Working Paper Series n° 618

June, 2010

"Non Parametric Models with Instrumental Variables"

Jean-Pierre FLORENS

Institut d'Economie Industrielle (IDEI) – Manufacture des Tabacs
Aile Jean-Jacques Laffont – 21, allée de Brienne – 31000 TOULOUSE – FRANCE
Tél. + 33(0)5 61 12 85 89 – Fax + 33(0)5 61 12 86 37 – **www.idei.fr** – contact@cict.fr

INSTITUT D'ECONOMIE INDUSTRIELLE

Non Parametric Models with Instrumental Variables

Jean-Pierre Florens*

June 2010

Abstract

This paper gives a survey of econometric models characterized by a relation between observable and unobservable random elements where these unobservable terms are assumed to be independent of another set of observable variables called instrumental variables. This kind of specification is usefull to address the question of endogeneity or of selection bias for example. These models are treated non parametrically and in all the example we consider the functional parameter of interest is defined as the solution of a linear or non linear integral equation. The estimation procedure then requires to solve a (generally ill-posed) inverse problem. We illustrate the main questions (construction of the equation, identification, numerical solution, asymptotic properties, selection of the regularization parameter) by the different models we present.

1 Introduction

Most of the econometric model take the form of a relation between a random element Y and two others random elements Z and U. Both Y and Z are observable (we have

*Toulouse School of Economics (IDEI and GREMAQ)

for example an *i.i.d.* sample $(y_i, z_i) i = 1, ..., n$ of (Y, Z)) but U is unobservable. In econometrics U may be view as a summary of all the missing variables of the model. The form of the relation may vary. Consider for example the three following cases:

i) $Y = \langle Z, \varphi \rangle + U$ where $\langle Z, \varphi \rangle$ denotes a scalar product between Z and a parameter φ (Z and φ may be infinite dimensional)

ii) $Y = \varphi(Z) + U$ where φ is an unknown function of Z.

iii) $Y = \varphi(Z, U)$ where φ is an unknown function of Z and U and is assumed to be increasing w.r.t. U.

The two first cases are said separable and the last one is non separable. We will say that Z is exogenous if the object of interest (the function φ) is characterized by an independence condition between Z and U. In the first case this condition reduces to a non correlation condition $E(ZU) = 0$ and $\langle Z, \varphi \rangle$ is the linear regression of Y relatively to Z. In the second case a mean independence condition $E(U|Z) = 0$ is assumed and φ is equal to the conditional expectation of Y given Z. In the last case it is usually assume that U and Z are fully independent. If moreover U is uniformly distributed between 0 and 1, $\varphi(Z)$ is the quantile function of Y given Z. At least in the two last cases, the exogeneity condition means that φ is determined by the conditional distribution of Y given Z.

As economics is not in general an experimental science, the exogeneity assumption creates an analogous statistical framework to treat economic data as in an experimental context. Essentially the econometrician mays treat the observations of Z as if they were fixed by an experimentalist and the mechanism generating the Z may be neglected in the estimation process of φ. This concept of exogeneity is fundamental in econometrics and has been analyzed from the beginning of econometric's researchs (see Koopmans and Reiersol (1950)) or more recently in connection to the concept of cut in statistical model (see Engle et al (1993), Florens and Mouchart (1985)...).

However in many important applications of statistics to economic data an exogeneity assumption is not valid in the sense that its not characterizes the parameter of interest. The elementary following example illustrates this point : assume Y and Z real and we are interested by the parameter φ of a linear relation $Y = \varphi Z + U$.

The variable Z is generated according to an equation $Z = \gamma W + V$ where W is an observable variable and V is an unobservable noise correlated with U. In that case $E(ZU) \neq 0$. There exists a parameter β such that $E((Y - \beta Z)Z) = 0$ but this parameter is different from φ.

We say that Z is endogenous if Z is not exogenous. This definition is not operational and should be precized in order to lead to a characterization of the parameter of interest.

The endogeneity of the Z variable may be illustrated by the notion of treatment model which is not specific to econometrics but which is very useful to motivate the interest to endogenous variables. Consider for example a deterministic variable $\zeta \in \mathbb{R}$ representing the level of a treatment and Y is a random element denoting the outcome of the treatment. Let us assume that the impact of the treatment ζ on Y may be formalized by a relation $Y = \varphi(\zeta) + U$ where $\varphi(\zeta)$ represents the mean effect of a level of treatment equal to ζ (i.e. $E(U) = 0$). In a non experimental design the level of the treatment Z assigned to an individual is not randomly determined but may depend on some characteristics of the patient observable by the person who fix the treatment but not by the statistician. In that case the model used by the statistician is $Y = \varphi(Z) + U$ but the assumption $E(U|Z) = 0$ is not relevant.

This example may be extended to macro econometric analysis. The aggregated consumption of some good may be written $Y = \varphi(\pi) + U$ where π is a fixed non random value of the price of this good. The function φ is in that case the average aggregated demand function. The observed price P is not at all randomly generated and follows for example from the equilibrium of a system of demand and supply (the supply verifies $S = \psi(\pi) + V$ and the statistician observes Y and P such that $Y = S$ or $\varphi(P) + V = \psi(P) + V$). In this situation the model becomes $Y = \varphi(P) + U$ but $E(U|P) \neq 0$.

In most of the case Z is endogenous because it is not fixed or randomized but is generated including a strategic component of the economic agents or Z follows from an equilibrium rule among the economic agents.

The three models we have introduced before are not well defined if we eliminate the independence assumption between Z and U. These assumptions should be replace by other assumption in order to characterize the function φ.

The more natural extension to models with exogenous variables is provided by models with instrumental variables (IV). We consider now three random elements (Y, Z, W) where W are the instruments and the model is still specified by a relation linking Y to Z and U but U is now assumed to verify an independence property with W and not with Z. This approach extends obviously the exogeneity case because W and Z may be taken equal but the interest of this framework is to separate the relevant variables in the model (Z) and the variables independent to the residual (W). In a general presentation Z and W may have common elements but contain specific variables. In the three models presented above the independence conditions now become $E(WU) = 0, E(U|W) = 0$ or $U \perp\!\!\!\perp W$ (U and W independent).

The IV approach is not the unique way to formalize the endogeneity condition. In separable models, we may introduce a control function approach.

Consider for example the second type of model and let us compute the conditional expectation $E(Y|Z, W) = \varphi(Z) + E(U|Z, W)$. A control function approach is based on the assumption that there exist a function $C(W, Z)$ such that $E(U|Z, W) = E(U|C) = \psi(C)$ and such that C is sufficiently separated of Z to allow the identification of the two components of the additive model $Y = \varphi(Z)+\psi(C)+\varepsilon$. For example we may assumed that $\frac{\partial}{\partial Z}\psi(C) = 0$. In that case $\varphi(Z)$ is obtained up to an additive constant by solving the equation $E(Y|Z, C) = \varphi(Z) + \psi(C)$ (see Newey et al. (1999) or Florens et al.(2008)).

In this paper we focus our attention on the instrumental variables approach in a non parametric context. This question has generated numerous researches in the last ten years in econometrics and this paper is just a survey of the main elements of this literature. The goal is to present the key points through different examples.

The strategy to examine this question is the following. First we derive from the independence condition between U and W a functional equation which link the unknown object of interest φ and the probability distribution of (Y, Z, W) (actually the conditional distribution of Y, Z given W). Under the hypothesis of correct specification we assume that a solution of this equation exists. The second question is the unicity or local unicity of this solution, or, in econometric terminology, the question of identification or local identification. This unicity property usually requires some dependence condition between the Z and the W variables. In the third step, we use the equation derived from the independence between U and W

to estimate φ. We replace the distribution of (Y, Z, W) by a non parametric estimate and we estimate φ as the solution of the estimated equation. Unfortunately this simple approach based on the resolution of an estimated functional equation belongs in general to the class of ill-posed inverse problem and this naive solution is not a consistent estimator. This difficulty is solved by a penalization technic and we essentially consider in this paper L^2 penalizations. The final element consists to examine the asymptotic properties of the estimator and to derive in particular its rate of convergence to the true function. This rate will basically depend upon the difficulty of the resolution of the equation (”degree of ill-posedness”) and of the regularity of φ relative to the problem (”degree of smoothness”). In the IV case the degree of ill-posedness is related to the dependence between the Z and the W. Intuitively low dependence means high degree of ill-posedness.

The penalized resolution of the equation requires the choice of some regularity parameter and the search of a data driven selection of this parameter is essential for the implementation of this approach. A comparison between a feasible estimator based on the data driven selection of the regularization parameter and a theoretical unfeasible estimator based on an optimal selection of the regularization parameter is important and may be conducted in the spirit of ”oracle” inequalities”. This last point will not be treated in the paper (see Cavalier (2010) in this volume or in a bayesian context Florens and Simoni (2010)).

This paper will review the instrumental variable analysis in the three kind of models which has been introduced. We also briefly introduce the extension to some dynamic models of the previous ideas, essentially developed a in static framework.

2 The linear model: vectorial or functional data

Let us start to recall the elementary model of instrumental variables which reduces to the well known two stages least squares method in the homoscedastic case.

We consider a random vector (Y, Z, W) where $Y \in \mathbb{R}, Z \in \mathbb{R}^p$ and $W \in \mathbb{R}^q$ (Z and W may have common elements) and the model verifies:

$$\begin{cases} Y = Z'\beta + U \\ E(WU) = 0 \end{cases} \tag{2.1}$$

where $\beta \in \mathbb{R}^p$ is the parameter of interest.

The condition (2.1) leads to the equation.

$$E(WZ')\beta = E(WY) \tag{2.2}$$

denoted $T\beta = r$ with T is a matrix operator from $\mathbb{R}^p$ to $\mathbb{R}^q$ and r is an element of $\mathbb{R}^q$. This system of linear equations is assumed to have a solution (well specification of the model) and this solution is unique (identification condition) if T is one to one, i.e. if $E(ZW')$ has a rank equal to p (which needs in particular $q \geq p$). This system is solve through the minimization of

$$\|T\beta - r\|^2 \tag{2.3}$$

where the norm is the euclidian norm in $\mathbb{R}^q$ and the solution is

$$\beta = (T^*T)^{-1}T^*r \tag{2.4}$$

where T^* denote the transpose of T.

We assume that an *i.i.d.* sample $(y_i, z_i, w_i)_{i=1,\dots,n}$ is available and the estimation of β is obtained by the replacement of T, T^* and r by their empirical counterparts:

$$\hat{\beta} = \left[\left(\frac{1}{n}\sum_{i=1}^{n} z_i w_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n} w_i z_i'\right)\right]^{-1}\left(\frac{1}{n}\sum z_i w_i'\right)\left(\frac{1}{n}\sum w_i y_i\right) \tag{2.5}$$

This estimator is not optimal in terms of its asymptotic variance. To find an optimal estimator we may start again from the moment condition $E(W(Y - Z'\beta)) = 0$ and the usual results (see Hansen (1982)) on GMMM (Generalized moments Method) implies that optimal estimation is deduced from the minimization of

$$\|B(T\varphi - r)\|^2 \text{ where } B = [Var(WU)]^{-\frac{1}{2}} \tag{2.6}$$

This minimization gives

$$\beta = (T^*B^*BT)^{-1}T^*B^*Br \tag{2.7}$$

If $Var(U|W) = \sigma^2$ (homoscedastic case) B^*B reduces to $[Var(W)]^{-1}$ or, using the empirical counter parts of these operators, we have:

$$\hat{\beta} = \left[\left(\frac{1}{n}\sum_{i=1}^{n} z_i w_i\right)\left[\left(\frac{1}{n}\sum_{i=1}^{n} w_i w_i'\right)]\right]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} w_i z_i'\right)\right]^{-1}\frac{1}{n}\sum_{i=1}^{n} z_i w_i\left(\frac{1}{n}\sum_{i=1}^{n} w_i w_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} w_i y_i \tag{2.8}$$

This estimation is consistent and verify

$$\sqrt{n}(\hat{\beta} - \beta) \Rightarrow N(0, \sigma^2 (T^* B^* B T)^{-1}) \tag{2.9}$$

This computation requires the inversion of two matrix operators and it is natural to consider questions coming from the possible ill conditioning of these matrix. The inversion of $Var(W)$ may be difficult if the dimension of W becomes large, in particular the inversion of the estimator of $VarW$ if the sample size is small compared to q the dimension of W. This difficulty may be solved by a regularization of the inversion of this variance and $\frac{1}{n}\Sigma w_i w_i'$ may be replace by $\alpha I + \frac{1}{n}\sum w_i, w_i'$ where α is a positive parameter going to 0 when $N \to \infty$. (see Carrasco, Florens (2000), Carrasco (2008)).

An other question come from the rank condition on $E(WZ')$ which determines the identification condition. A recent literature on the so called "weak instruments" (See a survey by Stock et al (2002)) consider cases where rank $(\frac{1}{n}\sum_{i=1}^{n} w_i z_i') = p$ but where this matrix converges to a non full rank matrix. The correct mathematical formalization of this situation is not very easy if the dimension of the vector W and Z are keeped fixed. This question is more easy to understand in the case where the dimension of Z and W are infinite.

The natural extension of the previous model (see Florens and Van Bellegem (2009)) considers $Y \in \mathbb{R}, Z \in \mathcal{F}$ and $W \in \mathcal{H}$ where $\mathcal{F}$ and $\mathcal{H}$ are two Hilbert spaces.

The model now becomes:

$$\begin{aligned} &Y = \langle Z, \varphi \rangle + U \qquad \varphi \in \mathcal{F} \langle , \rangle \text{ scalar product in } \mathcal{F} \\ &E(WU) = 0 \end{aligned} \tag{2.10}$$

where φ is the functional parameter if interest and where the condition (2.11) involves an expectation in the space $\mathcal{H}$. For example if $\mathcal{F}$ is the L^2 space of square integrable functions defined on [0,1] w.r.t. the uniform measure we have

$$\langle Z, \varphi \rangle = \int_0^1 Z(t)\varphi(t)dt \tag{2.11}$$

or if $\mathcal{F}$ is ℓ^2 space of square sommable sequences w.r.t. a measure $(\pi_j)_{j=0,1...}$ we may have

$$\langle Z, \varphi \rangle = \sum_{j=0}^{\infty} Z_j \varphi_j \pi_j. \tag{2.12}$$

The functional equation determined by condition (2.10) is now rewritten

$$E(W\langle Z, \varphi \rangle) = E(WY) \tag{2.13}$$

or $T\varphi = r$ where T is the covariance operator from $\mathcal{F}$ to $\mathcal{H}$. We still assume the model well specified (a solution exists to (2.13)) and identified (T is one to one).

The equation $T\varphi = r$ which characterizes φ is now a Fredholm equation of type I and is ill-posed because the covariance operator T is compact. In that case the generalized inverse solution (2.4) is not a continuous function of r and then does not lead to a consistent estimator.

The resolution of $T\varphi = r$ is then an ill-posed linear inverse problem which has the particularity that not only r is estimated but that the operator T is also unknown and estimated using the same data set as r.

The estimation of φ will be performed using a regularization technic and we will concentrate here on the estimation by a Tikhonov regularization which may include a smoothness constraint.

Let $L : \mathcal{F} \to \mathcal{F}$ a differential operator defined on a dense subset of $\mathcal{F}$ and self adjoint. For example let us take the operator Ion $L^2[0, 1]$ defined by:

$$I\varphi = \int_0^t \varphi(s)ds \tag{2.14}$$

and let us define L by $L^{-2} = I^*I$. We easily see that $\varphi \in \mathcal{D}(L^{-b})$ is equivalent to say that φ is b differentiable and satisfies some boundary conditions (e.g. in our example $\varphi \in \mathcal{D}(L^{-2})$ means that φ is twice differentiable and $\varphi(0) = \varphi'(0) = 0$).

Let us assume that $\varphi \in \mathcal{D}(L^{-b})$ an consider $s \leq b$.

We consider the following Tikhonov functional

$$\|T\varphi + r\|^2 + \alpha \|L^s \varphi\|^2 \tag{2.15}$$

The minimum φ^α is equal to :

$$\begin{aligned}\varphi^\alpha &= (\alpha L^{2s} + T^*T)^{-1}T^*r \\ &= L^{-s}(\alpha I + L^{-s}T^*TL^{-s})^{-1}L^{-s}T^*r\end{aligned} \tag{2.16}$$

and the estimator is obtained by replacing T, T^* and r by their empirical counterparts $\hat{T}, \hat{T}^*$ and $\hat{r}$.

At least three questions follows from this estimation mechanism: is the estimator easily computable, what are its asymptotic properties in relation in particular to a concept of strong or weak instruments and is it possible to extend the optimality argument presented in the finite dimensional case to the functional linear model. The answers of these question are given in Florens and Van Bellegem (2009) and we just summarize here the main results.

Let us first remark that the computation of the estimator of φ reduces to a matrix computations. To illustrate this points consider the case where $s = 0$ and consider the system $(\alpha I + \hat{T}^*\hat{T})\varphi = \hat{T}^*\hat{r}$, or equivalently:

$$\begin{aligned}\alpha\varphi + \tfrac{1}{n}\sum_{i=1}^{n} z_i\langle w_i, (\frac{1}{n}\sum_{j=1}^{n} w_j\langle z_i, \varphi\rangle) > \\ = \tfrac{1}{n}\sum_{i=1}^{n} z_i\langle w_i, (\frac{1}{n}\sum_{j=1}^{n} w_j y_j)\rangle\end{aligned} \tag{2.17}$$

This equation is estimated in two steps. First we take the scalar products of the two sides of this equation with any z_l ($l = 1, ..., n$) and we derive a linear system of n equations where the n unknowns are $\langle\varphi, z_l\rangle$. In a second step we may compute φ everywhere using the equation (2.17) and the previous computation of these scalar products. Note that we have to invert are $n \times n$ systems and that we assume to observe the scalar products $\langle z_i, z_j\rangle$ or $\langle w_i, w_j\rangle$ (and not necessarily the complete continuous trajectoires of the sample of W and Z).

The second question concerns the speed of convergence of the estimator. The main result is summarized by

$$\|\hat{\varphi}^\alpha - \varphi\|^2 \sim O(n^{-\frac{\beta}{\beta+1}}) \tag{2.18}$$

where $\beta = \frac{b}{a(1-\gamma)}$. We have defined b as the smoothness hypothesis on φ. The number a (the degree of ill-posedness) is defined by the property

$$\|T\varphi\| \sim \|L^{-a}\varphi\| \tag{2.19}$$

Intuitively L^{-1} is an integral operator and T is equivalent in terms of norms to L^{-a}.

The notation (2.19) is a shortcut of the property $C_1\|L^{-a}\varphi\| \leq \|T\varphi\| \leq C_2\|L^{-a}\varphi\|$ for two suitable constants C_1 and C_2.

The final term γ is specific to statistical inverse problems (different from inverse problems treated in numerical analysis). We have introduced an error term U and the element WU of $\mathcal{H}$ is assumed to have a variance Σ which is a trace class operator from $\mathcal{H}$ to $\mathcal{H}$. Let us consider the singular values decompositions of T^*T characterized by the (non null) eigen values λ_j^2 and the eigen vectors φ_j. The parameter γ is defined by the largest value in $[0,1]$ such that

$$\sum_{j=1}^{\infty} \frac{\langle \Sigma\varphi_j, \varphi_j\rangle^2}{\lambda_j^{2\gamma}} < \infty \tag{2.20}$$

This property is trivially satisfied for $\gamma = 0$ because Σ is trace class which correspond to the worth speed of convergence.

The last point we may consider concerns the optimality of our method in term of the asymptotic variance. If we follows the result obtained in the final dimensional case we should weight the difference $\hat{T}\varphi - \hat{r}$ in the norm by $\Sigma^{-\frac{1}{2}}$. This is impossible because Σ is a compact non invertible operator. A second regularization is then needed and we proof that the estimator

$$\begin{aligned} \hat{\varphi}^{\alpha,\nu} \; = \; & \left(\alpha I + \hat{T}^*\hat{\Sigma}^{\frac{1}{2}}(\nu I + \hat{\Sigma})^{-2}\hat{\Sigma}^{\frac{1}{2}}\hat{T}\right)^{-1} \\ & \hat{T}^*\hat{\Sigma}^{\frac{1}{2}}(\nu I + \hat{\Sigma})^{-2}\hat{\Sigma}^{\frac{1}{2}}\hat{r} \end{aligned} \tag{2.21}$$

is optimal in a large class of estimator. All the elements of this class converges at the same rate and (2.21) has the best asymptotic variance. The study of this estimator is complex because it depends on two regularization parameters α and ν and because Σ is unknown and estimated. One of the basic results is that ν may be chosen such that the speed given in (2.18) is preserved.

3 The additively separable model and its extensions

We still consider a random vector $(Y, Z, W) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ and we define the instrumental regression by the following properties (See Florens (2000)):

$$\begin{cases} Y = \varphi(Z) + U \\ E(U|W) = 0 \end{cases} \tag{3.1}$$

if Z and W are identical (3.1) characterizes φ as the conditional expectation of Y given Z. The interest of the model comes from the case where Z and W are not identical and for simplicity we first assume that Z and W have no common element. If (Y, Z, W) has a density f, the model (3.1) implies that φ should satisfies the integral equation:

$$\int \varphi(z) f(z|w) dz = \int y f(y|w) dy$$

which may be denoted

$$T\varphi = r \tag{3.2}$$

The choice of the spaces and then of the operator T and of r should be precised. In general we consider $L^2(Y, Z, W)$ Hilbert space of square integrable functions w.r.t. the true data generating process and $L^2(Z), L^2(W)$... the sub spaces of Z dependent or W dependent random variables. In that case r is assumed to be an element of $L^2(W)$, φ an element of $L^2(Z)$ and T is the conditional expectation operator from $L^2(Z)$ into $L^2(W)$. In that case the adjoint operator $T^*(L^2(W) \rightarrow L^2(Z))$ is simply the conditional expectation operator:

$$T^*(\psi) = E(\psi(W)|Z) \quad \psi \in L^2(W)$$

The difficulty behind this approach is that we have to estimate both r and T and we don't know the distribution which characterizes the spaces. It may be easier to specifies two given Hilbert spaces $\mathcal{E}$ and $\mathcal{F}$ and to assume that T operates from $\mathcal{E}$ to $\mathcal{F}$ and that $r \in \mathcal{F}$.

This approach has been follow in particular in Florens et al (2005). In this presentation however we consider that the relevent spaces are of the form $L^2(Z)$....

The first question following from equation (3.2) is the identification of φ or equivalently the unicity of the solution. Due to the linearity of T it is obvious that φ is identified if $T\varphi = 0$ implies $\varphi = 0$.

This property is the injectivity of the conditional expectation operator and is a dependence condition between Z and W. It means that there does not exist a function of Z orthogonal to any function of W. This property has been introduced in statistics under the name "completness" and has been studied under the name "strong identification" (see Florens et al (1989) chap. 5). For joint normal distribution T is one to one if and only if the rank of the covariance matrix between Z and W is equal to the dimension of Z.

In the general case the singular value decomposition of T may be used to characterize the identification condition. This condition is true if 0 is not an eigen value of T^*T.

Actually the statistical analysis of our problem requires that we may characterized the speed of decline to zero of the SVD of T. This speed of decline measures the dependence between Z and W. As we did in the previous section a natural tool is provided by a measurement deduced from an Hilbert scale defined from a differential operator L. We then assume that $\|T\varphi\| \sim \|L^{-a}\varphi\|$ and a defines the degree of ill-posedness of the problem. We may also assume that the singular values of T $(\lambda_j)_{j=1,...}$ declines at a geometric rate $(\lambda_j \sim \frac{1}{j^a})$ see Hall and Horowitz (2005) or at an exponential rate (which is the case for a jointly normal distribution from (Z, W)) .

As usual for non parametric statistic we need also to assume some regularity for the function we want to estimate. The Hilbert scale approach gives such a definition of regularity : φ has the regularity b if $\varphi \in \mathcal{D}(L^b)$. We can also assume some rate of decline for the Fourier coefficient of φ in the basis of the eigen vectors of T^*T. An elementary case is obtained by choosing $L = (T^*T)^{-\frac{1}{2}}$ which implies that the degree of ill-posedness is equal to 1 and the condition $\varphi \in \mathcal{D}[(T^*T)^{-\frac{b}{2}}]$ (or equivalently $\varphi \in \mathcal{R}(T^*T)^{\frac{b}{2}}$) is called a source condition. All these considerations are very common in the theory of inverse problems and we just applied this methodology to the conditional expectation operator.

The general principle of the estimation of φ is to estimate the r value of (3.2) and the operator T by usual non parametric technics and to solve any regularized version of equation (3.2).

The estimations is obtained by estimating the first order condition of the minimization of the Tikhonov functional see Carrasco et al (2007).

$$\|T\varphi - r\|^2 + \alpha\|\varphi\|^2 \tag{3.3}$$

which leads to

$$\varphi^\alpha = (\alpha I + T^* T)^{-1} T^* r \tag{3.4}$$

and to an estimator:

$$\hat{\varphi}^\alpha = (\alpha I + \hat{T}^* \hat{T})^{-1} \hat{T}^* \hat{r} \tag{3.5}$$

which may computed by matrix inversion only (see Florens et al (2003)).

More general estimation are derived from iterated Tikhonov method or from a minimization in an Hilbert scale penalization:

$$\|T\varphi - r\|^2 + \alpha\|L^s\varphi\|^2 \tag{3.6}$$

where $s \leq \beta$ which leads to an estimator.

$$\hat{\varphi}^\alpha = L^{-s}(\alpha T + L^{-s}\hat{T}^*\hat{T}L^{-s})^{-1}L^{-1}\hat{T}^*\hat{r} \tag{3.7}$$

We will consider only the case where $s = 0$ (usual L^2 Tikhonov method).

Let us first discuss the non parametric estimation part. The rhs r may be estimated by a usual kernel approach:

$$\hat{r} = \frac{\sum_{i=1}^{n} y_i K\left(\frac{w-w_i}{h_n}\right)}{\sum_{i=1}^{h} K\left(\frac{w-w_i}{h_n}\right)} \tag{3.8}$$

where K is a kernel of suitable order an h_n the bandwidth. The estimation of T is done by replacing $f(z|w)$ by its kernel estimation

$$\hat{f}(z|w) = \frac{\frac{1}{h^p}\sum_{i=1}^{n} K\left(\frac{z-z_i}{h_n}\right) K\left(\frac{w-w_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{w-w_i}{h_n}\right)} \tag{3.9}$$

where for simplicity we denote by K and h_n the different kernels and bandwidths. Then T is estimated by

$$\hat{T}\varphi = \int \varphi(z)\hat{f}(z|w)dz \tag{3.10}$$

and T^* by:

$$\hat{T}^*\varphi = \int \psi(w)\hat{f}(w|z)dw \tag{3.11}$$

where $\hat{f}(w|z)$ is defined analogously. Notice that $\hat{T}$ not the dual of $\hat{T}$. It may be proved (see Darolles et al (2003) that:

$$\|\hat{r} - \hat{T}_\varphi\|^2 \sim O\left(\frac{1}{nh_n^q} + h_n^{2\rho}\right) \tag{3.12}$$

$$\|\hat{T} - T\| \sim \|\hat{T}^* - T^*\|^2 \sim O\left(\frac{1}{nh_n^{p+q}} + h^{2\rho}\right) \tag{3.13}$$

where ρ represents the regularity of the joint distribution of the data.

Note that an alternative estimation of T would be

$$\hat{T}\varphi = \frac{\sum_{i=1}^n \varphi(z_i)K\left(\frac{w-w_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{w-w_i}{h_n}\right)} \tag{3.14}$$

and equivalently for T^*. These estimations gives excellent approximations and excellent results in the simulation but as this operators are not bounded in the L^2 spaces the available proofs of consistency do not applied to these estimations (see Fève and Florens (2010)).

As in the linear case the computation of $\hat{\varphi}^\alpha$ reduces to a matrix computation, at least if approximation (3.14) is used. Indeed in that case we have to solve:

$$\begin{aligned} \alpha\varphi(z) \quad &+ \quad \frac{\sum_j \frac{\sum_i \varphi(z_i)K\left(\frac{w-w_i}{h_n}\right)}{\sum K\left(\frac{w_j-w_i}{h_n}\right)} K\left(\frac{z-z_j}{h_n}\right)}{\sum_j K\left(\frac{z-z_j}{h_n}\right)} \\ &= \quad \frac{\sum_j \frac{\sum_i y_i K\left(\frac{w_j-w_i}{h_n}\right)}{\sum K\left(\frac{w_j-w_i}{h_n}\right)} K\left(\frac{z-z_j}{h_n}\right)}{\sum_j K\left(\frac{z-z_j}{h_n}\right)} \end{aligned} \tag{3.15}$$

which is solved in two steps: first for $z = z_1, ..., z_n$ and after for any value of z.

The last question is to consider the asymptotic properties of these estimators. Let us focussed on the usual Tikhonov estimation. The difference $\hat{\varphi}^\alpha - \varphi$ may be decomposed in three points:

$$\begin{aligned} \hat{\varphi}^* - \varphi &= (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*(\hat{r} - \hat{T}\varphi) && I \\ &+ [(\alpha I + \hat{T}\hat{T})^{-1}\hat{T}^*\hat{T} - (\alpha I + \hat{T}^*T)^{-1}T^*T]\varphi && II \\ &+ \varphi^\alpha - \varphi && III \end{aligned}$$

The norm $\|\varphi^\alpha - \varphi\|^2$ is the regularization bias and is known to be $O(\alpha^{\frac{b}{\alpha}})$ in the Hilbert scale approach.

The norm of the first term I verifies

$$|I\|^2 \le \|(\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\|^2 \|\hat{r} - \hat{T}\varphi\|^2$$

$$\sim O\left(\frac{1}{\alpha}\left(\frac{1}{nh^p} + h^{2\rho}\right)\right)$$

The norm of II requires some computations but under some regularity assumption is term is negligible w.r.t. to the other term. If h is chosen by an optimal rule we have $h_n = n^{-\frac{1}{p+2\rho}}$ and $\|I\|^2 \sim O\left(\frac{1}{\alpha} n^{-\frac{2\rho}{p+2\rho}}\right)$.

The optimal choice for α is then

$$\alpha \text{ proportional to } [n^{-\frac{2\rho}{p+2\rho}}]^{\frac{a}{b+a}} \tag{3.16}$$

which gives an optimal rate of convergence:
$\|\hat{\varphi}^\alpha - \varphi\|^2 \sim O(n - \frac{2\rho}{-p+2\rho} \times \frac{b}{a+b}))$

In some cases (see Chen and Reiss (2007) or Johannes et al (2007)) it is natural to assume that $\rho = b + a$ and the optimal rate simplifies to:

$$\|\hat{\varphi}^\alpha\|^2 \sim O(n^{-\frac{2b}{2(b+a)+\rho}}) \tag{3.17}$$

which has been shown to be minimax under some assumptions.

The main question following from this approach is the empirical determination of the regularization parameters, namely the bandwidths of the kernel estimation and the α for the Tikhonov regularization.

Several approaches has been proposed in the literature. The following rule has been proved to have good properties, both theoretically and by simulation (see Engle et al (2000) or Feve and Florens (2010)).

The principle is to compute α which minimizes

$$\frac{1}{\alpha}\|\hat{r} - \hat{T}\hat{\varphi}^{\alpha}\|^2 \tag{3.18}$$

where the norm is replaced by the empirical norm.

Indeed the minimization of $\|\hat{r} - \hat{T}\hat{\varphi}^{\alpha}\|^2$ leads to $\alpha = 0$ and multiplying by $\frac{1}{\alpha}$ is equivalent to penalize this quantity. The α obtained by this rule has the optimal speed of convergence (for a comparable rule see Loubes and Marteau (2010) or for a bayesian approach Florens and Simioni (2010)) and numerous simulations shows its relevance.

This separable model has many extensions which may be treated in the same spirit.

First, for dimensionality reason, we may consider some restrictions on the general form $Y = \varphi(Z) + U$, for example:

i) $Y = \varphi_1(Z_1) + \varphi_2(Z_2) + U$ (additive model) $Y = \varphi_1(Z_1) + Z_2'\beta_2 + U$ (semi parametric additive model) where Z_2 may be exogenous (Z_2 included in W) or not

ii) $Y = \varphi(\beta' Z) + U$ (single index form)

This model does not lead to a linear integral equation. These models has been treated in many papers (see Florens et al (2005), Ai and Chen (2003)).

Secondly we may consider the class of transformation models like:

$$\varphi(Y) = Z'\beta + U$$

(see Fève and Florens (2010)) or

$$\varphi(Y) = \psi(Z) + X'\beta + U$$

(see Florens and Sokullu (2010) when X is exogenous).

Third we may consider some test problems like testing that Z is exogenous ($\varphi(Z) = E(Y|Z)$) or that φ has a given parametric form (see Horowitz (2006), Blundell and Horowitz (2007)).

4 The non separable models

The last family of models we consider in the static case belongs to the class of non separable models. Let us still consider a random vector $(Y, Z, W) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ and we assume the following relation:

$$\begin{aligned} & Y = \varphi(Z, U) \quad U \in \mathbb{R} \\ & \text{where } \varphi(Z, .) \text{ is strictly increasing.} \\ & U \perp\!\!\!\perp W \text{ and } U \sim F_0 \text{ given} \end{aligned} \tag{4.1}$$

If $Z = W$ and U uniform this model is called the conditional quantile model and may be view as a way to describe the conditional distribution of Y given Z. If U is exponential and Y non negative this equation is a general characterization of duration model conditional to cofactors Z (see Horowitz (1996)). This model may be generalized by relaxing some assumptions as the monotonicity condition.

Our objective here is to relax the assumption $Z = W$ and to consider the instrumental variable generalization of the non separable models by considering the case where Z and W are distincts. (See Horowitz and Lee (2007)). A complete theory of these models is out of the scope of this survey but we want to give some elements about this specification.

First let us note that equation (4.1) leads a non linear integral equation where the unknown element is the fonction φ. Indeed:

$$\begin{aligned} U \perp\!\!\!\perp W \quad & \Leftrightarrow \quad \int Prob(U \leq u, Z = z | W = w) dz = F_0(u) \qquad (4.2) \\ & \Leftrightarrow \quad \int F(\varphi(z, u), z | w) dz = F_0(u) \end{aligned}$$

$$\begin{aligned} \text{where } F(y, z|w) \quad & = \quad Prob(Y \leq y, Z = z | W = w) \qquad (4.3) \\ & = \quad \frac{\partial^p}{\partial z_1, ..., \partial z_p} Prob(Y \leq y, Z \leq z | W = w) \end{aligned}$$

The fonction F is identifiable and estimable from the data, F_0 the c.d.f. of U is given and (4.2) appears as an equation which characterizes φ.

The next question is the identification question, i.e. the unicity of the solution of (4.2). As the equation is non linear it is natural to look at the local unicity of the solution which may be characterized by the one to one property of the linear approximation of the equation at the true value.

Let $f(y,z|w)$ the density of (Y,Z) given $W=w$ $(f(y,z|w)=\frac{\partial}{\partial y}F(y,z|w))$. The linearized version of the equation (4.2) denoted $T(\varphi)=F_0$ is based on the linear operator :

$$T'_\varphi(\tilde{\varphi})=\int \tilde{\varphi}(z,u)f(\varphi(z,u),z|w)dz \tag{4.4}$$

This operator is computed as the Gâteau derivative of T in φ and is shown to be the Frechet derivative under regularity conditions (see recalled in Nashed (1971)) The model is then locally identified if T'_φ is one to one for any φ. Assuming that the true φ is almost surely (as a function of Z and U) different from 0 we have:

$$\begin{aligned} T'_\varphi(\tilde{\varphi})=0 \quad &\Leftrightarrow \quad \int \frac{\tilde{\varphi}(z,u)}{\varphi(z,u)}\varphi(z,u)f(\varphi(z,u),z|w)dz=0 \\ &\Leftrightarrow \quad g(u|w)\int \frac{\tilde{\varphi}(z,u)}{\varphi(z,u)}g(z|w,w)dz=0 \end{aligned} \tag{4.5}$$

if $g(z,u|w)$ is the density of (Z,U) given W.

We say that Z is strongly identified by W given U if for any integrable function $\lambda(Z,U)$ we have $E(\lambda(Z,U)|W,U)=0$ implies $\lambda(Z,U)=0$ almost surely. It follows immediately that φ is locally identified if Z is strongly identified by W given U.

Let us now briefly discuss the estimation procedure of φ. The principle would be to construct a regularized solution of

$$min\|T(\varphi)-F_0\|^2 \tag{4.6}$$

where T is replaced by a non parametric estimator. This minimization is difficult and may lead to unconsistent estimator for some estimators of T. A better strategy is to estimate the first order conditions of the Tikhonov functional

$$\|T(\varphi)-F_0\|^2+\alpha\|\varphi\|^2 \tag{4.7}$$

i.e.

$$\alpha\varphi+T'^*_\varphi(T(\varphi)-F_0)=0 \tag{4.8}$$

where $T_\varphi^{\prime *}$ is the adjoint of T'_φ defined in (41). We have:

$$T_\varphi^{\prime *}(\psi) = \int \psi(w) f(\varphi(z,u), w|z) dw$$

where $f(y, w|z)$ is the joint density of (Y, W) given Z.

Numerous iterative methods exists for solving a non linear integral equation and are out of the scope of the paper (see Kaltenbacher et al (2008) for a survey of these methods).

For example we may consider the iterated following method. If $\hat{\varphi}_{k-1}^\alpha$ is the value of the estimator at step $k-1$ the new value $\hat{\varphi}_k^\alpha$ will be the solution of

$$\alpha\varphi + T_{\hat{\varphi}_{k-1}^\alpha}^{\prime\alpha}(T(\varphi) - F_0)) = 0 \tag{4.9}$$

The parameter α may be fixed or updated at each step. The algorithm is stopped at the convergence ($\hat{\varphi}_{k-1}^\alpha \simeq \hat{\varphi}_k^\alpha$) because the regularization is coming from the α parameter.

We don't consider in that section the extension of the analysis of the convergence rate of the estimator of φ neither then the optimal selection of the regularization parameter (see Gagliardini and Scaillet (2006), Chernozukov et al (2009) or Horowitz and Lee (2007)).

5 Some extensions to dynamic models

All the specifications we have considered have been introduced in an *i.i.d.* context. Their extension to some dynamic case with discrete time observations is natural. Take for example the model

$$Y_t = \varphi(Z_t) + U_t \tag{5.1}$$

where (Y_t, Z_t) is a joint markov process and

$$E(U|Y_{t-1}, Z_{t-1}) = 0 \tag{5.2}$$

All the theory of section 3 applies where the instruments W are now the lagged variables (Y_{t-1}, Z_{t-1}). In case of weakly dependent processes the main results of non parametric estimation apply and for example the analysis of the rate of convergence remains identical. Non stationary data main leads to unexpected conclusions: if (Y_t, Z_t) is a unit root process, Wang and Phillips (2009) verify that a usual estimation of the regression of Y_t given Z_t is a consistent estimators of φ.

We want to give a brief survey of a more theoretical approach for instrumental analysis for stochastic processes, possibly with a continuous time. (a complete presentation and examples are given in Florens and Simon (2010)).

We will briefly present two different approaches which correspond for stochastic processes to the extension of separable and non separable models.

The first extension considers a stochastic process Y_t $(t \geq 0)$ possibly with t continuous and two filtrations $\mathcal{Z}_t$ and $\mathcal{W}_t$. In general there exists two stochastic processes Z_t and W_t such that $\mathcal{Z}_t$ is generated by Y_t and Z_t and W_t by Y_t and W_t. In an intuitive presentation the idea is to decompose the variation of Y_t in this way:

$$dY_t = \lambda_t dt + dU_t \tag{5.3}$$

where λ_t depends on $\mathcal{Z}_t$ and where $E(dU_t|\mathcal{W}_t) = 0$. More formally if we integrate w.r.t. to t equation (5.3) we get

$$Y_t = \Lambda_t + U_t \tag{5.4}$$

where Λ_t is $\mathcal{Z}_t$ predictable and U_t satisfies the martingale condition $E(U_t - U_s|\mathcal{W}_s) = 0$.

This model may be identified by computing first the decomposition of Y_t w.r.t. to $\mathcal{W}_t$:

$$Y_t = H_t + M_t \tag{5.5}$$

where H_t is $\mathcal{W}_t$ predictable and M_t is a W_i martingale. We assume that (5.5) has a differential version

$$dY_t = h_t dt + dM_t \tag{5.6}$$

In that case λ_t is solution of

$$h_t = E(\lambda_t|\mathcal{W}_t) \quad \forall t. \tag{5.7}$$

The equation (5.7) generates a sequence of linear integral equations which may be treated (for each value of t) in the same way as in section 3. However h_t and λ_t

may depend on the complete past of Y_t and W_t for h_t and of Y_t and Z_t for λ_t and the statistical treatment of this problem is impossible unless some restrictions are improved to these processes.

This decomposition does not cover all the interesting cases and we propose another class of stochastic processes models with endogenous variables defined in the following way.

Let φ_t an increasing sequence of stopping times adapted to the filtration $\mathcal{Z}_t$ and U_t a process with a given distribution. We assume $(U_t)_t$ and $(W_t)_t$ independent (the complete paths of U and W are independent) and the model is defined by assuming

$$Y_{\varphi_t} = U_t, \tag{5.8}$$

i.e. the process Y stopped at φ_t is equal to U_t.

This model may be view as a non separable model and be used for counting processes (U_t is an homogenous Poisson process) or for diffusion (U_t is a Brownian motion). It is shown in Florens and Simon (2010) that φ_t is characterized as the solution of a non linear integral equation:

$$\int Q(dz) \int_0^{\varphi_t} k_t g(z|\mathcal{W}_s) ds = H_t^U \tag{5.9}$$

where H_t^U is the compensator U_t w.r.t. its own his history and is given, k_t is the intensity of Y_t w.r.t. $\mathcal{Z}_t$ and $\mathcal{W}_t$ and g is the density of the process $(Z_t)_t$ w.r.t. a dominating measure Q.

Several examples of the application of this formulae are given in Florens and Simon (2010). In this paper the local unicity of the solution of this sequence of equation is also discussed.

6 Conclusion

This paper present a different examples of econometric models based on an instrument variable assumption and shows that the functional parameter of interest is characterized as the solution of a linear or non linear integral equation. We have illustrated the main questions following of this characterization: unicity or local unicity of the solution, degree of ill-posedness and regularization, speed of convergence of the solutions and data driven selection of the regularization parameters.

All these points have been illustrated by a Monte Carlo analysis in Fève and Florens (2010) and an application may be founded e.g. in Blundell et al (2007).

References:

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71(6)*, 1795–1843.

Blundell, R., X. Chen and D. Kristensen (2007). Semi-nonparametric IV estimation of shape-invariant Engle Curves. *Econometrica. 75(6)*, 1613–1669.

Blundell, R. and J. Horowitz (2007). A non-parametric test of exogeneity. *Review of Economic Studies.Volume 74-4*, 1035–1058.

Carrasco, M. (2008). A regularization approach to the many instruments problem. Mimeo University of Montreal.

Carrasco, M. and J.P. Florens (2000). Generalization of the GMM in presence of a continuum of moment conditions. *Econometric Theory. 16*, 797–834.

Carrasco, M., J.P. Florens and E. Renault (2007). Linear inverse problems in structural econometrics: estimation based on spectral decomposition and regularization. In J.J. Heckman and E.E. Leamer (Eds.). *Handbook of Econometrics, Volume 6B*, North Holland Amsterdam.

Chen, X. and M. Reiss (2007). On rate optimality for ill posed inverse problems in econometrics. Forthcoming in *Econometric Theory.*

Chernozhukov, V., P. Gagliardini and O. Scaillet (2009). Nonparametric instrumental variable estimators of quantile structural effects. HEC, swiss finance institute, technical report.

Darolles, S., J.P. Florens and E. Renault (2003). Non Parametric instrumental regression. IDEI Working Paper.

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

Engle, R.F., Hendry, D.F. and J.F. Richard (1983). Exogeneity. *Econometrica.* 51 (2), 277-304.

Feve, F. and Florens, J. P (2009). The practice of nonparametric estimation by solving inverse problem: the example of transformation models. Discussion paper.

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

Florens, J.P. and M. Mouchart (1985). Conditioning in dynamic models. *Journal of Time Series Analysis. 53 (1)*, 15–35.

Florens, J.P. and G. Simon (2010). Endogeneity and instrumental variables in dynamic models. Working paper TSE.

Florens, J.P. and A. Simoni (2010). *Regularized Posteriors in Linear Ill-posed Inverse Problems.* Discussion paper.

Florens, J.P. and S. Sokullu (201) Semiparametric transformation models. Working paper TSE.

Florens J.P. and S. Van Belleghem (2009). Functional instrumental linear regression. Mimeo Toulouse.

Florens, J.P. , J. Johannes and S. Van Bellegem (2005). Instrumental regression in partially linear models. Forthcoming in *Econometric Journal.*

Florens, J.P., J. Johannes and S. Van Bellegem (2007). Identification and estimation by penalization in nonparametric instrumental regression. Forthcoming in *Econometric Theory.*

Florens, J.P., J.J. Heckman, C. Meghir and E. Vytlacil (2008). Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica 76 (5)*, 1191–1206.

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

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

Hansen, L.P (2002). Large sample properties of generalized moments estimators. *Econometrica. 50 (10)*, 1029–1054.

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

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

Horowitz, J.L. (2006). Testing a parametric model against a nonparametric alternative with identification through instrumental variables. *Econometrica 74*, 521–538.

Horowitz, J. and S. Lee (2007). Non parametric instrumental variables estimation of a quantile regression model. *Econometrica 75*, 1191–1208.

Johannes, J., S. Van Bellegem and A. Vanhems (2007). A unified approach to solve ill posed inverse problems in econometrics. Forthcoming in *Econometric Theory.* Li, Q.I. and J.S. Racine (2006). *Nonparametric Econometric.* Princeton University Press, Princeton.

Kaltenbacher, B. A. Neubauer and O. Scherzer (2008). *Iterative regularization methods for non linear ill-posed problems.* Random Series on Computational and Applied Mathematics. De Grugter, Berlin.

Koopmans, T.C. and O. Reiersol (1950). The identification of structural characteristics. *Annals of Mathematical Statistics. 21*, 165–181.

Loubes, J.M. and C. Marteau (2010). Paths toward adaptive estimation for instrumental variables regression. Working paper IMT.

Nashed, M.Z. (1971). Generalized inverses, normal solvability and iterations for singular operators equations. in L.B. Rall (ed). *Nonlinear functional analysis and application.* Academic Press. 311–359.

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

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

Stock, J.H., J.H. Wright and M. Yogo . A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics. 20 (4)*, 518–529.

Wang, Q. and P.C.B. Phillips (2009). Structural non parametric cointegration regression. *Econometrica. 77 (6)*, 1901–1948.

REFERENCES