Additive Nonparametric Instrumental Regressions: A Guide to Implementation

Working Paper 17-06

S. Centorrino, F. Fève and J. P. Florens

June, 2017

Stony Brook University

ADDITIVE NONPARAMETRIC INSTRUMENTAL REGRESSIONS: A GUIDE TO IMPLEMENTATION

S. CENTORRINO, F. FÈVE AND J. P. FLORENS

Stony Brook University - Economics Department and Toulouse School of Economics

Abstract. We present a review on the implementation of regularization methods for the estimation of additive nonparametric regression models with instrumental variables. We consider various versions of Tikhonov, Landweber-Fridman and Sieve (Petrov-Galerkin) regularization. We review data-driven techniques for the sequential choice of the smoothing and the regularization parameters. Through Monte-Carlo simulations, we discuss the finite sample properties of each regularization method for different smoothness properties of the regression function. Finally, we present an application to the estimation of the Engel curve for food in a sample of rural households in Pakistan, where a partially linear specification is described that allows one to embed other exogenous covariates.

Keywords: Nonparametric; Endogeneity; Instrumental Variables; Ill-posed inverse problem; Regularization; Petrov-Galerkin; Landweber-Fridman; Tikhonov; Simulations; Engel Curve.

JEL Codes: C01; C14; C18; C26.

1. Introduction

Instrumental variables are a widely used approach in econometrics and statistics to achieve identification and carry on inference in the presence of endogenous explanatory variables. However, in many empirical applications, it is often preferred to introduce a parametric structure of the function of interest. The implementation of some (linear or nonlinear) parametric models, that can be estimated using GMM, enormously simplifies the estimation exercise. This comes at the cost of imposing restrictions on the regression function which may not be justified by the economic theory and can lead to misleading inference and erroneous policy conclusions.

By contrast, a fully nonparametric specification of the main model *leaves the data to speak for themselves* and, therefore, does not impose any a priori structure on the functional form. A fully nonparametric approach can be a very useful exploratory tool for applied researchers in order to choose an appropriate parametric form and to test restrictions coming from the economic theory (e.g. convexity, monotonicity).

E-mail address: `samuele.centorrino@stonybrook.edu (corresponding author)`.

Date: This version: October 19, 2015; First version: July 15, 2013.

This paper has been previously circulated under the title “Implementation, Simulations and Bootstrap in Nonparametric Instrumental Variable Estimation”. Authors are deeply indebted to four anonymous referees for their useful comments and to point out some mistakes in the previous version of the manuscript.

However, while the nonparametric estimation of additive models with endogenous regressors using instrumental variables (also known as nonparametric instrumental regression) has recently received enormous attention in the theoretical literature (see, e.g. Darolles, Fan, Florens and Renault, 2011, Horowitz, 2011, and references therein), it remains quite unpopular among applied researchers.[1] This may be partially due to the theoretical difficulties that empirical researchers might encounter in approaching this topic. The regression function in nonparametric instrumental regressions is, in fact, obtained as the solution of an *ill-posed* inverse problem. Heuristically, this implies that the function to be estimated is obtained from a singular system of equations and, therefore, the mapping that defines it is not continuous. Hence, beside the conventional selection of the smoothing parameter for the nonparametric regression, the estimation of this type of models requires to transform this ill-posed inverse problem into a well-posed one. This transformation is achieved with the use of regularization methods that require the selection of a regularization constant.

Tuning of the latter parameter constitutes an additional layer of complication, and it has to be tackled with the appropriate methods. Data-driven techniques for the choice of regularization parameter in the framework of nonparametric instrumental regressions are presented in: Breunig and Johannes (2015), Centorrino (2015), Chen and Christensen (2015), Fève and Florens (2010), Florens and Racine (2012), Horowitz (2014*a*) and Liu and Tao (2014).[2] These works, however, focus on a specific regularization scheme and there is not, to the best of our knowledge, a paper which gives empirical researchers a broad picture about regularization frameworks that can be used in the context of nonparametric instrumental regressions.

Our intention is to give a unified and simple presentation of the several regularization procedures that can be considered when applied researchers would like to keep the flexibility of nonparametric estimation in the presence of endogenous regressors. Our aim is to narrow the gap between the theoretical literature on the topic, which has been growing extremely fast recently, with the empirical use of this framework.

We consider a simple framework with a scalar endogenous covariate, a scalar instrument and without additional exogenous regressors. We analyze the performances of several versions of Tikhonov (Darolles, Fan, Florens and Renault, 2011), Landweber-Fridman (Florens and Racine, 2012, Johannes, Bellegem and Vanhems, 2013) and Petrov-Galerkin (also known as sieve, see Cardot and Johannes, 2010, Chen and Pouzo, 2012, Horowitz, 2011, Johannes and Schwarz, 2011) regularizations in the case where both the smoothing and the regularization parameters are chosen using data-driven methods. We also discuss viable solutions for including additional exogenous covariates in the model and some of the relevant empirical challenges ahead.

The paper is structured as follows. In section (2), we present the main framework. We review carefully each regularization scheme and we discuss its practical implementation in section (3). In section (4), we describe the structure of the Monte-Carlo experiment. In

[1] The few notables exceptions we are aware of are Blundell, Chen and Kristensen (2007), Hoderlein and Holzmann (2011) and Sokullu (2015)

[2] There exists also a very large literature in mathematics about numerical criteria for the choice of the regularization parameter for integral equations of the first kind (Engl et al., 2000, Vogel, 2002).

section (5), we present an application to the estimation of the Engel curve for food using a cross section database of Pakistani households. Finally, section (6) concludes.

2. The main framework

We focus our general presentation on the simplest framework characterized by a triplet of random variables $(Y, X, W) \in \mathbb{R}^3$, satisfying the following model:

$$Y = \varphi(X) + U \tag{1a}$$

$$\mathbb{E}(U|W) = 0 \tag{1b}$$

This model is a regression type model, where the usual mean independence condition $\mathbb{E}(U|X) = 0$ is replaced by condition (1b). This specification has been extensively studied in econometrics in order to account for the possible *endogeneity* of X, i.e. the lack of mean-independence between the covariate X and the error term U. In particular, recent literature has investigated the nonparametric estimation of the function $\varphi(\cdot)$ in (1a) (see,e.g. Newey and Powell, 2003, Hall and Horowitz, 2005, Carrasco et al., 2007, Darolles et al., 2011, Chen and Pouzo, 2012, among others). As a solution to the endogeneity problem, this literature considers the role of one additional set of variables, W, which are exogenous with respect to U and they, therefore, satisfy condition (1b).

By taking the conditional expectation with respect to W in equation (1a), we obtain the following equality:

$$\mathbb{E}(\varphi(X)|W) = \mathbb{E}(Y|W). \tag{2}$$

Suppose now we observe an iid realization of size N of the triplet (Y, X, W) that satisfies our nonparametric regression model. The conditional expectation of Y given W can be easily estimated using the available data, as the regression of Y on W. Define:

$$r = \mathbb{E}(Y|W).$$

However, the unknown parameter φ enters equation (2) inside the conditional expectation operator. Therefore, φ is defined only implicitly by equation (2) and, in its current form, this equation cannot be used to compute an estimator as a direct function of the data.

To understand the approach undertaken for nonparametric regression, it is useful to start from the simplest parametric specification. Consider that $\varphi(X) = X\beta$ and define:

$$Y = X\beta + U \tag{3}$$

$$X = W\gamma + V, \tag{4}$$

where both X and W can be thought to include a constant term, and we impose $\mathbb{E}(U|W)$ and $\mathbb{E}(V|W)$ equal to 0.[3] In this example, we restrict our attention to linear functions of X and W and we are interested in the estimation of the parameter β. When we take the conditional expectation with respect to W in the first equation, we obtain the following restriction:

$$\mathbb{E}(X|W)\beta = r.$$

[3] The assumption of zero correlation would suffice in the linear parametric model. However, we keep the assumption of zero conditional expectation to maintain the comparison with the nonparametric specification.

Denote as $P_X = X\left(X'X\right)^{-1}X'$ and $P_W = W\left(W'W\right)^{-1}W'$, the orthogonal projectors onto the linear space of the X and the W respectively. Given the restrictions we have put on the data generating process, the conditional expectation is a projection on the space of linear functions. Thus, we have that:

$$P_W X\beta = r,$$

with $r = P_W Y$, removes the error component from the model. By projecting the last equation onto the space of X, we obtain:

$$P_X P_W X\beta = P_X r, \tag{5}$$

which can be easily seen to be satisfied for

$$\beta = \left(X'P_W X\right)^{-1} X'r.$$

This is nothing but the Two-Stage Least Square (TSLS) procedure to obtain an estimator of the parameter β. However, the linear specification imposes several linear restrictions, both on the shape of φ and on the specification of the model for the endogenous variable X. These restrictions may or may not be justified by economic theory. Therefore, a parametric specification might not be appropriate for some empirical applications. More generally, researchers may want to maintain some flexibility about the specification of the model and of the function φ.

As in related papers (Newey, 2013), we can extend the simple intuition about the linear model to our nonparametric estimator. The important difference is that our parameter of interest is now infinite dimensional (it is a function), so that we need to replace P_X and P_W with projectors onto an appropriately defined space of functions. Moreover, while a full rank condition on $X'P_W X$ is sufficient to ensure that the solution of (5) is well defined, in the nonparametric setting the inversion of a large dimensional matrix can create important numerical problems that need to be addressed with appropriate tools.

To restrict our nonparametric problem, we consider φ to belong to the space of square-integrable functions of X, that we denote as $\mathbb{L}^2_X$. The conditional expectation thus projects onto the space of square integrable functions. We define the projections onto this space as follows:

$$\begin{aligned} T: &\quad \mathbb{L}^2_X \to \mathbb{L}^2_W \\ T^*: &\quad \mathbb{L}^2_W \to \mathbb{L}^2_X \end{aligned}$$

with T projecting functions of X onto the space of functions of W and T^* doing the opposite operation, i.e. projecting functions of W onto the space of functions of X.[4]

We now follow the same procedure as above, by applying the projection twice, first onto the space of functions of W, to remove the error component and then onto the space of functions of X. We obtain:

$$T^*T\varphi = T^*r. \tag{6}$$

This last equation and equation (5) appear very similar at first glance. Therefore, one may be tempted to interpret T^*T as a very large dimensional matrix and just write the solution

[4] We define the two operators more formally below.

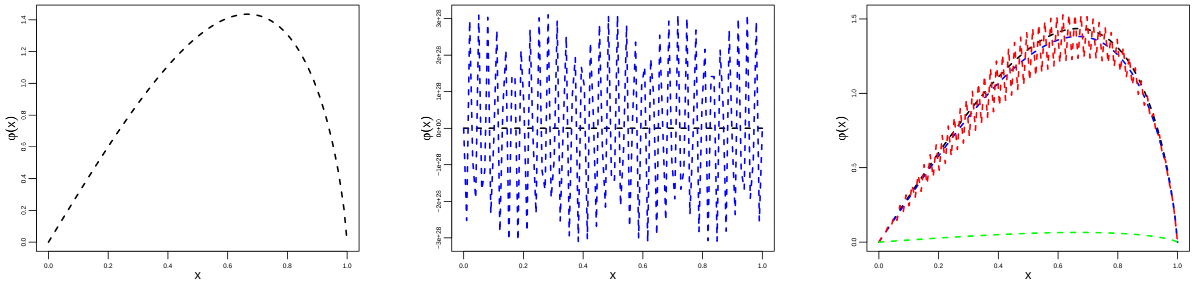
of the system of equations as:

$$\varphi = (T^*T)^{-1} T^*r.$$

While this solution can be assumed to exist and to be well defined in the population, it is generally not well-defined in the sample. This is mainly due to the fact that the smallest eigenvalues of T^*T are getting arbitrarily close to zero so that, in practice, the direct inversion leads to an explosive, non-continuous solution. Moreover, the fact that r is not observed and should be estimated introduces a further error which is magnified as the eigenvalues of T^*T get arbitrarily close to 0. In this sense, φ is not a continuous function of the data and the problem in equation (6) is labelled to be an *ill-posed* inverse problem (see the manuscript of Engl, Hanke and Neubauer, 2000, for a technical introduction).

The classical way to circumvent ill-posedness is to *regularize* T^*T. Regularization, in this context, boils down to the choice of a constant parameter, which transforms the ill-posed into a well-posed inverse problem. The choice of this tuning constant becomes thus a necessary ingredient for estimating nonparametrically the shape of the function φ in regression models with endogeneity.

Figure (1) illustrates this issue. The true known function is plotted in the left panel of the figure. The center panel shows the solution obtained by direct inversion of the integral operator. This solution is clearly explosive because the inverse mapping is not continuous. Finally, the right panel shows the regularized solution for several choices of the regularization parameter. Call our regularization parameter α. A large value of α *oversmooths* the inverse mapping. The function obtained is the flat green line in Figure (1), which is totally uninformative about the shape of the true regression function. A value of α that is too small, corresponds instead to *undersmoothing*. The oscillating red line obtained using a small value of α does not give any specific guidance about the shape of the true function. By contrast, with the right choice of α (blue line), we are able to retrieve a good numerical approximation of the true function.



(A) (B) (C)

Figure 1. The true function (A) and its numerical approximation by direct inversion of the operator (B) and using several values of the regularization parameter (C).

Therefore, the problem in (6) should be tackled using an appropriate regularization procedure. The heuristic idea is to replace the operator T^*T by a continuous transformation of it, so that the solution does not blow up. One could add to every eigenvalue a small constant term. This constant term *controls* the rate of decay of the eigenvalue to 0 (Tikhonov regularization). Another approach would be to replace the infinite dimensional matrix T^*T by a finite approximation of it and estimate the Fourier coefficients by projection on an arbitrary basis function of the instruments and the endogenous variable (sieve regularization). Finally, it is possible to avoid the inversion of the operator T^*T, by using an iterative method (Landweber-Fridman regularization). Note that all these methods require the tuning of a *regularization parameter*: the constant which controls the decay of the eigenvalues; the finite term at which the sum has to be truncated; and the number of iterations to reach a reasonable approximation of the operator's inverse.

One of the aims of this work is to gather and discuss data-driven choices of such parameters.

We conclude this initial presentation with more formal definitions of the objects used above. Uninterested readers can skip the remainder of this section.

2.1. **Theoretical underpinnings.** We assume that the triplet (Y, X, W) is characterized by its joint cumulative distribution function F, dominated by the Lebesgue measure. Denote as f its probability density function. We consider the space of square integrable function relative to the true F and we denote, for instance, by $\mathbb{L}^2_X$, the space of square integrable functions of X only. We further assume that $\mathbb{E}\left[Y^2|W=w\right] < \infty$ and $r \in \mathbb{L}^2_W$. The operator T is a conditional expectation operator and it defines the following linear mapping:

$$T: \quad \mathbb{L}^2_X \to \mathbb{L}^2_W$$
$$(T\varphi)(w) = \int \varphi(x) f(x|w) dx$$

In order to solve (2), we also require the adjoint of the operator T, T^*, which is defined as follows:

$$\langle T\varphi, \psi \rangle = \langle \varphi, T^*\psi \rangle \quad \text{where} \quad \varphi \in \mathbb{L}^2_X \quad \text{and} \quad \psi \in \mathbb{L}^2_W,$$

and

$$\left(T^*\psi\right)(x) = \int \psi(w) f(w|x) dw$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{L}^2_X$ or in $\mathbb{L}^2_W$ and ψ is a function in $\mathbb{L}^2_W$.

The operators T and T^* are taken to be compact (see, e.g. Carrasco et al., 2007, Darolles et al., 2011) and they, therefore, admit a singular value decomposition. That is, there is a nonincreasing sequence of nonnegative numbers $\{\lambda_i, i \geq 0\}$, such that:

(i) $T\phi_i = \lambda_i \psi_i$
(ii) $T^*\psi_i = \lambda_i \phi_i$

For every orthonormal sequence $\psi_i \in \mathbb{L}^2_W$ and $\phi_i \in \mathbb{L}^2_X$. Using the singular value decomposition of T, we can rewrite equation (2) as:

$$\sum_{j=1}^{\infty} \lambda_j \varphi_j \psi_j = \sum_{j=1}^{\infty} r_j \psi_j$$

where $\varphi_j = \langle \varphi, \phi_j \rangle$ and $r_j = \langle r, \psi_j \rangle$ are the Fourier coefficients of φ and r, respectively. We point out that compactness is not a simplifying assumption in this context but describes a realistic framework in which the eigenvalues of the operator are declining to zero. The assumption that the eigenvalues of the operator are bounded below by a strictly positive value is relevant for other econometric models, but it is never satisfied in the case of continuous nonparametric instrumental variable estimation.

Finally, a crucial assumption for identification is that the operator is T is injective, that is:

$$T\varphi \overset{a.s.}{=} 0 \quad \Rightarrow \quad \varphi \overset{a.s.}{=} 0 \tag{8}$$

(see Newey and Powell, 2003, Darolles et al., 2011, Andrews, 2011, D'Haultfoeuille, 2011). This *completeness condition* is assumed to hold throughout the paper. Canay, Santos and Shaikh (2013) have shown that it is not possible to construct a test for this assumption that uniformly controls size. However, a recent paper by Freyberger (2015) has proposed a testing procedure that links the outcome of such a test with the consistency of the nonparametric estimator. Another recent body of work has provided some genericity results (Chen, Chernozhukov, Lee and Newey, 2014), or sufficient conditions on the conditional distribution of X given W, such that completeness holds (Hu and Shiu, 2015). Moreover, Santos (2012) studies a testing procedure under partial identification, that is, when the completeness condition in (8) may fail. However, its implementation goes beyond the scope of the present paper.

Finally, under this set of assumptions, we can use Picard's theorem (see, e.g. Kress, 1999, p. 279) and write the solution to our inverse problem as:

$$\varphi = \sum_{j=1}^{\infty} \frac{r_j}{\lambda_j} \phi_j \tag{9}$$

The ill-posedness in (2) arises because of two main issues:

(i) The inverse operator T^{-1} is a non-continuous operator. The noncontinuity of T^{-1} is tantamount to the fact that the eigenvalues $\lambda_j \to 0$, as $j \to \infty$. This leads to a nonconsistent estimation of the function φ.

(ii) The right-hand side of the equation needs to be estimated. This approximation introduces a further error which renders the ill-posedness of the problem even more severe.

3. Implementation of the Nonparametric IV Estimator

Consider again the model in (1a):

$$Y = \varphi(X) + U, \text{ with } \mathbb{E}\left(U|W\right) = 0.$$

Our parameter of interest is the function φ, which, by taking the conditional expectation with respect to the instrument W, is implicitly defined by the following moment condition:

$$\mathbb{E}\left(\varphi(X)|W = w\right) = r(w),$$

where $r(w) = \mathbb{E}\left(Y|W = w\right)$.

To maintain the comparison with the linear model, we rewrite our moment condition in terms of the conditional expectation operator T, that is:

$$T\varphi = r.$$

By iterating the projection onto the space of the endogenous variable X, we finally obtain:

$$T^*T\varphi = T^*r. \tag{10}$$

We are given an iid sample from the triplet (Y, Z, W), denoted $\{(y_i, z_i, w_i), i = 1, \ldots, N\}$. From this sample, we aim to obtain an estimator of three objects: r, T and T^*. For each one of these objects, we wish to select a smoothing parameter. That is, either the bandwidth parameter, in case of local polynomial regression, or, the number of approximating bases, in the case of series methods.

However, notice that φ is unknown in equation (10) so that we end up selecting two smoothing parameters. One smoothing parameter is needed from the nonparametric regression of Y on W. This is also used to obtain an estimator of the conditional expectation operator T. Finally, we need to select a smoothing parameter for the estimation of T^*, the conditional expectation operator with respect to the endogenous variable X. We denote these estimators $\hat{r}$, $\hat{T}$ and $\hat{T}^*$.

We can, therefore, write the sample counterpart of equation (10) as:

$$\hat{T}^*\hat{T}\hat{\varphi} = \hat{T}^*\hat{r}. \tag{11}$$

This equation defines our estimator of φ as a solution of a large system of equations. We can interpret this system to be almost singular in finite samples. It will be at this point that the regularization procedure plays a role in transforming this problem into a well-posed one. Regularization, as nonparametric smoothing, consists of the selection of a constant parameter which is meant to remove the singularity in the system of equations.

To summarize, we have four objects to estimate in this model:

(i) The triplet (r, T, T^*) can be estimated using any nonparametric regression technique. These objects require the choice of *smoothing parameters.*

(ii) The function φ, which is found from the solution of an ill-posed inverse problem and, therefore, requires the selection of a *regularization parameter.*

Despite the fact that a correspondence between the smoothing and the regularization parameters clearly exists (Chen and Pouzo, 2012), their simultaneous choice is, to the best of our knowledge, not feasible.[5] The most judicious approach seems to select them sequentially. As a matter of fact, it appears that the regularization parameter adjusts to the choice of the smoothing parameter in a reasonable set of values.[6]

It is essential for practitioners to be able to access data-driven techniques for the selection of both types of parameters. As a matter of fact, these provide an objective decision rule for the selection of the tuning constants, given the sample. There is already a extensive body of literature about the selection of smoothing parameters for nonparametric regressions (for a review, see Härdle, 1990, Li and Racine, 2007). Hence, here we mainly focus our attention on the methods for the data-driven selection of the regularization parameter, after we have fixed the smoothing parameters using our preferred data-driven approach (e.g., least-squares cross-validation and AIC, just to name a few).

[5] The very recent paper by Liu and Tao (2014) has tackled the joint choice of the smoothing and the regularization parameter simultaneously in the case of sieve regularization.

[6] For a discussion on this topic, see also Fève and Florens (2010).

Given the smoothing parameter, an inadequate choice of the regularization parameter has a substantial impact on the final estimation as shown before: if we regularize too much, the estimated curve becomes flat as we *kill* the information coming from the data; if we do not regularize enough, the estimator oscillates around the true solution, but it does not ultimately give any guidance about the form of the regression function.

Another important choice is related to the *regularization procedure* the researcher wants to implement. This is essential to properly characterize the dependence of our estimator of the regression function φ on the nonparametric estimates of r, T and T^*.

The remainder of this section is divided into two parts. In the first part, we discuss the estimation of r, T and T^* by local linear kernel regressions and Legendre polynomials. In the second part, we review the regularization procedures we undertake in this paper and present, for each of them, one criterion for the data-driven choice of the regularization parameter.

3.1. Estimation of r, T and T^*.

3.1.1. *Local Linear Regressions.*

Seminal papers about the estimation of nonparametric instrumental regressions (Hall and Horowitz, 2005, Darolles et al., 2011) generally present their results when generalized kernels are used for estimation (Müller, 1991). These are theoretically useful to deal with the some of the issues related to local constant nonparametric regressions (e.g., boundary bias), but they are not generally implemented, to the best of our knowledge, in standard software.

Local polynomials are more popular in practice to deal with some of the shortcomings of local constant estimation. This is why we present the estimation of r, T and T^* using local polynomials. To simplify our exposition and without loss of generality, we consider polynomials of order 1. That is, local linear fitting.

We take the kernel function $K(\cdot)$ to be positive, symmetric and bounded over the support of the data and satisfying $\int K(t)dt = 1$. Further define $K_h(t) = h^{-1}K(t/h)$. The amount of local smoothness of our estimator is tuned by selecting the bandwidth parameter h.

Recall that r is defined as the conditional expectation of the dependent variable Y, given the instrument W. Therefore, for each $w_l, l = 1, \dots, N$, we can fit the following linear model:

$$y_i = a + b(w_i - w_l) + v_i, \text{ with } i = 1, \dots, N,$$

so that $\hat{r}(w_l) = \hat{a}$ and $\hat{b}$ can be used to estimate the first derivative of r. The local linear fit for r can be therefore written as

$$\min_{a,b} \sum_{i=1}^{N} \left(y_i - a - b(w_i - w_l)\right)^2 K_{h_w}\left(w_i - w_l\right),$$

with bandwidth $h = h_w$.

Define y to be the $N \times 1$ vector of observations of the dependent variable; $\bar{K}_W$, the $N \times N$ diagonal matrix of kernel weights $\{K_{h_w}\left(w_1 - w_l\right), \dots, K_{h_w}\left(w_N - w_l\right)\}$ and $\mathbf{W}$ to be the $N \times 2$ matrix with i-th row equal to $(1, w_i - w_l)$. We have that:

$$\hat{r}(w_l) = e_1' \left(\mathbf{W}'\bar{K}_W\mathbf{W}\right)^{-1}\left(\mathbf{W}'\bar{K}_W y\right) = M_{w_l} y,$$

with $e_1' = (1, 0)$ and M_{w_l} a $1 \times N$ vector.

We now turn to the estimation of T and T^*. Recall that both operators are defined to be compact. Hence, they can both be approximated by a finite dimensional linear smoother (Carrasco et al., 2007).

To approximate T using local linear regressions, we stack in a matrix of dimension $N \times N$, the linear smoothers M_{w_l}, for all $l = 1, \dots, N$, so that:

$$\hat{T} = \left[M'_{w_1}, \dots, M'_{w_N}\right]'.$$

In the same way, we can obtain an estimator of the conditional expectation of r given X, by fitting the following regression:

$$\min_{a,b} \sum_{i=1}^{N} \left(\hat{r}(w_i) - a - b(x_i - x_l)\right)^2 K_{h_x}\left(x_i - x_l\right),$$

with bandwidth $h = h_x$. Define $\bar{K}_X$ to be the $N \times N$ diagonal matrix of kernel weights $\{K_{h_x}\left(x_1 - x_l\right), \dots, K_{h_x}\left(x_N - x_l\right)\}$ and $\mathbf{X}$ to be the $N \times 2$ matrix with i-th row equal to $(1, x_i - x_l)$. Thus:

$$M_{x_j} = e_1'\left(\mathbf{X}'\bar{K}_X\mathbf{X}\right)^{-1}\mathbf{X}'\bar{K}_X,$$

and an estimator of T^* is obtained by stacking the linear smoothers from this last regression as before:

$$\hat{T}^* = \left[M'_{x_1}, \dots, M'_{x_N}\right]'.$$

The bandwidths h_w and h_x can be chosen using leave-one-out cross-validation (Li and Racine, 2004, 2007).

Notice that the use of least squares cross-validation in this context is only of practical relevance and it can be replaced by other methods. Possible alternatives include rule of thumb smoothing, maximum likelihood cross-validation, or a modified AIC criterion (Hurvich et al., 1998). Notice, that all these methods are known to balance the trade-off between variance and bias for nonparametric regressions. In practice, this also seems appropriate in the case of nonparametric instrumental regressions (see Fève and Florens, 2014, Centorrino, 2015, for a further discussion on the topic).

3.1.2. *Series (linear sieves).* Another simple way of estimating r, T and T^* is to use series regressions (Newey, 1997, Chen, 2007) rather than kernel smoothers. The underlying assumption is that functions in $\mathbb{L}^2_Z$ and $\mathbb{L}^2_W$ can be well approximated by a finite sum of basis functions which span those spaces. These basis functions can be chosen according to the joint distribution of X and W (Hoderlein and Holzmann, 2011). However, more frequently their choice remains arbitrary.

In this paper, we use Legendre polynomials, although other basis functions could equivalently be used (e.g., B-splines, wavelets, or Hermite polynomials, just to name a few).

Denote as $p_{J_N}(\cdot)$, the $J_N \times 1$ vector of Legendre polynomials of order J_N, with $0 < J_N < \infty$.

We can estimate r by a simple projection of the dependent variable Y on the polynomial basis of the instrument W. That is, we fit the following linear regression model

$$y_i = p_{J_N}(w_i)'\beta + v_i, \text{ for } i = 1, \dots, N,$$

where β is a $J_N \times 1$ vector of regression coefficients. This model can be estimated using ordinary least squares. Therefore:

$$\hat{r}(w_i) = p_{J_N}(w_i)'\hat{\beta},$$

with $\hat{\beta} = (\mathcal{W}_N'\mathcal{W}_N)^{-}\mathcal{W}_N'y$, where $\mathcal{W}_N = [p_{J_N}(w_1), \ldots, p_{J_N}(w_N)]'$ is a $N \times J_N$ matrix of regressors and $(\mathcal{W}_N'\mathcal{W}_N)^{-}$ denotes the generalized inverse of $\mathcal{W}_N'\mathcal{W}_N$.

As done before, since the operators T and T^* are conditional expectations, they can be estimated using linear smoothers. Thus, we approximate T as follows:

$$\hat{T} = \mathcal{W}_N(\mathcal{W}_N'\mathcal{W}_N)^{-}\mathcal{W}_N'.$$

In order to approximate T^*, we fit the regression of $\hat{r}$ on the polynomial basis of the endogenous variable X, so that:

$$\hat{T}^* = \mathcal{X}_N(\mathcal{X}_N'\mathcal{X}_N)^{-}\mathcal{X}_N',$$

where $\mathcal{X}_N = [p_{J_N}(x_1), \ldots, p_{J_N}(x_N)]'$.

In order to implement the series estimator, the order of the polynomial J_N is our smoothing parameter and ought to be chosen appropriately.

It is also worth mentioning that, in case of the estimation of the operator using linear sieves, once the basis function $p_{J_N}(\cdot)$ has been fixed, we are basically solving a two-stage least squares problem, where the dimension of the parameter to be estimated is J_N (see Horowitz, 2011). Therefore, if the smallest eigenvalue of

$$\mathcal{X}_N'\hat{T}\mathcal{X}_N$$

is bounded away from 0, we do not need to apply any further regularization in order to obtain our estimator of φ. In this sense, the dimension of the basis function J_N acts both as a smoothing parameter, in that it allows us to obtain an estimator of r; and as a regularization parameter, in that it permits the direct inversion of the low dimensional matrix $\mathcal{X}_N'\hat{T}\mathcal{X}_N$. Chen and Christensen (2015) and Liu and Tao (2014) have, therefore, explored ways to select two distinct values for J_N, one that is used for smoothing the nonparametric estimator of r and another to regularize the estimation of φ.

3.2. **Estimation of φ: Regularization procedures.**

In this paper we present three regularization procedures: Tikhonov, Landweber-Fridman and Petrov-Galerkin (more commonly known as sieve regularization).

3.2.1. *Tikhonov Regularization.*

The Tikhonov regularization method (TK henceforth, see Hall and Horowitz, 2005, Darolles et al., 2011) is based on the idea that adding a small but positive constant term to the eigenvalues of $\hat{T}^*\hat{T}$ would allow us to invert this matrix and obtain a well-defined solution from equation (11).

This leads us to consider our estimator of φ as the solution of the following system of normal equations:

$$\alpha\varphi + \hat{T}^*\hat{T}\varphi = \hat{T}^*\hat{r}. \tag{12}$$

It can be immediately seen that this condition implies:

$$\hat{\varphi}^{\alpha} = \left(\alpha I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*r, \tag{13}$$

where the superscript α stresses the dependence of the solution on the regularization parameter.

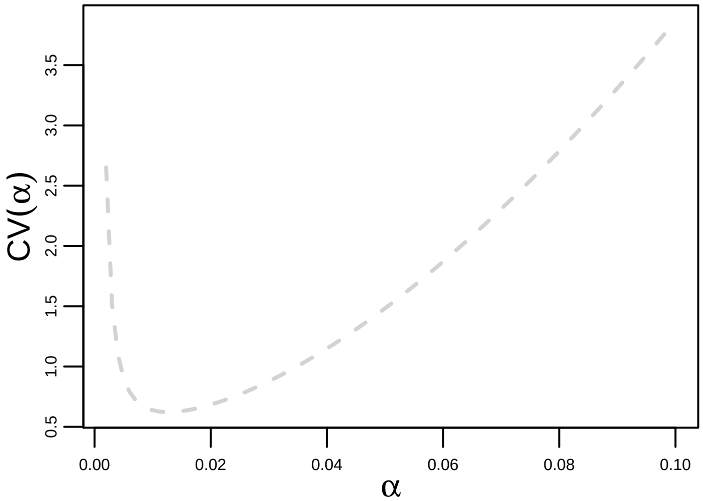


FIGURE 2. Criterion function for the optimal choice of α in Tikhonov regularization

In order to choose the regularization parameter α, we adopt the cross-validation approach developed in Centorrino (2015). This method consists of minimizing the following sum of squares:

$$CV_N(\alpha) = \sum_{i=1}^{N} \left[\left(\hat{T}\hat{\varphi}^{\alpha}_{(-i)} \right)(w_i) - \hat{r}(w_i) \right]^2 ,$$

where $\left(\hat{T}\hat{\varphi}^{\alpha}_{(-i)} \right)(w_i)$ is the estimator of $T\varphi$ obtained by removing the observation (w_i, x_i) from the sample.

Centorrino (2015) has proven that this criterion is order optimal in mean squared error and has shown its superior finite sample performances compared to other existing numerical methods for Tikhonov regularization. A typical shape of this criterion function can be found in figure (2).

The exact theoretical result for the case when the joint dimension of the instrument and the endogenous variable is equal to 2 is given in the following theorem:

Theorem 3.1. *The CV criterion is bounded in probability by:*

$$\left(\frac{\alpha_N + 1}{\alpha_N} \right)^2 \left[\frac{1}{\alpha_N} \left(\frac{1}{N} + h^{2\rho} \right) + \alpha_N^{\min(2u+1,2)} + \left(\frac{1}{Nh^2} + h^{2\rho} \right) \right],$$

where u is the degree of smoothness of the function φ and ρ is the order of the linear smoother.

When the bandwidth parameter, h_N, is chosen in such a way that

$$h_N = O_P\left(N^{-\kappa}\right), \text{ with } 0 < \kappa \leq 1,$$

then the minimization of the CV function leads to a choice of the regularization parameter α_N*, such that:*

$$\alpha_N^{CV} \approx N^{-\frac{\kappa}{(\min(2u,1)+2)}}.$$

3.2.2. *Landweber-Fridman Regularization.* The Landweber-Fridman (LF henceforth) regularization consists of avoiding the inversion of the matrix $\hat{T}^*\hat{T}$ by using an iterative approximation procedure (Johannes et al., 2013).

Consider a constant c, such that $c\|T^*T\| < 1$. If we multiply both sides of equation (11) by c, we obtain:

$$c\hat{T}^*\hat{T}\varphi = c\hat{T}^*\hat{r}. \tag{14}$$

The constant c is not a tuning parameter and its choice is not essential for estimation. Its only role is to ensure the convergence of the iterative method. Since the largest eigenvalue of T^*T is equal to 1, any choice of c that is strictly smaller than 1 would guarantee convergence. Of course, within the values of c that satisfy this restriction, the higher the value of c, the faster the iterative scheme would converge.

The Landweber-Fridman approach iterates equation (14) over φ, in order to find a fixed point of the system of equations.

By adding and subtracting φ on the left-hand side of (14), we obtain the recursive solution:

$$\hat{\varphi}_{l+1} = \hat{\varphi}_l + c\hat{T}^*\left(\hat{r} - \hat{T}\hat{\varphi}_l\right), \quad \forall l = 0, 1, \ldots \tag{15}$$

or equivalently:

$$\hat{\varphi}^{1/\alpha} = c\sum_{l=0}^{1/\alpha-1}\left(I - c\hat{T}^*\hat{T}\right)^l \hat{T}^*\hat{r}, \tag{16}$$

where $1/\alpha$ is the total number of iterations needed to reach the solution and plays the role of regularization parameter. As $1/\alpha$ diverges to infinity, our approximation becomes increasingly precise.

In order to implement the LF regularization, we, therefore, need to choose the number of iterations. Notice that, as shown in Centorrino (2015), a leave-one-out cross-validation criterion could be also used to select α for LF regularization, directly from equation (16), for given estimators of r, T and T^*.

However, Florens and Racine (2012) have shown that smoothing parameters for T and T^* could be updated at every iteration step in equation (15) and that this procedure seems to improve over the mean squared error of the estimator in finite samples.[7]

Therefore, we proceed as follows:

(i) From our estimators of the r, T and T^*, discussed above, we construct the initial condition $\hat{\varphi}_0 = c\hat{T}^*\hat{r}$.

(ii) Using $\hat{\varphi}_0$, we can update smoothing parameters for the estimation of T and T^*.

(iii) From equation (15), we compute $\hat{\varphi}_1$ as:

$$\hat{\varphi}_1 = \hat{\varphi}_0 + c\hat{T}^*\left(\hat{r} - \hat{T}\hat{\varphi}_0\right).$$

[7] We would like to thank Jeffrey S. Racine for insightful discussions on this topic.

(iv) For $l = 2, 3, \dots$, we repeat steps (ii) and (iii), until the following criterion used in Florens and Racine (2012) is minimized:

$$SSR(l) = l \sum_{i=1}^{N} \left[\left(\hat{T}\hat{\varphi}_l \right)(w_i) - \hat{r}(w_i) \right]^2 \quad , \quad l = 1, 2, \dots$$

i.e., we stop iterating when this objective function starts to increase. This criterion function minimizes the sum of square residuals and it is multiplied by the number of iterations in order to admit a minimum. A typical shape of this function is reported in figure (3). It can be seen that the function is only locally convex, so that, we need to check the criterion only after a certain number of iterations has been performed. In practice, one needs to run a sufficient number of initial iterations. The shape of the function can then be checked *ex-post* for local minima.

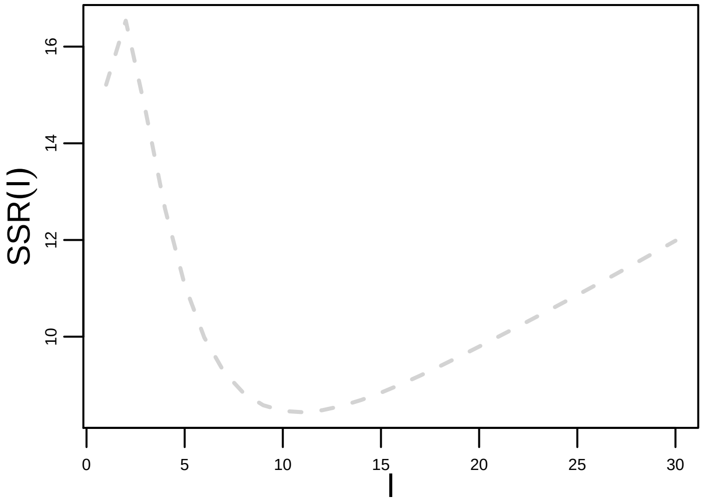


FIGURE 3. Stopping function for Landweber-Fridman regularization

3.2.3. *Sieve (Petrov-Galerkin) Regularization.* The Petrov-Galerkin regularization (GK henceforth, most commonly known as sieve regularization) consists on truncating the infinite sum in (9), by a finite approximation on an a basis (see, e.g. Blundell et al., 2007, Cardot and Johannes, 2010, Horowitz, 2011, Chen and Pouzo, 2012, Gagliardini and Scaillet, 2012).

As the name of the regularization procedure suggests, the approximation of the function φ is obtained using sieve estimation.

Consider the vector of Legendre polynomials of order J_N, as defined above. Our function φ can, therefore, be approximated as:

$$\varphi^{J_N}(x) = \sum_{j=1}^{J_N} p_j(x)\varphi_j = p_{J_N}(x)'\Phi,$$

where φ_j are the Fourier coefficients of φ and $\Phi = [\varphi_1, \dots, \varphi_{J_N}]$. This approximation drastically reduces the dimensionality of our problem: for an infinite dimensional parameter φ, we only require the estimation of a finite number, J_N, of Fourier coefficients. Therefore, as outlined above, the truncation constant J_N does not only act as a smoothing parameter for series estimation, but also as a regularization parameter.

Once we have obtained the estimators of r, T and T^* as detailed in section (3.1.2), we can write:

$$\hat{T}^*\hat{T}\mathcal{X}_N\Phi = \hat{T}^*\hat{r}.$$

As Φ is finite dimensional, the solution of this system of equations can be directly written as:

$$\hat{\Phi} = \left(\mathcal{X}_N'\hat{T}\mathcal{X}_N\right)^{-1}\mathcal{X}_N'\hat{r},$$

and the estimator of φ is given by:

$$\hat{\varphi}^{J_N} = \mathcal{X}_N\hat{\Phi}.$$

For the choice of the regularization parameter J_N, we follow the data-driven method proposed by Horowitz (2014*a*).[8]

Define $\mathcal{H}_{J_N,s}$, the space of functions with s square integrable derivatives, spanned by the family of basis function $p.(\cdot)$, whose Fourier decomposition is truncated at J_N. Define further:

$$\rho_{J_N} = \sup_{\nu \in \mathcal{H}_{J_N,s}, \|\nu\|=1} \left[\| \left(T^*T\right)^{\frac{1}{2}} \nu\|\right]^{-1}.$$

Blundell et al. (2007) call ρ_{J_N} the sieve measure of ill-posedness. As $N \to \infty$, to obtain consistency of the estimator, it is necessary that $\rho_{J_N}\left(J_N^3/N\right)^{\frac{1}{2}} \to 0$ and $\rho_{J_N}\left(J_N^4/N\right)^{\frac{1}{2}} \to \infty$.

The first step estimation consists of finding a value of J_N which satisfies these requirements. Such a value can be defined as:

$$J_N = \arg\min_{J=1,2,\dots}\left\{\rho_J^2 J^{3.5}/N \quad : \quad \rho_J^2 J^{3.5}/N - 1 \geq 0\right\}$$

i.e., J_N is the smallest integer such that $\rho_J^2 J^{3.5}/N \geq 1$. The method for determining this value of J_N has two steps:

(i) Obtain an estimator of ρ_J^2. Such an estimator can be obtained by noticing that $\hat{\rho}_J^{-2}$ is the smallest eigenvalue of the matrix $\hat{T}^*\hat{T}$, where the conditional expectation operators are estimated by a polynomial of order J.

(ii) Finally, define:

$$J_N = \arg\min_{J=1,2,\dots}\left\{\hat{\rho}_J^2 J^{3.5}/N \quad : \quad \hat{\rho}_J^2 J^3.5/N - 1 \geq 0\right\}$$

Hence, one can define:

$$\tilde{\varphi}^{J_N} = \mathcal{X}_N\hat{\Phi},$$

as a first step estimator of φ.

[8] Other recent papers have also proposed data-adaptive choices of the regularization parameter in such a context. See Chen and Christensen (2015) and Liu and Tao (2014).

Horowitz (2014*a*) then proposes to find the optimal data-driven value of the truncation parameter as the minimum of the following function:

$$S_N(J) = \frac{2}{3N^2}\log(N)\sum_{i=1}^{N}\left\{\left[y_i - \tilde{\varphi}^{J_N}(x_i)\right]^2 \sum_{j=1}^{J}\left[\left(\hat{T}\hat{T}^*\right)^{-1}\hat{T}p_j(x)\right](w_i)^2\right\} - \|\hat{\varphi}^J\|^2,$$

with $1 \leq J \leq J_N$. A typical shape of this criterion is drawn in figure (4).

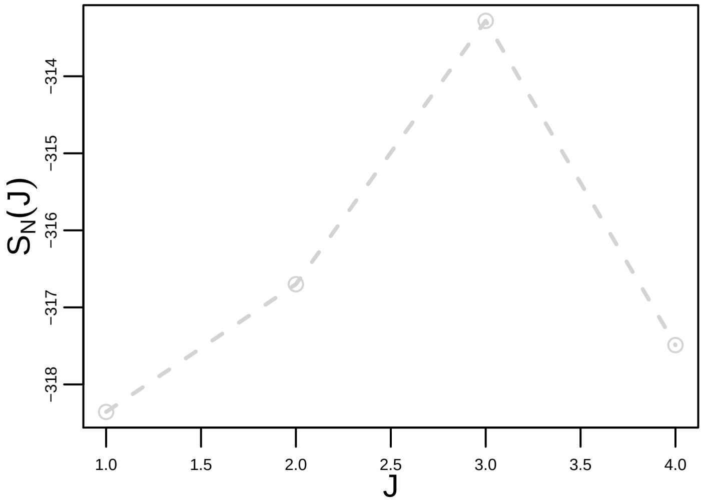


Figure 4. Choice of $\hat{J}_N$ for Galerkin regularization.

A final remark on GK regularization is about the variance of the estimator in finite samples. The GK estimation procedure is a nonparametric generalization of the TSLS estimator. Mariano (1972), in an influential paper, shows that the TSLS estimator in the normal case only possesses moments of order $p-q+1$, where p is the dimension of the endogenous variable and q the dimension of the instruments. Therefore, if one uses the same dimension for the matrices $\mathcal{W}_n$ and $\mathcal{Z}_n$, our GK would possibly have only finite mean, but infinite variance. In order to obtain a finite variance in our sample, we hence include an additional term in the matrix $\mathcal{W}_n$, so that its dimension is $J_N + 1$.

4. Monte-Carlo Simulations

In this section, we analyze the performances of the various estimators previously discussed using data-driven methods. In order to simplify our comparison, we analyze only three estimators: TK with local linear kernels; LF with local linear kernels; and GK with Legendre polynomials.[9]

In this simulation exercise, we fix the joint distribution of the endogenous variable and the instrument, and we focus our sensitivity analysis on the smoothness of the function φ. This

[9] In practice, we could combine Legendre polynomials with TK or LF regularization, but we do not explore it in this work. Interested readers are referred to Chen and Pouzo (2012) and Centorrino (2015).

is potentially more important for applied research, as economic theory may provide some guidance about the properties of the regression function while little may be known about the smoothness of the joint distribution of the data.[10]

The numerical example used in this paper is based on the framework adopted by Darolles et al. (2011), Florens and Simoni (2012) and Florens and Racine (2012). The main data generating process follows equation ($1a$):

$$Y = \varphi(X) + U,$$

where $\mathbb{E}(U|X) \neq 0$ so that endogeneity is present. We simulate independently the instrument W and two disturbances U and V. We then define the endogenous variable X as a function of W, U and V. In particular, we have the following:

$$\begin{aligned} W \sim & \mathcal{N}(0, 25) \\ V \sim & \mathcal{N}\left(0, (0.5)^2\right) \\ U \sim & \mathcal{N}\left(0, (0.25)^2\right) \\ X = & \frac{1}{1 + \exp\left(-(0.4W + 5U + V)\right)}. \end{aligned}$$

The main difference with the numerical examples reported in other papers is that the endogenous variable, X, is a nonseparable function of the instrument, W, and the disturbances, U and V. The companion code for this paper has been programmed in Matlab.

We work with a modest sample size of 1000 observations and we draw 1000 replications of the error terms, V and U, and the instrument, W. Since the regressor X is changing for each one of these replications, we evaluate the estimation of φ on a grid of 100 equispaced points in $(0, 1)$.

A typical sample of W and X is reported in figure (5).

We set the function φ as follows.

$$\begin{aligned} \varphi_1(x) &= 1 - 1.5x \\ \varphi_2(x) &= \sqrt{2}x^2 \\ \varphi_3(x) &= 0.5\sin(1.5\pi x), \end{aligned}$$

where the three functions have an increasing degree of smoothness. The first is a linear specification, and it is helpful to study the performance of the nonparametric estimator compared to the TSLS estimator. The second is quadratic and, while it may fail to be captured by a simple linear specification, it can be approximated by a second order polynomial. Finally, the last function is highly nonlinear and infinitely smooth. In this last case, the linear model is not appropriate for the estimation and the use of nonparametric methods becomes necessary.

Figure (6) reports the result of our estimation for one randomly chosen simulation. Each panel in the figure considers one specification for the regression function φ. The continuous

[10] Of course a nonparametric density estimator of the joint density of X and W can be obtained from the sample. However, besides the jointly normal case (Hoderlein and Holzmann, 2011), we are not aware of any other scenarios in which information about the smoothness properties of the joint density can be directly used for estimation.

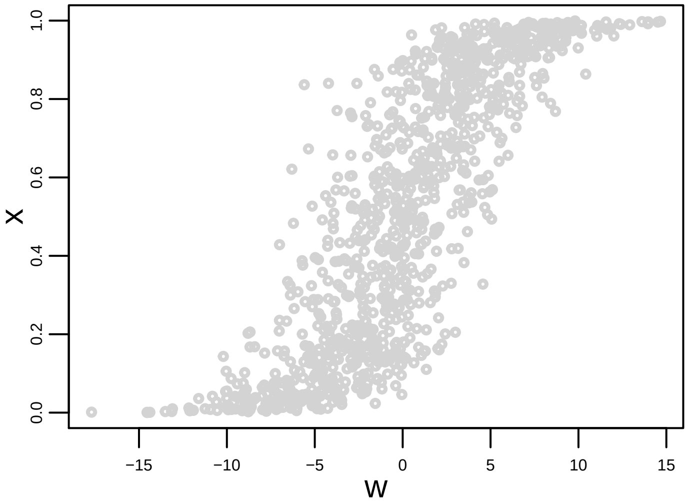


FIGURE 5. A typical draw from the joint distribution of X and W.

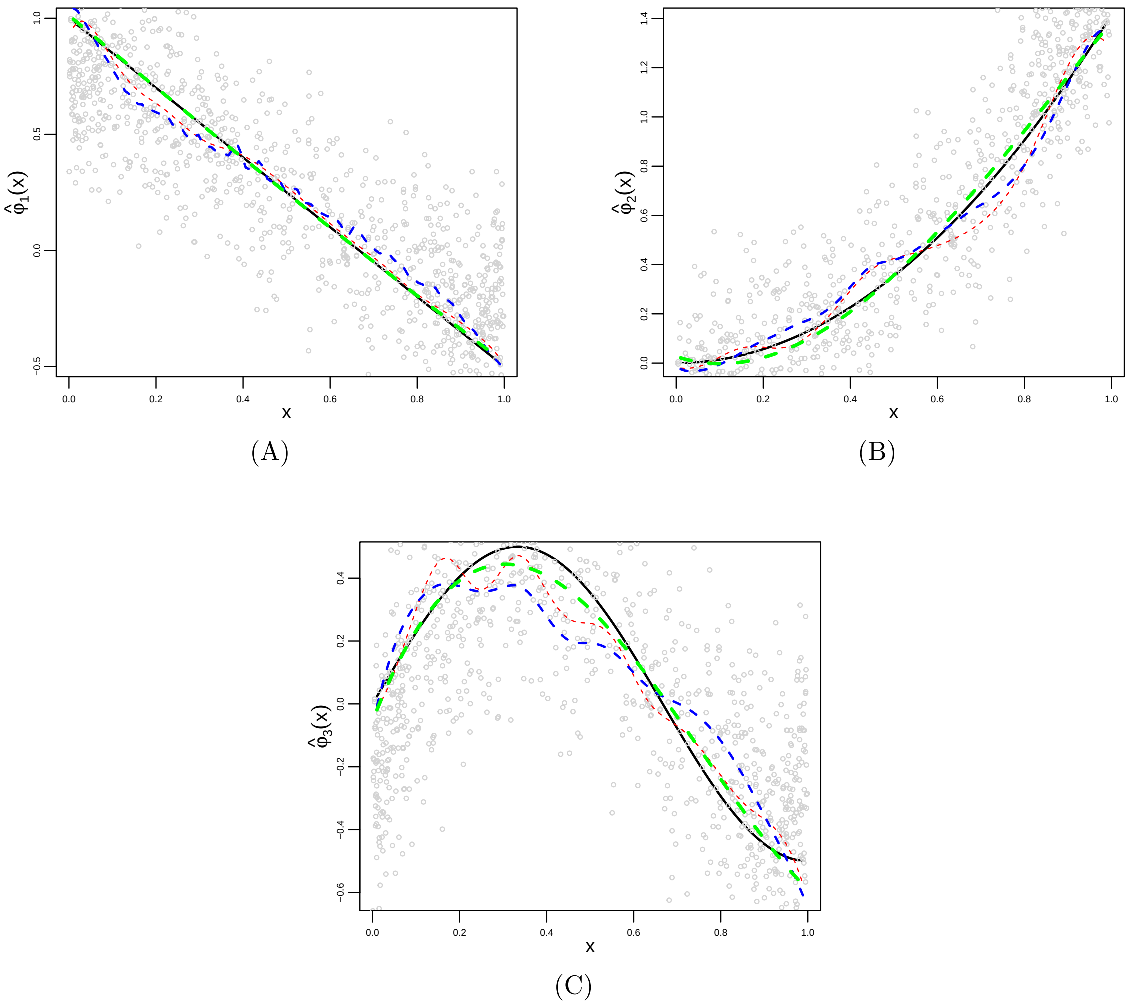
black line in each figure represents the true function; the dashed red line is the TK estimator with local linear kernels; the dashed blue line is the LF with local linear kernels; finally, the dashed green line is the GK with Legendre polynomials.

From this figure, it can be clearly noticed that the GK estimator has a comparative advantage with respect to the TK and LF estimators when the function has a lower degree of smoothness. A polynomial of relative low dimension can, in fact, easily capture the specification for both the linear and the quadratic model. This advantage is also somehow maintained when the order of smoothness of the function increases, although the fitting of the GK estimator is not as good as for the linear and the quadratic case. Both the TK and the LF estimator are more oscillating than the GK estimator. This should be expected as the kernel estimation is local in nature and a constant bandwidth over the entire support can cause some *bumps* in the estimated curve. The Tikhonov estimator with CV regularization parameter seems to have a higher variation than the LF estimator which is compensated by a smaller bias.

This comparison is further carried in terms of squared bias, variance, Mean Squared Error (MSE) and Mean Integrated Squared error over the entire set of simulations.

Table (1) reports the median squared bias, variance and MSE, along with the MISE for each estimator. For comparison, we have also reported these values for the TSLS estimator. While the latter constitutes an important benchmark for the linear case, it is clear that its performance is worsening as the complexity of the regression function increases. However, we believe that the comparison is still relevant as it gives us a hard measure about how much we improve over the linear estimator when φ is nonlinear.

The reported results confirm our initial intuition. The GK estimator dominates the TK and LF estimators in all the simulation studies. In fact, its squared bias and variance are smaller and so is the median MSE and MISE. The TK and LF have a clear squared



(A)

(B)

(C)

Figure 6. The true function (black line) against TK (dashed red line), LF (dashed blue line) and GK (dashed green line) estimators for the three specifications of the function φ over one simulated sample. The gray points represent the simulated data.

bias-variance trade-off. While the TK estimator has a smaller bias, which is in some cases comparable with the one of the GK estimator, its variance is the largest among the estimators studied in this work. This is most likely a feature of CV as a selection criterion for the choice of the smoothing and regularization parameters. Moreover, the variance of the LF estimator might be reduced by the fact that the smoothing parameters are updated at each iteration, which is not the case for the TK estimator. On the contrary, LF has the largest bias. Since the selection of the smoothing parameter for TK and LF is the same, we can conjecture that large squared bias may be due to an insufficient number of iterations. That is, to an insufficient regularization of the estimator.

		MISE	Median MSE	Median Bias^2	Median Variance
$\varphi_1(x)$	TK	0.00225	0.00235	0.00040	0.00180
	LF	0.00247	0.00245	0.00134	0.00107
	GK	0.00057	0.00058	0.00022	0.00035
	TSLS	0.00012	0.00011	0.00000	0.00011
$\varphi_2(x)$	TK	0.00301	0.00308	0.00037	0.00258
	LF	0.00376	0.00354	0.00209	0.00137
	GK	0.00068	0.00072	0.00019	0.00045
	TSLS	0.01229	0.01174	0.01155	0.00013
$\varphi_3(x)$	TK	0.00639	0.00521	0.00236	0.00267
	LF	0.01123	0.00839	0.00687	0.00211
	GK	0.00175	0.00146	0.00063	0.00074
	TSLS	0.05069	0.03546	0.03518	0.00020

Table 1. MISE and Median MSE, Squared Bias and Variance for each estimator and functional form of φ.

This is further explored in Table (2), where we report the summary statistics for the regularization parameter chosen using each one of the selection criteria presented above. In order to make the TK regularization comparable with LF and GK, we have reported the summary statistics for the parameter $1/\alpha$ instead.

Again, the summary statistics for the regularization parameters are broadly consistent with our previous discussion. The CV tends to regularize less than the SSR criterion for LF regularization. This perfectly explains the trade-off between squared bias and variance formerly discussed. For the GK estimator, we can see that the data-driven selection of the truncation constant delivers very little variation in the value of the regularization parameter. In particular, when the function is very smooth, the criterion always selects $J_N = 3$. This could also be a reason of the advantage of GK over TK and LF. While in the latter we require the choice of two smoothing and one regularization parameters, GK only needs us to pick a constant which serves both for smoothing and regularizing.

		Mean	Median	St.Dev	Min	Max
$\varphi_1(x)$	TK	65.9	59.6	28.6	5.6	189.1
	LF	10.9	10.0	3.2	5.0	44.0
	GK	2.1	2.0	1.0	1.0	3.0
$\varphi_2(x)$	TK	93.9	88.3	44.1	6.6	297.9
	LF	11.8	11.0	3.2	5.0	27.0
	GK	2.5	2.0	0.5	2.0	3.0
$\varphi_3(x)$	TK	48.6	44.8	20.9	3.8	160.1
	LF	14.8	13.0	6.6	4.0	46.0
	GK	3.0	3.0	0.0	3.0	3.0

Table 2. Summary statistics for the regularization parameter.

Finally, we also report in Table (3) some summary statistics for the computational time (in seconds). It is evident that the GK type regularization holds an advantage upon all other estimators. This is due to the fact that GK does not require a cross-validated tuning of the smoothing parameter, which can be computationally very costly. Moreover, the dimension of the estimated operator is reduced from the number of observations to the number of bases after truncation, which impacts computational time tremendously. Hence, although we only focus here on a fixed sample size, we expect that the gap in computational time between the GK regularization type and the other estimators spreads further as N increases. A final comment is about the difference between TK and LF regularization. TK regularization still holds an advantage in terms of computational time. This is because the choice of the smoothing parameter is performed only once in TK, while for LF it has to be repeated as many times as the number of iterations. Furthermore, the sample size considered in our simulation study is relatively mild and the inversion of the regularized operator does not require excessive CPU memory. However, as the sample size increases, the computation of the inverse operator becomes very costly and this computational advantage may disappear.

	Mean	Median	St.Dev	Min	Max
TK	96.93	91.57	33.87	42.73	391.32
LF	1149.90	982.79	482.73	302.74	5289.23
GK	0.84	0.81	0.23	0.22	2.88

Table 3. CPU time for each estimator (in seconds).

5. Including additional exogenous variables: an empirical application to the estimation of the Engel curve for food in rural Pakistan

While the Monte-Carlo study is useful to compare the performances of different estimators in the simple framework of one endogenous variable and one instrument, applied researchers are often aiming at estimating models with several other exogenous controls, Z. In general, we could flexibly extend our model in such a way that:

$$Y = \varphi(X, Z) + U, \tag{18}$$

with $\mathbb{E}\left[U|Z, W\right] = 0$. In this case, we could simply extend the definition of our conditional expectation operators so that:

$$\begin{aligned} T :&\mathbb{L}^2_{X,Z} \to \mathbb{L}^2_W \\ T^* :&\mathbb{L}^2_W \to \mathbb{L}^2_{X,Z}, \end{aligned}$$

and φ would still be defined as the solution of the following equation:

$$T\varphi = r.$$

Notice that estimation is carried exactly in the same way as detailed before, with the minor exception that now T^* defines the conditional expectation over functions of both X and Z.

Three points are worth mentioning. In the nonparametric setting, Z and W do not have elements in common. Compactness of the operator would be lost in this case (Carrasco et al., 2007). Furthermore, empirical researchers may be worried that, in order to identify

φ, we would need as many instruments as the joint dimension of X and Z. This is however not the case. Despite the fact that now φ is a joint function of X and Z, it is still only one object that we are trying to identify, so only one (sufficiently strong) instrument would be enough for identification and estimation.[11] Finally, the endogenous variable X should not necessarily be scalar. We can, therefore, include in this more general framework cases in which an empirical researcher faces more than one continuous endogenous regressor.

However, a model like the one in equation (18) suffers from the curse of dimensionality common to nonparametric regressions, especially when the dimension of the exogenous variable Z gets large. Moreover, its interpretation may not be straightforward. Therefore, in this section we look into simpler strategies for the estimation of models with several exogenous controls.

In nonparametric statistics and econometrics, in order to avoid the curse of dimensionality, it is common to construct additive specifications of the more general model in (18) (Horowitz and Mammen, 2004, Horowitz, 2014*b*). That is, one might consider the following specification for the function φ:

$$\varphi(X, Z) = \varphi_x(X) + \sum_{j=1}^{k} m_j(Z_j),$$

where k is the dimension of Z. However, to the best of our knowledge, methods to estimate such an additive structure in regression models with endogenous variables using instruments are not available yet.[12] However, recent papers have provided estimation procedures for the following semiparametric structure (Ai and Chen, 2003, Florens et al., 2012):

$$\varphi(X, Z) = \varphi_x(X) + Z\gamma,$$

that has been first studied by Robinson (1988) in the purely exogenous context. In this partially linear specification, we maintain the component of φ that depends on X unspecified and we model other exogenous controls to enter in a linear fashion. This modeling strategy is sufficiently flexible to allow us to include many exogenous controls without incurring in the curse of dimensionality. Moreover, the interpretation of the results from this model remains straightforward.

In this last section, we present an application of this partially linear specification to the estimation of the Engel curve for food. The empirical study is not original to this work: the goal of this session is, therefore, to discuss how nonparametric instrumental regressions can be flexibly embedded in models with several other exogenous controls, using a partially linear specification. The database is the one used in Bhalotra and Attfield (1998) and consists of 9740 rural households in Pakistan with less than 20 members.

The Engel curve relationship describes the expansion path for commodity demands as the household's budget increases. To estimate its shape, it is sufficient to regress the share of the household's budget spent for a given commodity (or group of commodities) over the total

[11] Notice, however, that the completeness condition for identification would be particularly strong as the joint dimension of X and Z increases. Our instrument should be sufficiently informative to capture the variation of all square-integrable functions of both X and Z.

[12] A recent paper by Ozabaci et al. (2014) looks into the estimation of additive models with endogenous regressors using a control function approach.

budget. However, as pointed out in Blundell et al. (2007), the total budget is likely to be determined jointly with the share of expenditure across consumption goods. Hence, it is an endogenous regressor. Blundell et al. (2007) suggest using other sources of income as a suitable instrument for total expenditure.

We thus denote as the random variable Y, the share of expenditure in a given consumption good; as X, the total log expenditure of the household; and, as W the log gross income of the household head.

Blundell et al. (2007) devise and apply a sieve minimum distance framework to the shape-invariant estimation of this curve using a sample of British households. This specification allows for a nonparametric modeling of the endogenous variable X, minus a parametric component which *scales* the function according to some household characteristics; and a linear parametric component, which explicitly controls for household's demographics. Bhalotra and Attfield (1998) use a partially linear model, in which X enters in a nonlinear fashion, and household's characteristics are modeled parametrically. In the results reported in the paper, they do not explicitly control for potential endogeneity of X. They claim that, when using a control function approach with W as a control variable, their results do not differ substantially. However, the control function is taken to be linear in W, while substantial nonlinearity may actually be present in the relation between income and total expenditure.

Here, we follow the partially linear specification of Bhalotra and Attfield (1998):

$$Y = \varphi(X) + Z\gamma + U, \qquad \mathbb{E}(U|W,Z) = 0, \tag{19}$$

where φ represents the shape of the Engel curve and Z is a vector of controls that enter the model linearly. These controls include the size of the household, the order of birth of male/female children and the literacy of the household head and his/her spouse. The main difference with respect to the original paper is that we are going to explicitly take into account the endogeneity of X and estimate the (possibly) nonlinear shape of the function using nonparametric instrumental regressions. In order to reduce the computational cost and some of the heterogeneity in the sample, we only consider households living in the region of Punjab. This choice is justified by the fact that this province accounts for around 60% of the sample and the results obtained in Bhalotra and Attfield (1998) are mostly driven by its demand paths. Moreover, we also trim 1% of our observations at low densities for Y. We, therefore, end up using a sample of 5635 observations.

In this database, food, as a broad aggregate of 82 commodities, accounts on average for about 51% of the total household expenditure in Punjab (see table 4).

	Mean	St.Dev	Min	Max
Log PC Expenditure	5.60	0.47	4.22	7.91
Log PC Income	5.62	0.51	3.98	8.00
Budget share food	0.51	0.10	0.17	0.75

Table 4. Summary statistics

In the original work of Bhalotra and Attfield (1998), it is shown that the Engel curve for food is decreasing, as predicted by Engel's law and has a quadratic shape. This latter result is of great interest as a quadratic Engel curve seems to be a feature of developing economies. However, as reported by Blundell et al. (2007), neglecting potential endogeneity in the estimation can lead to incorrect estimates of the Engel curve shape.

Our goal is to assess the robustness of previous results and provide some additional evidence using the nonparametric instrumental variable approach.

In order to estimate the partially linear model in (19), we use the following backfitting approach. For a given value of γ, φ is defined as the solution of the following equation:

$$\mathbb{E}\left[Y - Z\gamma | W\right] = \mathbb{E}\left[\varphi(X)|W\right].$$

In the same way, for a given value of φ, we can obtain γ as the OLS regression of the dependent variable $Y - \varphi(X)$ over Z. We iterate this backfitting approach until convergence of the following sample criterion:

$$SSR(\varphi, \gamma) = \sum_{i=1}^{N} \left(y_i - \varphi(x_i) - z_i\gamma\right)^2.$$

We obtain confidence bands for the nonparametric estimator of φ and standard errors for γ using pairwise bootstrap. We sample with replacement from the joint distribution of (Y, X, Z, W). This bootstrap procedure has been shown to be uniformly valid in the case of GK regularization by Horowitz and Lee (2012), although to the best of our knowledge, theoretical results about the validity of bootstrap confidence bands in Nonparametric IV using TK or LF regularizations do not yet exist. Therefore, for some of the estimators reported here, the confidence bands are only meant to give an idea about the variability of the estimation, but they should not be interpreted as valid.

We also compare the nonparametric specification with a parametric quadratic model, which also takes into account the endogeneity issue and is estimated using a control function approach:

$$Y = \beta_1 X + \beta_2 X^2 + Z\gamma + V\delta + U \tag{20}$$

$$X = \zeta(Z, W) + V \tag{21}$$

$$\mathbb{E}\left(U|X, V\right) = \mathbb{E}\left(U|V\right). \tag{22}$$

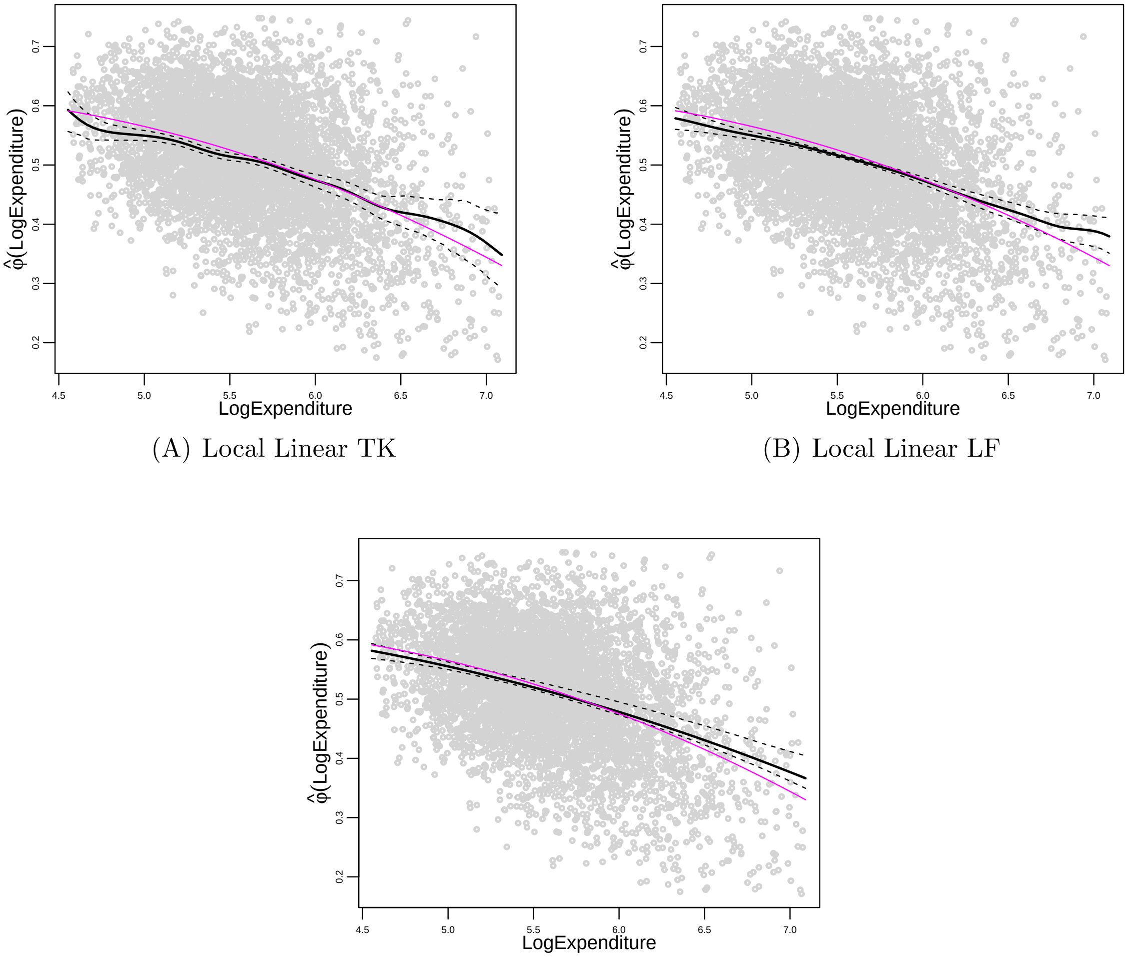
The link function ζ is estimated using local constant kernels and rule-of-thumb bandwidths, to reduce the computational time, given the high dimensionality of Z. The coefficients $(\beta_1, \beta_2, \gamma, \delta)$ are instead estimated using simple OLS. The results for the estimation of β_1, β_2 and δ are summarized in table (5), where the standard errors reported are heteroscedasticity robust. We can see that all coefficients are significant. The one associated with the quadratic component is very small but significantly negative.[13]

	Intercept	Log PC Exp	Log PC Exp^2	$\hat{V}$
Coefficient	0.39	0.14	-0.02	-0.06
Std.Error	0.14	0.05	0.00	0.01

TABLE 5. Results from model (20). Dependent variable: share of budget for food.

[13] The results for the estimation of γ in all models are not reported here.

The results of the semiparametric estimation of the Engel curve for Pakistan data are reported in Figure (7). For each estimator, we present the estimation outcome and its 95% bootstrap confidence intervals. Moreover, for graphical comparison, we also draw the quadratic fitting obtained using the control function approach in (20), which is represented by the continuous magenta line.



(A) Local Linear TK

(B) Local Linear LF

(C) Legendre GK

Figure 7. Estimation of the Engel curve for food in rural Pakistan. The black continuous line denotes the estimator of φ; the dashed black lines are the 95% confidence interval; and the continuous magenta line is the quadratic specification that uses control functions.

Our results are largely consistent across the different frameworks and also confirm the quadratic shape of the Engel curve found in Bhalotra and Attfield (1998). However, compared to the control function estimators, all the IV estimators show a much less pronounced quadratic shape. This can be due to the bias introduced by the regularization procedure.

In fact, the quadratic control function estimator is between the confidence bounds of the TK estimator, which has the largest variance within the IV estimators, as suggested by our simulation study and, therefore, possibly a smaller regularization bias. However, the control function estimator is outside the bootstrap confidence intervals for both GK and LF estimators, especially towards the boundary of the support. Hence, the nonparametric IV estimators seem to support the main result about the quadratic shape of the Engel curve. However, they point out that the marginal effects could be locally closer to being constant than those predicted by a control function approach.

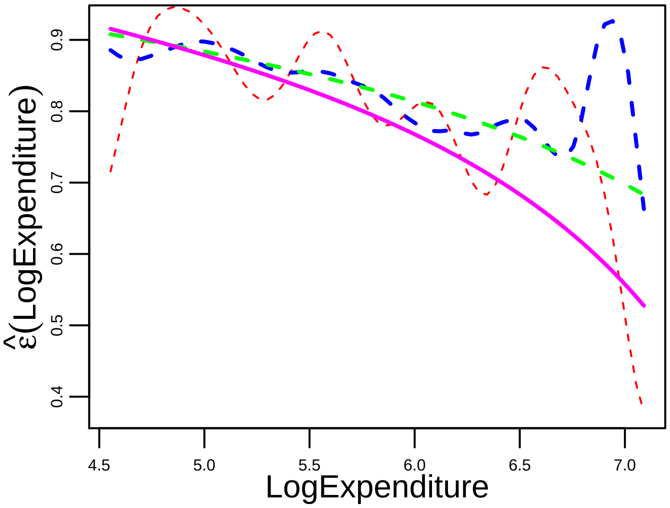


FIGURE 8. Estimation of expenditure elasticity. TK (dashed red line), LF (dashed blue line), GK (dashed green line) and quadratic control function (magenta line).

In order to explore this point further, we also consider the derivative of each estimator and compute an estimate of the expenditure elasticity ($\varepsilon(x)$) as:

$$\hat{\varepsilon}(x) = 1 + \frac{\hat{\varphi}'(x)}{\hat{\varphi}(x)}.$$

We briefly point out here that the derivative of each estimator is obtained using directly the first order coefficient of the local linear approximation, for TK and LF, and by differentiation of the Legendre polynomial basis for GK.[14]

The elasticities are again broadly consistent with existing evidence (see Figure 8). They are strictly included between 0 and 1, as food is a necessity good. They are also relatively

[14] A further remark is that we compute here a self-consistent first derivative of the estimator of φ and not an estimator of the first derivative, as it is usually achieved by Hilbert Scale or Sobolev penalization in the nonparametric instrumental regression literature (see Blundell et al., 2007, Chen and Pouzo, 2012, Florens et al., 2011, Florens and Racine, 2012).

higher than the values obtained for developed countries (Banks et al., 1997, find an average elasticity of 0.3 in a sample of UK households).

We can also see that the TK estimator displays more variation than the control function and the other nonparametric estimators. However, the figure also suggests that the control function approach may be very sensitive to outliers in the sample. In fact, the elasticity estimated from control functions follows the trend of the TK estimate, but without capturing the local variability present in the data, due to the quadratic restriction over the entire support of X. This may confirm our intuition that the shape of the control function estimator may be driven by outliers and, therefore, locally underestimate the elasticity.

6. Conclusions

This paper presents an overview of the practical implementation of nonparametric instrumental regressions. We consider the small sample properties of various estimators in a single endogenous covariate and single instrument framework. A simulation study shows the performances of these estimators and provides a review of some of the data-driven approaches that have been proposed so far for the selection of the regularization parameter. Finally, an application to the estimation of the Engel curve for food in a sample of households in rural Pakistan shows its practical usefulness and how additional exogenous covariates can be flexibly embedded using a partially linear specification.

Our intention is to give a unified and simple presentation of the several regularization procedures that can be considered when applied researchers would like to keep the flexibility of nonparametric estimation in the presence of endogenous regressors.

We present what we consider to be *the state-of-the-art* of the literature on Additive Nonparametric Instrumental Regression and we acknowledge that ongoing and future research can fill some of the blind spots of our work.

Our findings suggest that GK (sieve) regularization holds an advantage over TK and LF regularization. We guess that something may be lost in the sequential choice of the smoothing and the regularization parameters in the latter regularization schemes. Future research should, therefore, try to tackle the joint selection of these parameters in both the aforementioned regularization schemes.

Moreover, we have not deeply discussed any estimation procedure that accounts for more than one endogenous variable. While Florens et al. (2012) study a partially linear specification, as the one used in our empirical application, where Z could also contain endogenous regressors, we are not aware of any other existing work that studies a more flexible nonparametric approach for multiple endogenous variables. For instance, one could easily imagine an extension of the nonparametric additive regression model with both multiple endogenous and exogenous regressors.

Finally, while confidence intervals are useful to infer the variability of the estimator, a much wider set of inference procedures may be needed. A recent contribution in this respect is the work of Chen and Pouzo (2015), in which the authors present valid procedures for the construction of Wald and Quasi Likelihood Ratio tests within the framework of sieve regularization.

References

Ai, C. and Chen, X. (2003), 'Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions', *Econometrica* **71**(6), 1795–1843.

Andrews, D. W. K. (2011), 'Examples of L^2-Complete and Boundedly-Complete Distributions', *Cowles Foundation Discussion Paper* **1801**.

Banks, J., Blundell, R. and Lewbel, A. (1997), 'Quadratic Engel Curves and Consumer Demand', *Review of Economics and Statistics* **79**(4), pp. 527–539.

Bhalotra, S. and Attfield, C. (1998), 'Intrahousehold Resource Allocation in Rural Pakistan: a Semiparametric Analysis', *Journal of Applied Econometrics* **13**(5), 463–480.

Blundell, R., Chen, X. and Kristensen, D. (2007), 'Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves', *Econometrica* **75**(6), 1613–1669.

Breunig, C. and Johannes, J. (2015), 'Adaptive Estimation of Functionals in Nonparametric Instrumental Regressions', *Econometric Theory* **FirstView**, 1–43.

Canay, I. A., Santos, A. and Shaikh, A. M. (2013), 'On the Testability of Identification in Some Nonparametric Models with Endogeneity', *Econometrica* **81**(6), 2535–2559.

Cardot, H. and Johannes, J. (2010), 'Thresholding Projection Estimators in Functional Linear Models', *Journal of Multivariate Analysis* **101**(2), 395 – 408.

Carrasco, M., Florens, J.-P. and Renault, E. (2007), Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization, *in* J. Heckman and E. Leamer, eds, 'Handbook of Econometrics', Elsevier, pp. 5633–5751.

Centorrino, S. (2015), 'Data Driven Selection of the Regularization Parameter in Nonparametric Instrumental Regressions', *Mimeo - Stony Brook University* .

Chen, X. (2007), Large Sample Sieve Estimation of Semi-Nonparametric Models, *in* J. J. Heckman and E. E. Leamer, eds, 'Handbook of Econometrics', Vol. 6, Part B, Elsevier, pp. 5549 – 5632.

Chen, X., Chernozhukov, V., Lee, S. and Newey, W. K. (2014), 'Local Identification of Nonparametric and Semiparametric Models', *Econometrica* **82**(2), 785–809.

Chen, X. and Christensen, T. (2015), 'Optimal Uniform Convergence Rates and Adaptive Estimation of Nonparametric Instrumental Variables Models', *Mimeo - NYU* .

Chen, X. and Pouzo, D. (2012), 'Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals', *Econometrica* **80**(1), 277–321.

Chen, X. and Pouzo, D. (2015), 'Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models', *Econometrica* **83**(3), 1013–1079.

Darolles, S., Fan, Y., Florens, J. P. and Renault, E. (2011), 'Nonparametric Instrumental Regression', *Econometrica* **79**(5), 1541–1565.

D'Haultfoeuille, X. (2011), 'On the Completeness Condition in Nonparametric Instrumental Problems', *Econometric Theory* **27**, 460–471.

Engl, H. W., Hanke, M. and Neubauer, A. (2000), *Regularization of Inverse Problems*, Vol. 375 of *Mathematics and Its Applications*, Kluwer Academic Publishers, Dordrecht.

Fève, F. and Florens, J.-P. (2010), 'The Practice of Non-parametric Estimation by Solving Inverse Problems: the Example of Transformation Models', *Econometrics Journal* **13**(3), S1–S27.

Fève, F. and Florens, J.-P. (2014), 'Non Parametric Analysis of Panel Data Models with Endogenous Variables', *Journal of Econometrics* **181**(2), 151 – 164.

Florens, J.-P., Johannes, J. and Van Bellegem, S. (2011), 'Identification and Estimation by Penalization in Nonparametric Instrumental Regression', *Econometric Theory* **27**(3), 472–496.

Florens, J.-P., Johannes, J. and Van Bellegem, S. (2012), 'Instrumental Regressions in Partially Linear Models', *The Econometrics Journal* **15**(2), 304–324.

Florens, J.-P. and Racine, J. (2012), 'Nonparametric Instrumental Derivatives', *Mimeo - Toulouse School of Economics* .

Florens, J.-P. and Simoni, A. (2012), 'Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior', *Journal of Econometrics* **170**(2), 458 – 475.

Freyberger, J. (2015), 'On Completeness and Consistency in Nonparametric Instrumental Variable Models', *Mimeo - University of Wisconsin Madison* .

Gagliardini, P. and Scaillet, O. (2012), 'Tikhonov Regularization for Nonparametric Instrumental Variable Estimators', *Journal of Econometrics* **167**(1), 61 – 75.

Hall, P. and Horowitz, J. L. (2005), 'Nonparametric Methods for Inference in the Presence of Instrumental Variables', *Annals of Statistics* **33**(6), 2904–2929.

Härdle, W. (1990), *Applied Nonparametric Regression*, Econometric Society Monographs, Cambridge University Press.

Hoderlein, S. and Holzmann, H. (2011), 'Demand Analysis as an Ill-posed Inverse Problem with Semiparametric Specification', *Econometric Theory* **27**, 609–638.

Horowitz, J. L. (2011), 'Applied nonparametric instrumental variables estimation', *Econometrica* **79**(2), 347–394.

Horowitz, J. L. (2014*a*), 'Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter', *Journal of Econometrics* **180**(2), 158 – 173.

Horowitz, J. L. (2014*b*), Nonparametric Additive Models, *in* J. Racine, L. Su and A. Ullah, eds, 'The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics', Oxford University Press, pp. 129–148.

Horowitz, J. L. and Lee, S. (2012), 'Uniform Confidence Bands for Functions Estimated Nonparametrically with Instrumental Variables', *Journal of Econometrics* **168**(2), 175 – 188.

Horowitz, J. L. and Mammen, E. (2004), 'Nonparametric estimation of an additive model with a link function', *The Annals of Statistics* **32**(6), 2412–2443.

Hu, Y. and Shiu, J.-L. (2015), 'Nonparametric Identification Using Instrumental Variables: Sufficient Conditions For Completeness', *Mimeo - John Hopkins University* .

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998), 'Smoothing Parameter Selection in Nonparametric Regression using an Improved Akaike Information Criterion', *Journal of the Royal Statistical Society Series B* **60**, 271–293.

Johannes, J., Bellegem, S. V. and Vanhems, A. (2013), 'Iterative regularization in nonparametric instrumental regression', *Journal of Statistical Planning and Inference* **143**(1), 24–39.

Johannes, J. and Schwarz, M. (2011), 'Partially adaptive nonparametric instrumental regression by model selection', *Journal of the Indian Statistical Association* **49**(2), 149–175.

Kress, R. (1999), *Linear integral equations*, Applied mathematical sciences, Springer-Verlag.

Li, Q. and Racine, J. (2004), 'Cross-Validated Local Linear Nonparametric Regression', *Statistica Sinica* **14**, 485–512.

Li, Q. and Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Liu, C.-A. and Tao, J. (2014), 'Model Selection and Model Averaging in Nonparametric Instrumental Variables Models', *Working Paper* .

Mariano, R. S. (1972), 'The Existence of Moments of the Ordinary Least Squares and Two-Stage Least Squares Estimators', *Econometrica* **40**(4), pp. 643–652.

Müller, H.-G. (1991), 'Smooth Optimum Kernel Estimators Near Endpoints', *Biometrika* **78**(3), pp. 521–530.

Newey, W. K. (1997), 'Convergence rates and asymptotic normality for series estimators', *Journal of Econometrics* **79**(1), 147 – 168.

Newey, W. K. (2013), 'Nonparametric Instrumental Variables Estimation', *American Economic Review* **103**(3), 550–56.

Newey, W. K. and Powell, J. L. (2003), 'Instrumental Variable Estimation of Nonparametric Models', *Econometrica* **71**(5), 1565–1578.

Ozabaci, D., Henderson, D. J. and Su, L. (2014), 'Additive Nonparametric Regression in the Presence of Endogenous Regressors', *Journal of Business Economics and Statistics* **32**(4), 555–575.

Robinson, P. M. (1988), 'Root-N-Consistent Semiparametric Regression', *Econometrica* **56**(4), pp. 931–954.

Santos, A. (2012), 'Inference in nonparametric instrumental variables with partial identification', *Econometrica* **80**(1), 213–275.

Sokullu, S. (2015), 'A Semiparametric Analysis of Two-Sided Markets: An application to the Local Daily Newspapers in the U.S.', *Journal of Applied Econometrics* **Forthcoming**.

Vogel, C. (2002), *Computational Methods for Inverse Problems*, Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics.