heckman, Meghir , Vytlacil (2001)] or [19, Florens, Larribeau (1995)]);

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

vi) 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 $\int \left(\hat{\varphi}_N^{\alpha_N} - m_N\right)^2 \pi(z)dz$ (where π is a suitable weighing 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.

A Proofs

A.1 Proof of Corollary 2.1

i) $\Longleftrightarrow$ *ii*): We start with:

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

We have:

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

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

i) $\Longleftrightarrow$ *iii*): This property can be deduced from [22, Florens-Mouchart-Rolin (1990), theorem 5.4.3] or [33, Luenberger (1969), theorem 3 section 6.3] since $\mathcal{R}(T_F^*) = \mathcal{N}(T_F^*)^{\perp}$.

A.2 Proof of Proposition 2.2

We need to prove that the null space of:

$$(\nu, \mu) \rightarrow E[\nu(Z) + \mu(X) \mid X, W],$$

is reduced to zero. If

$$E[\nu(Z) + \mu(X) \mid X, W] = 0,$$

then we also have:

$$E[\nu(Z) + \mu(X) \mid X] = 0,$$

and the difference of the two equalities gives:

$$E[\nu(Z) \mid X, W] = E[\nu(Z) \mid X].$$

The implication gives $\nu(Z) = E[\nu(Z) \mid X]$. We use the measurable separability assumption to obtain: $\nu(Z) = a$ a.s. and $\mu(X) = a$, which gives $\nu = \mu = 0$ using the location constraint.

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 $\zeta \in L^2_Z$, we have:

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

where

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

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

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

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

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:

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

Then we have:

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

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

A.4 Discussion of Assumptions A.5 and A 6

We introduce the following assumptions to obtain usual rate of convergence and a functional central limit theorem from the properties of the kernel density estimate.

Assumption 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 2: *The probability density function f is continuous on $\mathcal{X}$.*

Assumption 3: *The probability density function f is bounded[8] from below by $\varepsilon > 0$.*

Assumption 4: *The kernels K_y, K_z, K_w are bounded, symmetric, of order[9] r, Lipschitzian, and satisfy $\lim_{\|u\|\to\infty} \|u\| K_y(u) = 0$, $\lim_{\|u\|\to\infty} \|u\|^p K_z(u) = 0$ and $\lim_{\|u\|\to\infty} \|u\|^q K_w(u) = 0$.*

Assumption 5: *The p.d.f. f is d-continuously differentiable on $\mathcal{X}$, and there is a b such that $\|f\|_\infty < b$ and $\|f^{(d)}\|_\infty < b$.*

Assumption 6: *As $N \to \infty$, $Nh_N^p \to \infty$, $Nh_N^{2r} \to 0$.*

Lemma A.1 *: Under Assumptions 1-6, $\forall \lambda \in L_F^2(Z)$, we get :*

$$\|(T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F)\lambda\|^2 = O_p\left(\left(\frac{1}{Nh_N^p} + h_N^{2\min(r,d)}\right)\|\lambda\|^2\right).$$

Proof. To be added ■

Lemma A.2 *: Under the identification assumption A.2 and the technical Assumptions A.1 and 1-6,*

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

Proof. To be added ■

[8] This assumption can be relaxed by using data dependent trimming. This is out of the scope of this paper.

[9] 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.

B Generalization allowing to consider non bounded densities

B.1 Motivation

The theory presented in the paper concerns the solution of the linear equation $T_F\varphi = r_F$ using the estimation of the function r_F and the linear operator T_F. We compare the solution $\hat{\varphi}_N^{\alpha_N}$ of the equation:

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

to the solution of

$$T_F\varphi = r_F.$$

This analysis is simplified if the functions belongs to an Hilbert space, and if the integral operator is an Hilbert Schmidt operator. The main work consists then to transform the convergence properties of $\|T^*_{\hat{F}_N} T_{\hat{F}_N} - T^*_F T_F\|^2_{HS}$ and $\|T^*_{\hat{F}_N} r_{\hat{F}_N} - T^*_{\hat{F}_N} T_{\hat{F}_N}\varphi\|^2$ into the convergence properties of $\|\varphi^{\alpha_N} - \varphi\|^2$. This is the mathematical part of the work. The statistical part consists to find the estimators $T_{\hat{F}_N}$ and $r_{\hat{F}_N}$, which satisfy the assumptions made in the mathematical part. We choose not to discuss the different smoothing alternatives, and use simple kernel smoothing.

As an introduction, let us examine the convergence properties of $\|T_{\hat{F}_N}\varphi - T_F\varphi\|$, where φ is a function of z and $T_F\varphi$ denotes $E[\varphi(Z) \mid W]$. It is sufficient to study the behavior of:

$$E\|T_{\hat{F}_N}\varphi - T_F\varphi\|^2 = \int (E(T_{\hat{F}_N}\varphi - T_F\varphi)(w))^2 f(w)dw.$$

The term $(E(T_{\hat{F}_N}\varphi - T_F\varphi))^2$ in the integral can be decomposed in a variance term:

$$\frac{1}{Nh_N^p}\int K^2(u)du \frac{1}{f(w)} V[\varphi(Z) \mid W = w],$$

and a bias term [....] to

$$\frac{1}{4}h_N^4\sigma_K^2\left[\frac{1}{f(w)}\int (\varphi(z) - E[\varphi(Z) \mid W])\left(\partial^2 f\right)dz\right],$$

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

Let us consider for example le variance term in the computation of $E\|T_{\hat{F}_N}\varphi - T_F\varphi\|^2$. It is equal to:

$$\frac{1}{Nh_N^p}\int K^2(u)du \int V[\varphi(Z) \mid W = w]dw.$$

This integral does not converge in the general case if f is not bounded from below by a strictly positive number. To circumvent this difficulty, a natural idea consists to modify the norm used to define the functional space. Let us introduce a new density function, denoted $h(w)$. We must now study the behavior of:

$$\int V[\varphi(Z) \mid W = w] \frac{h(w)}{f(w)} dw,$$

with the convention $f(w) = 0$ implies $h(w) = 0$. If we assume in addition that $\frac{h(w)}{f^2(w)}$ is bounded from above, the previous integral is bounded from above by:

$$\sup \frac{h(w)}{f^2(w)} \int V[\varphi(Z) \mid W = w] f(w) dw,$$

and then converges. We propose to generalize this approach to the problem studied in the paper.

B.2 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:

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

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

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

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

The two constants are then necessarily greater than one.

Remark B.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:

$$\int \frac{f^2(z,w)}{f^2(z)f^2(w)} g(z)f(w)dzdw < \infty. \tag{B.3}$$

Remark B.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 $(B.3)$ from the conditions $(B.2)$.*

Definition B.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 B.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..*

B.3 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 dominate case, T^*_F satisfies:

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

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:

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

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:

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

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

B.4 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:

$$\frac{K\left(\frac{w-w_i}{h_N}\right)}{\sum_i K\left(\frac{w-w_i}{h_N}\right)} \in L^2_H(W) \text{ and } \frac{K\left(\frac{z-z_i}{h_N}\right)}{\sum_i K\left(\frac{z-z_i}{h_N}\right)} \in L^2_G(Z)$$

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 B.

B.5 Asymptotic properties

We must check if the assumptions A.5 and A.6 are satisfied. This imposes some additional regularity assumptions. 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.:

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

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:

$$\frac{1}{Nh^p_N} \sum_i K\left(\frac{z-z_i}{h_N}\right) \frac{f(u,w_i)h(w_i)}{f^2(w_i)} - \int \frac{f(z,w)f(u,w)}{f^2(w)} h(w)dw, \tag{B.8}$$

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:

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

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

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

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

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

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:

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

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

C Numerical implementation

For the numerical implementation, it is more convenient to use a matrix formulation. First, we explicit the integral operator in terms of the *pdf* estimator $\hat{f}_N$. Definition (4.1) becomes:

$$\alpha_N \hat{\varphi}_N^{\alpha_N}(z) + \int \hat{\varphi}_N^{\alpha_N}(u)\, \hat{c}_N(u,z)\, du = \int y \hat{d}_N(y,z) dy, \tag{C.1}$$

where $\hat{c}_N(u,z)$ and $\hat{d}_N(y,z)$ can be expanded in replacing the *pdf* estimator $\hat{f}_N$ by it expression. Hence, we obtain:

$$\hat{c}_N(u,z) = \sum_{i=1}^{N}\sum_{n=1}^{N} a_{i,n} K_{h_{z,N}}(z_i - u) \frac{K_{h_{z,N}}(z_n - z)}{\sum_{l=1}^{N} K_{h_{z,N}}(z_l - z)},$$

where

$$a_{i,n} = \int \frac{K_{h_{w,N}}(w_i - w)\, K_{h_{w,N}}(w_n - w)}{\sum_{l=1}^{N} K_{h_{w,N}}(w_l - w)} dw, \tag{C.2}$$

for $i, n = 1, ..., N$, denotes the generic term of a $N \times N$ matrix A_N. Thus the integral in the left term of $(C.1)$ becomes:

$$\sum_{n=1}^{N} \frac{K_{h_{z,N}}(z_n - z)}{\sum_{l=1}^{N} K_{h_{z,N}}(z_l - z)} \sum_{i=1}^{N} a_{i,n} \xi_i,$$

where the terms:

$$\xi_i = \int \hat{\varphi}_N^{\alpha_N}(u)\, K_{h_{z,N}}(z_i - u)\, du, \tag{C.3}$$

correspond to the convolution of the solution $\hat{\varphi}_N^{\alpha_N}$ with the kernel $K_{h_{z,N}}$ at point Z_i. The same approach can be used with the right hand term of equation $(C.1)$. If we replace *pdf* estimator $\hat{f}_N$ by it expression, we obtain:

$$\hat{d}_N(y,z) = \sum_{i=1}^{N}\sum_{n=1}^{N} a_{i,n} K_{h_{y,N}}(y_i - y) \frac{K_{h_{z,N}}(z_n - z)}{\sum_{l=1}^{N} K_{h_{z,N}}(z_l - z)},$$

where $a_{i,n}$ is defined in $(C.2)$. Thus, the right term of $(C.1)$ becomes:

$$\sum_{n=1}^{N} \frac{K_{h_{z,N}}(z_n - z)}{\sum_{l=1}^{N} K_{h_{z,N}}(z_l - z)} \sum_{i=1}^{N} Y_i a_{i,n},$$

using the simplification $\int y K_{h_{y,N}}(y_i - y) = y_i$ coming from the properties of kernel K_0. After these simplifications, equation $(C.1)$ can be multiplied by $K_{h_{z,N}}(z_j - z)$ and then integrated with respect to z. Hence, we obtain:

$$\alpha_N \xi_j + \sum_{i=1}^{N}\sum_{n=1}^{N} a_{i,n} b_{n,j} \xi_i = \sum_{i=1}^{N}\sum_{n=1}^{N} a_{i,n} b_{n,j} Y_i,$$

where

$$b_{n,j} = \int \frac{K_{h_{z,N}}(z_n - z)\, K_{h_{z,N}}(z_j - z)}{\sum_{l=1}^{N} K_{h_{z,N}}(z_l - z)} dz, \tag{C.5}$$

for $n, j = 1, ..., N$, denotes the generic term of a $N \times N$ matrix B_N. We denote by $c_{i,j}$ the sum $\sum_{n=1}^{N} a_{i,n} b_{n,j}$. This is the generic term of a $N \times N$ matrix C_N which is equal by definition to $A_N \times B_N$. We obtain:

$$\alpha_N \xi_j + \sum_{i=1}^{N} c_{i,j} \xi_i = \sum_{i=1}^{N} c_{i,j} Y_i,\ j = 1, ..., N. \tag{C.6}$$

The previous system of N equations can be written in a matrix form using the $N \times N$ matrix C_N. If we denote by I_N the $N \times N$ identity matrix and by Y_N the $N \times 1$ vectors of Y_i, the $N \times 1$ vector ξ_N such that each component ξ_j is solution of $(C.6)$ is obtained in solving the following system:

$$(\alpha_N I_N + C_N^{'})\xi_N = C_N^{'} Y.$$

For a suitable choice of the parameter α_N, the matrix $\alpha_N I_N + C_N^{'}$ is invertible, and the solution ξ_N of $(C.6)$ can be computed using the following formula:

$$\xi_N = (\alpha_N I_N + C_N^{'})^{-1} C_N^{'} Y_N.$$

The last step consists now to recover $\hat{\varphi}_N^{\alpha_N}$ from ξ_N. As it is emphasized, each term $\xi_{j,N}$ corresponds to the convolution product of $\hat{\varphi}_N^{\alpha_N}$ by the kernel K_z. A solution would be to use deconvolution techniques. In our particular problem, we use a different approach. We start from equation $(C.1)$ and we remark that the solution $\hat{\varphi}_N^{\alpha_N}(z)$ can be expressed as follow:

$$\alpha_N \hat{\varphi}_N^{\alpha_N}(z) = \sum_{n=1}^{N} \frac{K_{h_{z,N}}(z_n - z)}{\sum_{l=1}^{N} K_{h_{z,N}}(z_l - z)} \left[\sum_{i=1}^{N} y_i a_{i,n} - \sum_{n=1}^{N} a_{i,n} \xi_i) \right],$$

which gives the last expression for the estimated instrumental regression function:

$$\hat{\varphi}_N^{\alpha_N}(z) = \frac{1}{\alpha_N} \sum_{n=1}^{N} \beta_{h_{z,N}}(z_n - z) \sum_{i=1}^{N} a_{i,n} (y_i - \xi_i). \tag{C.7}$$

where

$$\beta_{h_{z,N}}(z_n - z) = \frac{K_{h_{z,N}}(z_n - z)}{\sum_{l=1}^{N} K_{h_{z,N}}(z_l - z)},$$

corresponds to a weight term. To a practical point of view, we need first to compute the N terms $\hat{\xi}_N$ in inverting the $N \times N$ matrix $\alpha_N I_N + C_N^{'}$. Then, we can implement the formula $(C.7)$ to compute the estimator $\hat{\varphi}_N^{\alpha_N}$ for any z. The remaining difficulties concern the computation of the terms

appearing in both matrices A_N and B_N. From equations $(C.2)$ and $(C.5)$ we see that the computation of each term $a_{i,n}$ and $b_{n,j}$ of these matrices requires an integration. If we use gaussian kernel, we need to numerical compute each integral, which is time consuming and generate computation error.

An alternative approach consists in approximating the terms ξ_i and $a_{i,n}$ defined in $(C.2)$-$(C.3)$ and appearing in $(C.7)$. Indeed, we have:

$$\int f(z) K_{h_{z,N}}(z_i - z) dz \backsimeq f(z_i),$$

for any function f defined on $\mathbf{R}^p$, when the bandwidth h_{1N} goes to zero. In applying this approximation for the function $\hat{\varphi}_N^{\alpha_N}$, we obtain $\xi_i \backsimeq \hat{\varphi}_N^{\alpha_N}(z_i)$. Using the same trick, one can make the following approximation:

$$\int g(w) K_{h_{w,N}}(w_n - w) dw \backsimeq g(w_n),$$

for any function g defined on $\mathbf{R}^q$, when the bandwidth $h_{w,N}$ goes to zero. We apply this in $(C.2)$ to the function $K_{h_{w,N}}(w_i - w) / \sum_{l=1}^N K_{h_{w,N}}(w_l - w)$ and we obtain:

$$a_{i,n} = \frac{K_{h_{w,N}}(w_i - w_n)}{\sum_{l=1}^N K_{h_{w,N}}(w_l - w_n)},$$

which is denoted in the following $\beta_{h_{w,N}}(w_i - w_n)$ since it corresponds to weight term. Hence, we obtain the following formula for our instrumental regression function estimator:

$$\hat{\varphi}_N^{\alpha_N}(z) = \frac{1}{\alpha_N} \sum_{n=1}^N \beta_{h_{z,N}}(z_n - z) \sum_{i=1}^N \beta_{h_{w,N}}(w_i - w_n)(y_i - \hat{\varphi}_N^{\alpha_N}(z_i)).$$

The numerical implementation of such solution requires the numerical inversion of an N-dimensional matrix.

References

[1] Aït-Sahalia, Y. (1995), *The Delta and Bootstrap Methods for Nonparametric Kernel Functionals,* Discussion Paper, MIT.

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

[3] 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.

[4] Basmann, R.L. (1959), *A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equations*, Econometrica **25**, 77-83.

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

[6] Blundell, R., Chen X. and J., Powell (2001),

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

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

[9] Carrasco, M., and J.P., Florens (2000), *GMM in Continuous Time,* Econometric Theory, **16**, 797-834.

[10] Carrasco, M., and J.P., Florens (2000), *Efficient GMM Estimation Using the Empirical Characteristic Function,* Discussion Paper, GREMAQ.

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

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

[13] 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.

[14] Darolles, S., Florens, J.P., and C., Gouriéroux (1998), *Kernel Based Nonlinear Canonical Analysis*, Discussion Paper CREST 9858.

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

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

[17] 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.

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

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

[20] Florens, J.P., Mouchart, M., and J.M., Rolin (1974), *Bayesian Inference in Error-in-variables Models*, Journal of Multivariate Analysis, **4**, 419-432.

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

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

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

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

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

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

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

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

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

[30] Kress, R. (1998), *Linear Integral Equations,* Springer.

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

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

[33] Luenberger (1969), *Optimization by Vector Space Methods*, Wiley, New York.

[34] Malinvaud (1970),

[35] 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.

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

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

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

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

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

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

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

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

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

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

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