Nonparametric Estimation of Semiparametric Transformation Models

Jean-Pierre FLORENS
Toulouse School of Economics

Senay SOKULLU
University of Bristol

July 11, 2012

Abstract

In this paper we develop a nonparametric estimation technique for semiparametric transformation models of the form: $H(Y) = \varphi(Z) + X'\beta + U$ where H, φ and β are unknown and the variables (Y, Z) are endogenous. Identification of the model and asymptotic properties of the estimator are analyzed under the mean independence assumption between the error term and the instruments. We show that the estimators are consistent and $\sqrt{N}$-convergence rate for $\hat{\beta}$ can be attained. The simulations demonstrate that our nonparametric estimates fits the data well.

Keywords: Nonparametric IV Regression, Inverse problems, Tikhonov Regularization, Regularization Parameter

JEL Classification: C13, C14, C30

1 Introduction

In this paper we focus on nonparametric estimation of a semiparametric transformation model. The model we study is given by the following relation:

$$H(Y) = \varphi(Z) + X'\beta + U, \quad \mathbb{E}[U|X, W] = 0 \tag{1}$$

where $Y, Z \in \mathbb{R}$ are endogenous, $X \in \mathbb{R}^q$ is exogenous, $W \in \mathbb{R}^p$ is a vector of instruments and $U \in \mathbb{R}$ is the error term. We aim to estimate the functions of interest, H and φ, and the parameter of interest, β, by nonparametric instrumental regression using the mean independence condition given in (1). We also study the identification of the model and asymptotic properties of the estimators.

The model given in (1) is a hybrid of transformation models and partially linear models that both has been studied extensively in econometrics. Transformation models have the form $H(Y) = X'\beta + U$. These models have been used in applied econometrics not only to improve the performance of the estimators but also to help to interpret the model. One well-known example is Box and Cox (1964) who propose a power transform of the dependent variable which may lead to normality in a linear regression. Horowitz (1996) gives other examples such as parametric and semiparametric proportional hazard rate model, log-linear regression and accelerated failure time models. Transformation models still get a lot of attention in econometrics, however, examples with nonparametric specifications are rare. A semiparametric partially linear model can be written as $Y = \varphi(Z) + X'\beta + U$. Use of partially linear models in applied econometrics is especially common when it is not clear how to specify the effect of one variable parametrically. Florens, Johannes, and Van Bellegem (2009) study the estimation of β in $Y = \varphi(Z) + X'\beta + U$. Their main example is the model of Engle, Granger, Rice, and Weiss (1986) in which the effect of temperature is specified nonparametrically in the electricity demand. A more recent example for partially linear specification comes from Bontemps, Simioni, and Surry (2008) where they look at the impact of agricultural pollution on the prices of residential houses and use a nonlinear nonparametric specification for the effect of pollution on the price of houses.

In this paper we study the nonparametric estimation of a semiparametric transformation model in equation (1) which includes nonparametric specifications on both sides of the equation. Hence we are extending transformation models to a general case where the transformation of the dependent variable is specified nonparametrically and the right hand side of the equation includes a parametric as well as a nonparametric part. The equation we propose to estimate is motivated by the estimation of demand systems in network industries, where the effect of size of the network on demand can be ambiguous. For a brief illustration, we consider the example of the magazine market. The magazine market is a two-sided market where the magazine is a platform serving to readers and advertisers. The demands of two end users depend on each other and hence indirect network externalities exist. Since the advertisers would like to reach as many readers as they can, they would prefer a magazine with many readers. Additionally, if the readers like advertisements, they would like to read a magazine with more advertisements. However, when the number of advertising pages increases too much in a magazine, it may have a nuisance effect on the readers and the network effect may start to decrease and even become negative. If we want to model the demand of readers for the magazine, it is better to specify this indirect network effect nonparametrically to be able to capture nonlinearities and nonmonotonicities. Assuming that the demand function of readers is additive in its arguments, we can write[1]:

$$Y = F(\varphi(Z) + X'\beta + U) \tag{2}$$

where Y and Z are the market shares of the magazine on the readers' and advertisers' side, respectively. F is the demand function of readers. $\varphi(Z)$ is the network externality function that depends on the number of advertisers, in other words, it is the effect of number of advertisements on readers' demand for the magazine. X are observed and U are unobserved

[1] For more information on the derivation of the demand equation see Bass (1969).

magazine characteristics for readers. In Kaiser and Song (2009), X includes number of content pages, cover price and frequency of the magazine while U is assumed to be a content related quality shock. β is a parameter to be estimated. Under the assumption that the demand is one-to-one, we can take the inverse of the demand function and obtain equation (1): $F^{-1}(Y) = \varphi(Z) + X'\beta + U$ where $H(Y) = F^{-1}(Y)$. Note that this specification allows us to specify both the demand and the network effect functions nonparametrically.

To the best of our knowledge estimation of equation (1) has not been studied yet nor has it been used in an empirical application. We present the identification, estimation and asymptotic properties of this model using the mean independence condition between the error terms and instruments. So, we are not only introducing a very general form of nonparametric model but also using a relatively weak condition in its estimation. The estimation we propose depends on nonparametric instrumental variable regression. It is well known in the nonparametric IV literature that this estimation problem is an ill-posed inverse problem. More broadly, the solution of our main identifying equation needs the inversion of an infinite dimensional operator with infinitely many eigen values which are very close to zero. Hence, it needs a modification, or in the terminology of ill-posed inverse problems, we need to regularize the problem. In this paper, we solve the ill-posed inverse problem we encounter by *Tikhonov Regularization* which can be thought of as the nonparametric counterpart of *Ridge Regression*. We show that we get consistent estimators as well as a $\sqrt{N}$-convergence rate for β with nonparametric IV regression.

We investigate the performance of our estimation procedure by means of a Monte Carlo simulation. Since we regularize the ill-posed inverse problem in the estimation, the practical implementation requires the choice of 2 tuning parameters: Bandwidth and the regularization parameter. We present a way to choose the optimal regularization parameter and use it in the simulations for a given bandwidth. Simulations show that, when the regularization parameter is chosen optimally, our estimated curves fit the actual ones very well. However, in cases where we choose the regularization parameter arbitrarily, we may have very oscillating or very flat curves, as the theory suggests. This result also proves the importance of the selection of the regularization parameter in inverse problems which is encountered very often in nonparametric estimation.

This paper differs from the existing literature in the sense that it covers a very general case, as it considers a semiparametric transformation model. When the nonparametric estimation is considered, Darolles, Fan, Florens, and Renault (2011), Horowitz (2011), Hall and Horowitz (2005), Newey and Powell (2003) and Ai and Chen (2003) are the first papers coming to mind as related literature. Darolles, Fan, Florens, and Renault (2011), Horowitz (2011), Hall and Horowitz (2005) and Newey and Powell (2003) consider models without finite dimensional parameters and all use the nonparametric IV regression to estimate the functions of interest using the mean independence condition. We also use the mean independence between the error term and instruments to identify and estimate the equations of interest. Florens, Johannes, and Van Bellegem (2009) and Ai and Chen (2003) consider the partially linear, semiparametric model. The former uses nonparametric instrumental variables regression based on kernels and get over the ill-posed inverse problem by Tikhonov regularization, while the latter restricts the set of functions to be a compact set to avoid the inverse problem and estimates the parameter and functions of interest by minimum distance sieve estimation. We follow the approach of Florens, Johannes, and Van Bellegem (2009) and recover our functions of interest by regularizing the ill-posed inverse problem. In most examples of transformation models, parametric models are used and the estimations are also done parametrically. Horowitz (1996) presents semi parametric estimation of transformation models, though it makes a parametric specification for the right hand side. Linton, Sperlich, and Keilegom (2008) also estimate a semiparametric transformation model but they assume a parametric transformation of the dependent variable. Chiappori, Komunjer, and Kristensen (2011) show the identification and estimation of a nonparametric transformation model where they specify the equations on both sides nonparametrically. Different from our paper, they do not have a partially linear model and they assume that the error terms are independent of the exogenous variables conditional on the endogenous variables. Feve and Florens (2010) estimate a simplified version of our model with nonparametric instrumental regression, where they have a nonparametric transform explained by a parametric linear model. Therefore, this paper is generalizing the nonparametric transformation models. Moreover, compared to all the aforementioned papers, we are using a weaker assumption on the error terms and instruments.

The paper is organized as follows: In *Section 2* we introduce a simple model where $X \in \mathbb{R}$ and β is normalized to 1. After studying the identification, estimation and asymptotic properties of this simpler model we generalize it to the model in equation (1) in *Section 3*. A data based method for the selection of optimal regularization parameter is discussed in *Section 4* while we present the results of a small Monte Carlo simulation exercise in *Section 5*. Finally, *Section 6* concludes. All the proofs are presented in the appendices.

2 A Semiparametric Transformation Model and Its Nonparametric Estimation

In this section we study a simpler version of the transformation model in (1) for ease of exposition. In this simpler version we restrict X to be a scalar and normalize β to 1. The relationship between the variables is given by the following equation:

$$H(Y) = \varphi(Z) + X + U \tag{3}$$

$$\mathbb{E}[U|X, W] = 0$$

This is a semiparametric transformation model in which we have two endogenous variables, $Y, Z \in \mathbb{R}$, and an exogenous variable $X \in \mathbb{R}$. $U \in \mathbb{R}$ is the error term and $W \in \mathbb{R}^p$ is a vector of instruments. Moreover, $H(.)$ is a one-to-one monotone function.

Y, Z, X, W generate the random vector Ξ which has a cumulative distribution function

F. Then for each F, we can define the subspaces of real valued functions as $L^2_F(Y)$, $L^2_F(Z)$, $L^2_F(X)$ and $L^2_F(W)$ which depend only on Y, Z, X and W, respectively, and which belong to a common Hilbert space denoted by L^2_F. [2]

2.1 Identification

The identification of the model is based on the conditional independence of the error term and the instruments rather than full independence. Hence, our approach differs from the existing literature not only by the nonparametric specification of functions on both sides of the transformation model but also by a weaker assumption for identification.

We consider the random vector Ξ defined above. We assume that this vector satisfies the following assumptions:

Assumption 1 *There exists two square integrable functions H and φ such that:*

$$H(Y) = \varphi(Z) + X + U$$

with

$$\mathbb{E}[U|X, W] = 0$$

We have normalized the equation by assuming that the coefficient of X is equal to 1. Under this constraint, we want to consider the unicity of H and φ under the mean independence condition. In order to verify this unicity we assume two regularity conditions on the joint distribution of (Y, Z, X, W).

Assumption 2 ***Completeness.*** *The distribution of (Y, Z) given (X, W) is complete in*

[2]Throughout the paper, all function spaces are assumed to be L^2 spaces relative to the density of the data generating process. This choice of L^2 space is motivated by two reasons: First, the conditional expectation operator is well-defined in an L^2 space and second (different from the L^p spaces where $p \neq 2$) the L^2 spaces are Hilbert spaces which simplifies the use of adjoint operators. A theory in Banach spaces may be developed but it would be more abstract and not really motivated by applications. The choice of the density for the L^2 definition is also motivated by the simplicity of the computation of the adjoint operators.

the following sense:

$$\forall m(Y,Z) \in L^2_F(Y) \times L^2_F(Z), \quad \mathbb{E}[m(Y,Z)|X,W] = 0 \quad a.s \quad \Rightarrow \quad m(Y,Z) = 0 \quad a.s$$

This assumption (also called *strong identification*) has a long history in statistics in the analysis of the relation between sufficiency and ancillarity (See Florens, Mouchart, and Rolin, 1990, Chapter 5) and it is essential in the study of instrumental variables estimation (See Darolles, Fan, Florens, and Renault, 2011; Feve and Florens, 2010; Florens, Johannes, and Van Bellegem, 2009; Newey and Powell, 2003). More recently DHaultfoeuille (2011), Hu and Shiu (2011) and Andrews (2011) have analyzed this assumption and the primitive conditions that lead to the complete distributions.[3] From an intuitive point of view, this assumption can be seen as the nonparametric counterpart of *rank condition* in the parametric IV estimation.

Assumption 3 ***Separability.*** *Y and Z are measurably separable i.e., $\forall m(Y) \in L^2_F(Y)$ and $\forall l(Z) \in L^2_F(Z)$:*

$$m(Y) = l(Z) \Rightarrow m(.) = l(.) = constant$$

Assumption 3 is also standard in nonparametric estimation. It means that there is not an exact relation between Y and Z, or put it differently, $X+U$ in equation (3) is not equal to a constant. It is essentially a support condition on Y and Z, and it prevents the existence of a deterministic relation between Y and Z. In particular, if the support of the joint distribution of Y and Z is the product of the two marginal supports *Assumption 3* is satisfied. A more precise analysis of measurable separability condition is given in Florens, Heckman, Meghir, and Vytlacil (2008).

Finally, we want to normalize the function φ:

Assumption 4 ***Normalization.*** *If $\varphi(Z)$ is constant a.s. then $\varphi(Z) = 0$ a.s.*
For simplicity, we will assume that $\varphi(.)$ is normalized by the condition $\mathbb{E}[\varphi(Z)] = 0$. Under

[3] In the appendices, we present a discussion about this assumption

this assumption, we consider as the parametric space:

$$\mathcal{E}_0 = (H, \varphi) \in L^2_F(Y) \times L^2_F(Z) \quad such \quad that \quad \mathbb{E}[\varphi(Z)] = 0$$

Theorem 1 *Under the assumptions 1-4, the functions $H(Y)$ and $\varphi(Z)$ are identified.*

We may remark that these assumptions can be weakened in some cases. Imagine that $Y \perp W|W_1$ and $Z \perp W|W_2$. This means that the instruments can be grouped into 2 components acting separately on Y and Z. We assume also that W_1 and W_2 are measurably separable which in particular means that W_1 and W_2 have no elements in common. In this case:

$$\mathbb{E}[H(Y) - \varphi(Z)|W] = 0 \Rightarrow \mathbb{E}[H(Y)|W_1] = \mathbb{E}[\varphi(Z)|W_2] = c$$

where c is a constant and it is equal to 0 because $\mathbb{E}[\varphi(z)] = 0$. Then if Y is strongly identified by W_1 and Z is strongly identified W_2, we get the identification result.

2.2 Estimation

After showing that the model is identified, we can now continue with the estimation. Let us define the operator:

$$T : \mathcal{E}_0 = \left\{ L^2_F(Y) \times \tilde{L}^2_F(Z) \right\} \mapsto L^2_F(X, W) : T(H, \varphi) = \mathbb{E}[H(Y) - \varphi(Z)|X, W]$$

where $\tilde{L}^2_F(Z) = \{\varphi \in L^2_F(Z) | \mathbb{E}(\varphi) = 0\}$, and the inner product is defined as[4]:

$$\langle (H_1, \varphi_1), (H_2, \varphi_2) \rangle = \langle H_1, H_2 \rangle + \langle \varphi_1, \varphi_2 \rangle$$

The adjoint operator of T, T^*, satisfies:

$$\langle T(H, \varphi), \psi \rangle = \langle (H, \varphi), T^* \psi \rangle$$

[4]See appendix for the verification of our inner product definition.

for any $(H,\varphi) \in \mathcal{E}$ and $\psi \in L^2_F(X,W)$. From this equality it follows immediately that

$$T^*\psi = (\mathbb{E}[\psi|Y], \mathbb{E}[\psi|Z])$$

However, as already defined, our parametric space is $\mathcal{E}_0$. Let us denote the restriction of T to $\mathcal{E}_0$ by T_0 ($T_0 = T_{|\mathcal{E}_0}$) and the projection of $\mathcal{E}$ under $\mathcal{E}_0$ by $\mathbb{P}$, $\{\mathbb{P}(H,\varphi) = (H, \varphi - \mathbb{E}(\varphi))\}$. Then we have the following lemma to characterize the adjoint operator T^* of T:

Lemma 2 *Let us define the operator $T_0 : \mathcal{E} \to \mathcal{F}$ with the adjoint $T_0^* : \mathcal{F} \to \mathcal{E}$. Moreover, let us define $T = T_{0|\mathcal{E}_0}$, where $\mathcal{E}_0 \in \mathcal{E}$. Then, $T^* = \mathbb{P}T_0^*$ where $\mathbb{P}$ is the projection operator on $\mathcal{E}_0$.*

Then we can write the adjoint operator of T as:

$$T^* = \begin{pmatrix} \mathbb{E}(\phi|Y) \\ \mathbb{P}\mathbb{E}(\phi|Z) \end{pmatrix}$$

where $\mathbb{P}$ is the projection of $L^2_F(Z)$ on $\tilde{L}^2_F(Z)$ ($\mathbb{P}\varphi = \varphi - \mathbb{E}(\varphi)$).

Now, we can rewrite our estimation problem as:

$$T(H,\varphi) = r \tag{4}$$

where $r = \mathbb{E}[X|X,W]$.

Estimation of H and φ requires nonparametric estimation of operator T which has an infinite dimensional range and in general is compact.[5] This, in turn, gives us an ill-posed inverse problem, since the inversion of the estimator of T, lead to noncontinuities of the resulting estimators with respect to joint distribution of data.[6] To get a stable solution, we

[5]The compactness is satisfied in particular when the joint density Ξ is square integrable. (See Darolles, Fan, Florens, and Renault, 2011)

[6]Engl, Hanke, and Neubauer (1996) define a problem as well-posed if the conditions below hold:
(i) For all admissible data a solution exist.
(ii) For all admissible data the solution is unique.
(iii) The solution continuously depends on the data.

therefore need to regularize our problem. For this we have chosen the Tikhonov Regularization as it is easy to work with. Basically, we control the norm of the solution by a penalty term, α, which we call *the regularization parameter*.[7] The solution of (4) is then given by the minimization of the following problem:

$$\min_{H,\varphi}(\|r - T(H,\varphi)\|^2 + \alpha \|(H,\varphi)\|^2) \tag{5}$$

Thus,

$$(H(Y), \varphi(Z))' = (\alpha I + T^*T)^{-1}T^*X \tag{6}$$

where I is the identity operator in $L^2_F(Y) \times L^2_F(Z)$. Note that, we do not perform the minimization on the estimated operators. Instead, we first solve the inverse problem, thus minimize the norm with a penalty and perform the estimation on the solution. We can write the solution in (6) as follows:

$$(\alpha I + T^*T)(H,\varphi) = T^*X$$

Equivalently,

$$\begin{pmatrix} \alpha H + \mathbb{E}\left[\mathbb{E}(H|X,W)|Y\right] - \mathbb{E}\left[\mathbb{E}(\varphi|X,W)|Y\right] \\ -\alpha\varphi + \mathbb{PE}\left[\mathbb{E}(H|X,W)|Z\right] - \mathbb{PE}\left[\mathbb{E}(\varphi|X,W)|Z\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X|Y) \\ \mathbb{PE}(X|Z) \end{pmatrix} \tag{7}$$

As we do not know the true distribution of our variables, we need to estimate them first. So, we estimate conditional expectations with kernels. As a result, this brings about the second source of distortion in our problem. The first one is due to the regularization parameter α, and the second one is coming from the bandwidths of the kernels. We need to make two remarks here. First, although in equation (7) it seems that we use the same α's for the regularization of different equations, they do not necessarily be the same. In fact, in

[7] The choice of α is very important since it characterizes the balance between the fitting and the smoothing, and in the following sections we will introduce a data based selection rule for it.

the simulations we perform, we have seen that we can not get good fits for equal values of α's. Second, as is shown in Darolles, Fan, Florens, and Renault (2011), the dimension of instruments does not have a negative effect on the speed of convergence, on the contrary, the speed of convergence increases with the dimension of instruments. Therefore, if we have a large number of instruments this will not cause a problem through the curse of dimensionality, but it will instead increase the speed of convergence of our estimator.

Remember that $T(H, \varphi) = \int (H(y) - \varphi(z))f(y, z|x, w)dydz$. Denote $\hat{f}$ as the nonparametric kernel estimator of f. We may define $\hat{T}(H, \varphi)$ by replacing f by $\hat{f}$ in this expression. Analogously,

$$T^*(\psi) = \left(\int \psi(x, w)f(x, w|y)dxdw, \mathbb{P} \int \psi(x, w)f(x, w|z)dxdw \right)'$$

may be estimated by replacing f by its estimator. Let $\hat{T}^*$ be this estimator. In order to avoid numerical integration $\hat{T}$ (and $\hat{T}^*$) would be approximated by:

$$\hat{T}(H, \varphi) \simeq \frac{\sum_j (H(y_j) - \varphi(z_j))K_x\left(\frac{x-x_j}{h_x}\right)K_w\left(\frac{w-w_j}{h_w}\right)}{\sum_j K_x\left(\frac{x-x_j}{h_x}\right)K_w\left(\frac{w-w_j}{h_w}\right)}$$

for some bandwidth parameters h_y, h_z, h_x and h_w. The order of the approximation error is analogous to the order of the bias of the kernel estimation and it does not change the speed of convergence.

Let $A_{xw}(w)$ be the matrix whose (i,j)th element is:

$$A_{xw}(w)(i, j) = \frac{K_x\left(\frac{x_i-x_j}{h_x}\right)K_w\left(\frac{w-w_j}{h_w}\right)}{\sum_j K_x\left(\frac{x_i-x_j}{h_x}\right)K_w\left(\frac{w-w_j}{h_w}\right)}$$

A_y and A_z are the matrices with the (i,j)th elements:

$$A_y(i,j) = \frac{K_y\left(\frac{y_i - y_j}{h_y}\right)}{\sum_j K_y\left(\frac{y_i - y_j}{h_y}\right)}$$

$$A_z(i,j) = \frac{K_z\left(\frac{z_i - z_j}{h_z}\right)}{\sum_j K_z\left(\frac{z_i - z_j}{h_z}\right)}$$

Moreover, P is the matrix with $\frac{n-1}{n}$ on the diagonal and $-\frac{1}{n}$ elsewhere. Our estimated functions are the solutions of the following system:

$$\begin{pmatrix} \alpha_N \hat{H} + A_y A_{xw} \hat{H} - A_y A_{xw} \hat{\varphi} \\ -\alpha_N \hat{\varphi} + P A_z A_{xw} \hat{H} - P A_z A_{xw} \hat{\varphi} \end{pmatrix} = \begin{pmatrix} A_y X \\ P A_z X \end{pmatrix}$$

More precisely,

$$\begin{pmatrix} \hat{H} \\ \hat{\varphi} \end{pmatrix} = \begin{pmatrix} \alpha_N I + A_y A_{xw} & -A_y A_{xw} \\ P A_z A_{xw} & -(\alpha_N I + P A_z A_{xw}) \end{pmatrix}^{-1} \begin{pmatrix} A_y X \\ P A_z X \end{pmatrix} \tag{8}$$

Equation (8) is a system of $2N$ equations in $2N$ unknowns which means that we can recover $\hat{H}$ and $\hat{\varphi}$. Here we denote α by α_N, where N is the sample size, since in the estimation the optimal value of the regularization parameter depends on the sample size, see Engl, Hanke, and Neubauer (1996).

The solution given by equation (8) requires the inversion of a $2N \times 2N$ matrix. So, with very large N, this will be difficult to compute. In cases with large N, instead of Tikhonov regularization, we can use the *Landweber-Friedman* regularization scheme that does not need the inversion of this matrix, see Carrasco, Florens, and Renault (2007).

2.3 Consistency and Rate of Convergence

In our estimation process, we are estimating the functions through the estimation of the operators. For this reason, to be able to talk about the consistent estimation of the functions of interest, first we have to estimate the operators T^*T and T^*X consistently. To show this, we are going to make a set of assumptions.

Let us begin with the definition of *singular value decomposition*.

Definition 1 *Let $\{\lambda_j, \phi_j, \psi_j\}$ be the singular system of the operator T such that:*

$$T\phi_j = \lambda_j \psi_j \quad and \quad T^*\psi_j = \lambda_j \phi_j$$

where λ_j denote the sequence of the nonzero singular values of the compact linear operator T, ϕ_j and ψ_j, for all $j \in \mathbb{N}$, are orthonormal sequences of functions in $\mathcal{E}$ and $L^2_F(X, W)$, respectively. We can moreover write the singular value decomposition for each $\varphi \in \mathcal{E}$:[8]

$$T\varphi = \sum_{j=1}^{\infty} \lambda_j \langle \varphi, \phi_j \rangle \psi_j$$

Assumption 5 *Source Condition: There exists $\nu > 0$ such that:*

$$\sum_{j=1}^{\infty} \frac{\langle \Phi, \phi_j \rangle^2}{\lambda_j^{2\nu}} = \sum_{j=1}^{\infty} \frac{[\langle H, \phi_{j1} \rangle + \langle \varphi, \phi_{j2} \rangle]^2}{\lambda_j^{2\nu}} < \infty$$

where $\Phi = (H, \varphi)$.

By this assumption we define a regularity space for our functions. In other words, as stated in Carrasco, Florens, and Renault (2007), we can say that the unknown value of $\Phi_0 = (H_0, \varphi_0)$ belongs to the space Ψ_ν where

$$\Psi_\nu = \left\{ \Phi \in \mathcal{E} \quad such \quad that \quad \sum_{j=1}^{\infty} \frac{\langle \Phi, \phi_j \rangle^2}{\lambda_j^{2\nu}} < \infty \right\}$$

[8] For more on singular value decomposition, see Carrasco, Florens, and Renault (2007).

In fact, assuming that $\Phi_0 \in \Psi_\nu$ just adds a smoothness condition to our functional parameter of interest. As was pointed out by Carrasco, Florens, and Renault (2007), this regularity assumption will give us an advantage in calculating the rate of convergence of the regularization bias.

Assumption 6 *There exists $s \geq 2$ such that:*

- $\left\|\hat{T} - T\right\|^2 = O\left(\frac{1}{Nh_N^{p+2}} + h_N^{2s}\right)$
- $\left\|\hat{T}^* - T^*\right\|^2 = O\left(\frac{1}{Nh_N^{p+2}} + h_N^{2s}\right)$

where s is the minimum between the order of the kernel and the order of the differentiability of f, p is the dimension of the instrument vector W and h_N is the bandwidth.

Assumption 7

$$\left\|\hat{T}^* X - \hat{T}^*\hat{T}\Phi\right\|^2 = O\left(\frac{1}{N} + h_N^{2s}\right)$$

Assumption 8

$$\lim_{N\to\infty} \alpha_N = 0$$

$$\lim_{N\to\infty} \frac{h_N^{2s}}{\alpha_N^2} = 0$$

$$\lim_{N\to\infty} \alpha_N^{2-\nu} N h_N^{p+2} \to \infty \quad or \quad \nu \geq 2$$

$$\lim_{N\to\infty} \alpha_N^2 N \to \infty$$

$$\lim_{N\to\infty} N h_N^{p+2} \to \infty$$

Theorem 3 *Let us define $\Phi = (H(Y), \varphi(z))$. Let s be the minimum between the order of the kernel and the order of the differentiability of f and ν be the regularity of Φ. Under assumptions 5 to 8:*

- $\left\|\hat{\Phi}_N^\alpha - \Phi\right\|^2 = O\left(\frac{1}{\alpha^2}\left(\frac{1}{N} + h_N^{2s}\right) + \frac{1}{\alpha^2}\left(\frac{1}{Nh_N^{p+2}} + h_N^{2s}\right)\left(\alpha^{\min\{\nu,2\}}\right) + \alpha^{\min\{\nu,2\}}\right)$
- $\left\|\hat{\Phi}_N^\alpha - \Phi\right\| \to 0$ *in probability.*

Optimal speed of convergence is obtained by the calculation of optimal α. To do this we equalize the first and the third term of the rate of convergence above, as the second term is negligible. Then we obtain that the optimal α_N is proportional to $N^{-1/[\min\{\nu,2\}+2]}$. Moreover, under the assumption that $h^{2s} = O_p(1/N)$ if the conditions $[(p+2)(\nu+2)]/2\nu \leq s$ when $\nu \leq 2$, and $(p+2)/4 \leq s$ when $\nu > 2$, are satisfied, we can obtain the following optimal speed of convergence:

$$\left\| \hat{\Phi}^{\alpha} - \Phi \right\|^2 \sim O\left(N^{-\frac{\min\{\nu,2\}}{\min\{\nu,2\}+2}} \right)$$

This rate of convergence follows from an argument similar to that of Darolles, Fan, Florens, and Renault (2011). Under more specific assumptions (for example, geometric rate of decline of λ_j, $\langle H, \varphi_{j1} \rangle$ and $\langle \varphi, \varphi_{j2} \rangle$ as in Hall and Horowitz (2005)) it may be improved upon, and shown to be minimax, see Chen and Reiss (2007); Breunig and Johannes (2009). As we want to focus on the semiparametric specification, we do not reproduce this discussion which is not specific to our model.

3 Semiparametric Transformation Model: The General Case

In this section, we generalize the simple model of Section 2. We examine the identification and estimation of H, φ and β in equation (1) as well we study the asymptotic properties of the estimators. We show that, in the semiparametric transformation models with many explanatory variables, we can get $\sqrt{N}$-consistency for the estimated parameters. In other words, we show that the nonparametrically estimated parameters of a partially linear transformation model can still attain $\sqrt{N}$-convergence rate. Remember that, equation (1) is:

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

For identification, we need to normalize one element of the vector β to 1. Then, we can write the model in (1) as the following:

$$H(Y) = \varphi(Z) + X_0 + X_1'\beta + U \tag{9}$$

where $X = \{X_0, X_1\}$. Moreover, $Y, Z, X_0, U, V \in \mathbb{R}$, $X \in \mathbb{R}^q$ and $W \in \mathbb{R}^p$.

Since the estimation and asymptotic properties of the nonparametric parts are already discussed in *Section 2*, in this section we restrict our attention to β.

3.1 Identification

Identification of this general model is not very different from the previous one, nonetheless we need some additional assumptions.

Assumption 9 *(Y,Z,X) are strongly identified by (X,W).*[9]

$$\mathbb{E}[g(Y, Z, X)|X, W] = 0 \Rightarrow g(Y, Z, X) = 0 \quad a.s.$$

Remark: Assumption 9 may be weakened by considering only the function $g(Y, Z, X)$ satisfying:

$$g(Y, Z, X) = g_1(Y, Z) + g_2(X)$$

Assumption 10 (Y, Z) *and* X_1 *are measurably separable:*

$$m(Y, Z) = l(X_1) \Rightarrow m(.) = l(.) = constant$$

Theorem 4 *Under the assumptions 1-4, 9 and 10 the functions* $H(Y)$ *and* $\varphi(Z)$ *and the parameter* β *are identified.*

[9]This assumption can also be stated as: (Y, Z) are strongly identified by W conditional on X.

3.2 Estimation

We can now proceed with the estimation. Let us keep the operator T the same as in *Section 2*, and introduce an additional operator $T_X : \mathbb{R}^k \to L^2_F(X,W) : \beta \mapsto X_1'\beta$. Equivalently its adjoint is defined $T^*_X : L^2_F(X,W) \to \mathbb{R}^k : g \mapsto \mathbb{E}[X_1 g(X,W)]$ which follows from the following relation:

$$\langle T_X\beta, g(X,W)\rangle = \langle \beta, T^*_X g(X,W)\rangle$$

Then we can write:

$$T(H,\varphi) - T_X\beta = X_0 \tag{10}$$

The normal equations are:

$$T^*T(H,\varphi) - T^*T_X\beta = T^*X_0 \tag{11}$$

$$T^*_X T(H,\varphi) - T^*_X T_X\beta = T^*_X X_0 \tag{12}$$

From equation (11), we get $(H,\varphi) = (\alpha_N I + T^*T)^{-1}(T^*T_X\beta + T^*X_0)$ and if we substitute it into equation (12), we obtain an expression for the β:

$$\beta = (T^*_X(P^\alpha - I)T_X)^{-1}T^*_X(I - P^\alpha)X_0$$

where $P^\alpha = T(\alpha I + T^*T)^{-1}T^*$. Then the estimator is given by:

$$\hat{\beta} = \left(\widehat{T^*_X T}(\alpha_N I + \hat{T}^*\hat{T})^{-1}\widehat{T^*T_X} - \widehat{T^*_X T}\right)^{-1}\left(T_X - \widehat{T^*_X T}(\alpha_N I + \hat{T}^*\hat{T})^{-1}\widehat{T^*T_X}\right)X_0$$

Note that the α_N in this generalized version and the α_N in the simple version need not necessarily be the same.

3.3 Consistency and Rate of Convergence of $\hat{\beta}$

We can continue with the asymptotic properties of $\hat{\beta}$. In a semiparametric context the $\sqrt{N}$-convergence of the parametric component is a standard question (See Ichimura and Todd, 2007), and is generally addressed in cases where the nonparametric component is a density or a regression function. Usually this $\sqrt{N}$-convergence requires assumptions to distinguish the nonparametric and parametric part of the model. In this paper, the nonparametric component is estimated by solving an inverse problem. In the sequel we present the assumptions that are needed to prove the parametric rate of convergence.

Let $\{\lambda_j, \phi_j, \psi_j\}$ be the singular system of the operator T as defined before and let $\{\mu_j, e_j, \tilde{\psi}_j\}$ be the singular system of the operator T_X, such that for each $\beta \in \mathbb{R}^k$ we can write:

$$T_X \beta = \sum_{j=1}^{k} \mu_j \langle \beta, e_j \rangle \tilde{\psi}_j$$

Assumption 11 *Source Condition: There exists* $\eta > 0$ *such that:*

$$\max_{i=1,\ldots,k} \sum_{j=1}^{\infty} \frac{\left\langle \tilde{\psi}_i, \psi_j \right\rangle^2}{\lambda_j^{2\eta}} < \infty$$

This source condition explains the collinearity between (Y, Z) and (X). Indeed, *Assumption 11* is false if the range of T is included in the linear space generated by the elements of X_1. In contrast, if the range of T (the space of $\mathbb{E}(H|X, W) - \mathbb{E}(\varphi|X, W)$ for all H and φ) is orthogonal to X_1, then *Assumption 11* is directly satisfied because the term $\left\langle \tilde{\psi}_i, \psi_j \right\rangle$ cancels out. This assumption says that the degree of collinearity is not too high compared to the singular values of T. In other words, any linear function of X_1 has Fourier coefficients in the basis ψ_j declining sufficiently fast. In fact, the values of η gives a measurement of collinearity, i.e., $\eta = 0$ if there is perfect collinearity and $\eta = \infty$ if there is no collinearity.

Assumption 12 *There exists $s \geq 2$ such that:*

$$\left\| \widehat{T^* T_X} - T^* T_X \right\|^2 = O\left(\frac{1}{N h_N} + h_N^{2s} \right)$$

where s is the minimum between the order of the kernel and the order of the differentiability of f.

Assumption 13

$$\left\| \widehat{T_X^* T_X} - T_X^* T_X \right\|^2 = O\left(\frac{1}{N} \right)$$

$$\left\| \widehat{T_X^* T} - T_X^* T \right\|^2 = O\left(\frac{1}{N} \right)$$

Assumption 14

$$\lim_{N \to \infty} \alpha_N^{\min\{\eta,2\}} h^{2s} = 0$$

$$\lim_{N \to \infty} \frac{\alpha_N^{\min\{\eta,2\}}}{N h^{p+q+1}} \to 0$$

$$\lim_{N \to \infty} \frac{\alpha_N^{\min\{\eta,2\}}}{N} \to 0$$

$$\lim_{N \to \infty} \frac{h^{2s}}{N} \to 0$$

Now, we can state the theorem about the $\sqrt{N}$-consistency of $\hat{\beta}$.

Theorem 5 *Under the assumptions 5, 6, 7, 8, 11, 12, 13, 14:*

$$\sqrt{N} \left\| \hat{\beta} - \beta \right\| = O_p(1)$$

4 Data Based Selection of α_N

Selection of the regularization parameter is crucial. It is of great importance because it characterizes the balance between fitting and smoothing. As a result, it may have important impacts on the estimation. For example, a regularization parameter which is selected too high gives very flat curves. In contrast, a regularization parameter which is too small results in oscillating curves. Engl, Hanke, and Neubauer (1996) propose a heuristic selection rule, called *the discrepancy principle. The discrepancy principle* is based on the comparison between the residual and the assumed bound for the noise level. Moreover, it has been proven that the regularization method where α is defined via this rule is convergent and of optimal order.

In this section we define an adaptive method for the selection of the optimal regularization parameter, namely *the method of residuals. The method of residuals* is similar to *the discrepancy principle* in the sense that the aim is to minimize the error from the estimation. Indeed it is based on the minimization of some function of the squared norm of residuals. Since the squared norm of residuals can be shown to reach its minimum at $\alpha_N = 0$, we can not directly use it, hence we need a function of it. This function is obtained by first dividing the norm by α_N^2 and second by calculating the residuals from an estimation obtained by *Tikhonov Regularization of order 2.* One of the drawbacks of Tikhonov regularization (of order one) is that since its qualification is 2, when the function being estimated is very regular, i.e. $\nu > 2$, we can not improve more on the rate of convergence. So, iterated Tikhonov regularization is developed to get over this problem.[10] This second modification to the squared norm of residuals is especially done for cases where $\nu > 2$.

For the simple model in equation (3) introduced in *Section 2*, it can be shown that the functions of residuals defined above is decreasing in α. Moreover as can be seen in Equation 7, we have two regularization parameters for two unknown functions and these two parameters

[10] For more information on Iterated Tikhonov Regularization see Engl, Hanke, and Neubauer (1996), Chapter 5.

need not necessarily be the same. To get over the stated problems, we first assume that there is a constant ratio between the 2 regularization parameters, i.e. $\alpha_\varphi = c\alpha_H$ for $c \geq 0$.[11] We then propose to choose optimal values for α_H and c in two steps:

- First, we choose α_H according to method of residuals defined above where the estimation problem is given by:

$$G(Y, Z) = X + U$$

where $G(Y, Z) = H(Y) - \varphi(Z)$. Under the mean independence condition in (1) the main identifying equation can be written as:

$$T_G G(Y, Z) = X$$

where $T_G : L^2_F(Y, Z) \mapsto L^2_F(X, W) : T_G G = \mathbb{E}[G(Y, Z)|X, W]$. The adjoint T^*_G is defined as: $T^*_G : L^2_F(X, W) \mapsto L^2_F(Y, Z) : T^*_G \phi = \mathbb{E}[\phi(X, W)|Y, Z]$. Then $\hat{G}$ is given by:

$$\hat{G} = (\alpha I + T^*_G T_G)^{-1} T^*_G X$$

Let A_{yz} be the matrix with $(i, j)th$ entry:

$$A_{yz}(z)(i, j) = \frac{K_y\left(\frac{y_i - y_j}{h_y}\right) K_z\left(\frac{z - z_j}{h_z}\right)}{\sum_j K_y\left(\frac{y_i - y_j}{h_y}\right) K_z\left(\frac{z - z_j}{h_z}\right)}$$

$\hat{G}$ obtained from estimation with Tikhonov regularization of order 1 can be written as:

$$\hat{G}^\alpha_{(1)} = (\alpha I + A_{yz} A_{xw})^{-1} A_{yz} X$$

and $\hat{G}$ obtained from estimation with Tikhonov regularization of order 2 is given by:

$$\hat{G}^\alpha_{(2)} = (\alpha I + A_{yz} A_{xw})^{-1} (A_{yz} X + \alpha \hat{G}^\alpha_{(1)})$$

[11] α_φ represents the α in front of φ and α_c represents the α in front of H in equation 7.

Finally, the residuals from the estimation obtained by Tikhonov Regularization of order 2 are given by:

$$\hat{\epsilon}^{\alpha}_{(2)} = A_{yz}X - A_{yz}A_{xw}\hat{G}^{\alpha}_{(2)}$$

Then the optimal α_H is given by the minimization of the following problem:

$$\alpha^*_H = \underset{\alpha}{\operatorname{argmin}} \frac{1}{\alpha^2}\|\hat{\epsilon}^{\alpha}_{(2)}\|^2$$

- In the second step, we plug α^*_H in our original estimation problem:

$$\begin{pmatrix} \hat{H} \\ \hat{\varphi} \end{pmatrix} = \begin{pmatrix} \alpha^*_H I + A_y A_{xw} & -A_y A_{xw} \\ +PA_z A_{xw} & -(c\alpha^*_H I + PA_z A_{xw}) \end{pmatrix}^{-1} \begin{pmatrix} A_y X \\ PA_z X \end{pmatrix} \tag{13}$$

  and choose c which minimizes the squared norm of residuals obtained from an estimation regularized by Tikhonov regularization of order 2.

Another issue in the choice of regularization parameter is its sensitivity to the choice of bandwidth. The method of residuals is defined for given bandwidths. In our simulations, we choose the bandwidth by a *rule of thumb* and then optimize on regularization parameter. Feve and Florens (2010) use an iterative approach in their simulations where they first choose α for an arbitrary bandwidth and then they iterate the optimization to choose the bandwidth. They conclude that the results do not change drastically when an iterative scheme is used since α adapts itself for any a priori selection of bandwidth. The simultaneous choice of the regularization parameter and bandwidth is still an open question in the literature and is left for future work.

5 A Simulation Analysis

This section presents a Monte Carlo simulation analysis of our estimation method. We first explain the data generating process and then present our results.

We generate the data according to a simultaneous equations model where Y and Z are derived simultaneously. We simulate the following model:

$$H(Y) = \varphi(Z) + X\beta + U \tag{14}$$

$$G(Z) = \psi(Y) + W\gamma + V \tag{15}$$

$H(.)$ and $G(.)$ are chosen to be inverse of the logistic survival function, i.e., $H(t) = S^{-1}(t)$ where $S^{-1}(t) = log((1-t)/kt)$. Moreover $\varphi(.)$ and $\psi(.)$ are chosen to be:

$$\varphi(x) = Ax^a$$

$$\psi(x) = Bx^b$$

Then the simulated simultaneous transformation model is given by the following:

$$log\left(\frac{1-Y}{kY}\right) = \varphi(Z) + X'\beta + U \tag{16}$$

$$log\left(\frac{1-Z}{kZ}\right) = \psi(Y) + W'\gamma + V \tag{17}$$

We associate the parameters with the following values: $k = 0.1$, $A = 1$, $a = 2$, $B = 1$ and $b = 0.5$. As in the simpler model β and γ are normalized to 1. (X, W) are drawn from a joint normal distribution with mean $\mu = (2, 3)$ and covariance

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

U and V are drawn independently from a standard normal distribution. Given the associate values of parameters and sample of (U,V,X,W), we solve the equation system in (16) to get the values of Y and Z to use in the estimation process. We generate 100 samples of sizes 200, 500 and 1000. We perform the simulation for different sample sizes to control for the effect of sample size on the estimation.

In the estimation process, all the kernels are Gaussian and the bandwidths of the kernels are computed by a *rule of thumb*. For the regularization parameter, we use the data based selection rule defined in *Section 5* at each estimation. The simulation is performed by user written code in MATLAB.

Results are given in Figures 1 to 5 in *Appendix D*. Figure 1 shows the estimated functions over the true ones for a single sample of 200 and Figure 2 presents a Monte Carlo analysis for a sample size of 200. Our results are very satisfactory. As can be seen in Figure 1 we can get very close to the true values of our estimated functions. There are a couple of issues worth noting. Firstly, the estimated functions are very sensitive to both the regularization parameter and the bandwidth of the kernel estimator. Secondly, as already stated in the previous section, we are estimating two functions simultaneously and thus we need two different regularization parameters, which makes the estimation even harder. In the simulations, we see that $\hat{H}(.)$ is also very sensitive to the regularization parameter we use for $\hat{\varphi}(.)$ and using the same regularization parameter for both functions does not give good estimates, see Figures 3 and 4. Another point worth noticing is the change of optimal α parameter with the sample size. The simulations show that an optimal α which gives very good estimates with a sample size of 200, cannot do the same with a sample size of 1000, see Figure 5. This also supports the theory that the optimality of the regularization parameter is related to the sample size.

In addition to those, we know that we need a strong regularization, i.e., a large regularization parameter for the operators whose smallest eigenvalue is close to 0. In our case, not only the smallest eigen value was very close to zero but also the 3rd largest one. In our simulations, the optimal α for $\hat{H}$, α_H, lies between the ranges $[10^{-3}, 10^{-1}]$ while that for $\hat{\varphi}$, α_φ was in $[10^1, 10^3]$.

6 Conclusion

In this paper, we have considered the nonparametric estimation of a semiparametric transformation model. The equation we introduce is motivated by empirical study of network industries and can be applicable to many models in economics. We have studied the identification and estimation of the model and the asymptotic properties of the estimators. Furthermore, we have presented a data based selection rule for the regularization parameter for a fixed bandwidth. Development of a rule for the simultaneous selection of the two is left for future work.

The contributions of the paper are manyfold. First, it considers a transformation model where both left hand side and right hand side functions are introduced nonparametrically. Second, for the right hand side we make a partially linear specification and show that under some assumptions, we can obtain $\sqrt{N}$-convergence rate in the nonparametric estimation of the parametric part. Third, all the results of this very general model holds under the assumption of mean independence which is weaker than the full independence condition.

There may be many other possible extensions of the paper. First of all, estimation of a system of semiparametric transformation model with a full information approach is worth studying. Moreover, the estimation method and its asymptotic properties can be generalized to other nonparametric techniques different from kernels, which would be very useful when working with high dimensional variables. Nonparametric tests of specification for transformation models is still underdeveloped in the literature. Finally, estimation of a structural economic model by applying the method developed here will be an interesting application.

Appendices

A Illustration of Completeness Assumption

Assumption 2, the so called *Completeness* assumption is crucial in the identification of nonparametric IV models. In this section, we will give an illustration of primitive conditions needed in the case of a simultaneous equations system. Let us consider the model in (14) presented in the simulation and assume that we normalize the β and γ to one:

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

$$G(Z) = \psi(Y) + W + V \tag{19}$$

Let us define the following variables:

$$\zeta = H(Y) - \varphi(Z)$$

$$\eta = G(Z) - \psi(Y)$$

To show that (Y, Z) is complete for (X, W) we need the following assumptions:

Assumption 15 *The function $h : (Y, Z) \mapsto (\zeta, \eta)$ is a bijection.*

Assumption 16 *(U, V) is independent of (X, W).*

Assumption 17 *Fourier transform of joint distribution of (U,V) is strictly positive, i.e.,*

$$\int\int e^{-i(t\mu+s\nu)} f_{u,v}(\mu, \nu) d\mu d\nu \neq 0$$

Lemma 6 *Under assumptions 15-17, (Y, Z) is complete for (X, W).*

Proof. Under *Assumption 15*, if (ζ, η) is complete for (X, W), then (Y, Z) is complete for (X, W) as well. (See Florens, Mouchart, and Rolin, 1990, Chapter 5) So, it is enough to show that (ζ, η) is complete for (X, W), i.e:

$$If \quad \mathbb{E}[\phi(\zeta, \eta)|X, W] = 0 \quad a.s. \Rightarrow \phi(\zeta, \eta) = 0 \quad a.s$$

Let us write the expectation:

$$\int\int \phi(\zeta, \eta) f_{\zeta,\eta}(\zeta, \eta|X, W) d\zeta d\eta = 0$$

By *Assumption 16*:

$$\int\int \phi(\zeta, \eta) f_{U,V}(\zeta - X, \eta - W) d\zeta d\eta = 0$$

We apply Fourier Transform:

$$\int\int\int\int e^{i(tX+sW)} \phi(\zeta, \eta) f_{U,V}(\zeta - X, \eta - W) d\zeta d\eta dx dw = 0$$

Let $\mu = \zeta - X$ and $\nu = \eta - W$. Then:

$$\int\int\int\int e^{it(\zeta-\mu)} e^{is(\eta-\nu)} \phi(\zeta, \eta) f_{U,V}(\mu, \nu) d\zeta d\eta d\mu d\nu = 0$$

$$\underbrace{\int\int e^{(it\zeta+is\eta)} \phi(\zeta, \eta) d\zeta d\eta}_{\mathcal{F}_\phi(t,s)} \underbrace{\int\int e^{-i(t\mu+s\nu)} f_{U,V}(\mu, \nu) d\mu d\nu}_{\mathcal{F}_f(t,s)} = 0 \tag{20}$$

Equation 20 can be equal to zero either $\mathcal{F}_\phi(t, s)$ or $\mathcal{F}_f(t, s)$ equals zero. By *Assumption 16*, $\mathcal{F}_f(t, s)$ is different than zero, so $\mathcal{F}_\phi(t, s) = 0$. Since Fourier transform is injective, this implies that $\phi(\zeta, \eta) = 0$ and so (Y, Z) is complete for (X, W). ■

Note that *Assumption 16* is stronger than what we require in our identification theorem, however we present Lemma 6 just as an illustration of completeness assumption. Moreover, *Assumption 16* can be relaxed under a location-scale model. Proof of completeness with such a construction can be found in Hu and Shiu (2011).

B Verification of Inner Product Space

Definition 2 *Let H be a complex vector space. A mapping $\langle , \rangle : H \times H \mapsto \mathbb{C}$ is called an inner product in H if for any $\phi, \psi, \xi \in H$ and $\alpha, \beta \in \mathbb{C}$ the following conditions are satisfied:*

- $\langle \phi, \psi \rangle = \overline{\langle \psi, \phi \rangle}$
- $\langle \alpha\phi + \beta\psi, \xi \rangle = \alpha \langle \phi, \xi \rangle + \beta \langle \psi, \xi \rangle$
- $\langle \phi, \phi \rangle \geq 0$ *and* $\langle \phi, \phi \rangle = 0 \Leftrightarrow \phi = 0$

Remember that our inner product space is defined by

$$\langle (H_1, \varphi_1), (H_2, \varphi_2) \rangle = \langle H_1, H_2 \rangle + \langle \varphi_1, \varphi_2 \rangle$$

Let us start with the first condition, conjugate symmetry:

$$\langle (H_1, \varphi_1), (H_2, \varphi_2) \rangle = \langle H_1, H_2 \rangle + \langle \varphi_1, \varphi_2 \rangle$$

$$\left\langle \overline{(H_2, \varphi_2), (H_1, \varphi_1)} \right\rangle = \left\langle \overline{H_2, H_1} \right\rangle + \langle \overline{\varphi_2, \varphi_1} \rangle$$

Then we can write:

$$\langle H_1, H_2 \rangle + \langle \varphi_1, \varphi_2 \rangle = \left\langle \overline{H_2, H_1} \right\rangle + \langle \overline{\varphi_2, \varphi_1} \rangle$$

Second condition is the condition of linearity:

$$\begin{aligned} \langle \alpha(H_1, \varphi_1) + \beta(H_3, \varphi_3), (H_2, \varphi_2) \rangle &= \alpha \langle (H_1, \varphi_1), (H_2, \varphi_2) \rangle + \beta \langle (H_3, \varphi_3), (H_2, \varphi_2) \rangle \\ &= \alpha \left(\langle H_1, H_2 \rangle + \langle \varphi_1, \varphi_2 \rangle \right) + \beta \left(\langle H_3, H_2 \rangle + \langle \varphi_3, \varphi_2 \rangle \right) \end{aligned}$$

Finally the last condition is the condition of positive definiteness, which is verified by the positive definiteness of each term.

$$\langle (H_1, \varphi_1), (H_1, \varphi_1) \rangle = \underbrace{\langle H_1, H_1 \rangle}_{\geq 0} + \underbrace{\langle \varphi_1, \varphi_1 \rangle}_{\geq 0}$$

so:

$$\langle (H_1, \varphi_1), (H_1, \varphi_1) \rangle = \langle H_1, H_1 \rangle + \langle \varphi_1, \varphi_1 \rangle \geq 0$$

and it is equal to zero if each term is equal to zero separately, which can be the case only if $H_1 = 0$ and $\varphi_1 = 0$.

C Proofs of Theorems

C.1 Theorem 1

Proof. By Assumption 1

$$\mathbb{E}[H(Y) - \varphi(Z) - X|X, W] = 0$$

Let us recall two more functions $H^*(Y)$ and $\varphi^*(Z)$. By Assumption 1 again, we can write:

$$\mathbb{E}[H(Y) - \varphi(Z) - X|X, W] = 0 \quad \mathbb{E}[H^*(Y) - \varphi^*(Z) - X|X, W] = 0$$

If we take the difference of the two expectations:

$$\mathbb{E}[(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) + (X - X)|X, W] = 0$$

then by Assumption 2:

$$(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) = 0$$

by Assumption 3

$$(H(Y) - H^*(Y)) = (\varphi(Z) - \varphi^*(Z)) = c$$

finally by Assumption 4:

$$c = 0$$

then:

$$H(Y) = H^*(Y) \quad and \quad \varphi(Z) = \varphi^*(Z)$$

■

C.2 Lemma 2

Proof. Note that, we can write:

$$x \in \mathcal{E}_0, \quad \langle T_0 x, y \rangle = \langle T x, y \rangle$$

$$\langle T_0 x, y \rangle = \langle x, T_0^* y \rangle$$

$$\langle T x, y \rangle = \langle x, T^* y \rangle$$

Moreover, $x \in \mathcal{E}_0$ and $z \in \mathcal{E}$

$$\langle x, z \rangle = \langle x, \mathbb{P} z \rangle$$

then

$$\langle x, T_0^* y \rangle = \langle x, \mathbb{P} T_0^* y \rangle = \langle x, T^* y \rangle$$

then

$$\mathbb{P} T_0^* = T^*$$

■

C.3 Theorem 3

Proof. Remember that the solution of our problem was given by

$$\Phi = (\alpha I + T^*T)^{-1}T^*X$$

For the proof, we will decompose our equation into three parts as was done in Darolles, Fan, Florens, and Renault (2011) and look at the rates of convergence term by term.

$$\hat{\Phi}^\alpha - \Phi = \underbrace{(\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*X - (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\Phi}_{I}$$
$$+ \underbrace{(\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\Phi - (\alpha I + T^*T)^{-1}T^*T\Phi}_{II}$$
$$+ \underbrace{(\alpha I + T^*T)^{-1}T^*T\Phi - \Phi}_{III}$$

The first term (I) is the estimation error about the right hand side (X) of the equation, the second term (II) is the estimation error coming from the kernels and the third term (III) is the regularization bias coming from regularization parameter α.

Now, let's first examine the first term:

$$I = (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*X - (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\Phi$$

$$I = (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*(X - \hat{T}\Phi)$$

$$\|I\|^2 = \left\|(\alpha I + \hat{T}^*\hat{T})^{-1}\right\|^2 \left\|\hat{T}^*X - \hat{T}^*\hat{T}\Phi\right\|^2$$

where the first term is $O\left(1/\alpha^2\right)$ by Darolles, Fan, Florens, and Renault (2011) and the second term is $O\left(N^{-1} + h_N^{2s}\right)$ by Assumption 7.

Now, let us look at the second term II:

$$II = (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\Phi - (\alpha I + T^*T)^{-1}T^*T\Phi$$

$$= \left[\left[I - (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\right] - \left[I - (\alpha I + T^*T)^{-1}T^*T\right]\right]\Phi$$

$$= \left[\alpha(\alpha I + \hat{T}^*\hat{T})^{-1} - \alpha(\alpha I + T^*T)^{-1}\right]\Phi$$

$$= (\alpha I + \hat{T}^*\hat{T})^{-1}(\hat{T}^*\hat{T} - T^*T)\alpha(\alpha I + T^*T)^{-1}\Phi$$

$$\|II\|^2 = \left\|(\alpha I + \hat{T}^*\hat{T})^{-1}\right\|^2 \left\|(\hat{T}^*\hat{T} - T^*T)\right\|^2 \left\|\alpha(\alpha I + T^*T)^{-1}\Phi\right\|^2$$

The first term in (II) is $O(1/\alpha^2)$ by Darolles, Fan, Florens, and Renault (2011) while the second one is of order $O\left((Nh_N^{p+2})^{-1} + h_N^{2s}\right)$ as a result of relation $\left\|\hat{T}^*\hat{T} - T^*T\right\| = O\left(max\left\|\hat{T} - T\right\|, \left\|\hat{T}^* - T^*\right\|\right)$ by Assumption 6 and by Florens, Johannes, and Van Bellegem (2009). Finally, the third is equal to $O(\alpha^{(\nu\wedge 2)})$ by Darolles, Fan, Florens, and Renault (2011).

The third term can be examined more straightforwardly:

$$III = (\alpha I + T^*T)^{-1}T^*T\Phi - \Phi$$

$$= \Phi^\alpha - \Phi$$

and $\|III\|^2 = \|\Phi^\alpha - \Phi\|^2$ is $O(\alpha^{\nu\wedge 2})$ by Assumption 5. Finally if we combine all what we have:

$$\left\|\hat{\Phi}_N^\alpha - \Phi\right\|^2 = O\left(\frac{1}{\alpha^2}\left(\frac{1}{N} + h_N^{2s}\right) + \frac{1}{\alpha^2}\left(\frac{1}{Nh_N^{p+2}} + h_N^{2s}\right)\left(\alpha^{(\nu\wedge 2)}\right) + \alpha^{(\nu\wedge 2)}\right)$$

The proof of the second part of the theorem follows by Assumption 8. ■

C.4 Theorem 4

Proof.

$$H(Y) - \varphi(Z) - X_0 - X_1'\beta = U$$

$$\mathbb{E}[H(Y) - \varphi(Z) - X_0 - X_1'\beta | X, W] = 0 \quad by \quad Assumption \quad 1$$

Let us recall two more functions $H^*(Y)$, $\varphi^*(Z)$ and β^* such that:

$$H^*(Y) - \varphi^*(Z) - X_0 - X_1'\beta^* = U$$

Then, again by Assumption 1:

$$\mathbb{E}[H^*(Y) - \varphi^*(Z) - X_0 - X_1'\beta^*|X, W] = 0$$

If we take the difference of the two expectations:

$$\mathbb{E}[(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) - (X_1'\beta - X_1'\beta^*)|X, W] = 0$$

Then, by Assumption 9:

$$(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) - (X_1'\beta - X_1'\beta^*) = 0$$

By Assumption 10:

$$(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) = (X_1'\beta - X_1'\beta^*) = constant$$

Finally by Assumptions 3 and 4 we get the identification:

$$H(Y) = H^*(Y) \quad \varphi(Z) = \varphi^*(Z) \quad X_1'\beta = X_1'\beta^*$$

■

C.5 Theorem 5

Proof.

$$\hat{\beta}-\beta=\underbrace{\hat{M}_{\alpha}^{-1}}_{I}\left\{\underbrace{[\hat{T}_X^{*}-\widehat{T_X^{*}T}(\alpha I+\hat{T}^{*}\hat{T})^{-1}\hat{T}^{*}](X_0-\hat{T}(H,\varphi)+T_X\beta)}_{II}+\underbrace{[\hat{T}_X^{*}-\widehat{T_X^{*}T}(\alpha I+\hat{T}^{*}\hat{T})^{-1}\hat{T}^{*}]\hat{T}(H,\varphi)}_{III}\right\}$$

where $\hat{M}^{\alpha}=\widehat{T_X^{*}T}(\alpha I+\hat{T}^{*}\hat{T})^{-1}\widehat{T^{*}T_X}-\widehat{T_X^{*}T_X}$.

To prove our final result, we will show the following asymptotic convergences:

$$I=\left\|\hat{M}_{\alpha}^{-1}-M_{\alpha}^{-1}\right\|=O_p\left(\alpha^{\frac{\eta\wedge 2}{2}}\left(\frac{1}{\sqrt{N}}+\frac{1}{\sqrt{Nh^{p+q+1}}}+h^s+\frac{1}{Nh}+h^s\right)+\frac{1}{\sqrt{N}}\right)$$

$$II=O_p(A+B)$$

where

$$A=\left[\frac{1}{\sqrt{N}}\left(1+\frac{1}{\sqrt{\alpha}}\right)\left(\frac{1}{\sqrt{Nh^{p+q+1}}}+h^s\right)\right]\frac{1}{\sqrt{\alpha}}\left(\frac{1}{Nh^{p+q+1}}+h^s\right)$$

$$B=\alpha^{\frac{\eta\wedge 2}{2}}\left(\left(\frac{1}{Nh^{p+q+1}}+h^s\right)+\left(\frac{1}{\sqrt{Nh^{p+q+1}}}+h^s\right)\right)$$

$$III=O_p\left(\frac{1}{\sqrt{N}}+\sqrt{\alpha}\left(\frac{1}{\sqrt{Nh^{p+q+1}}}+h^s\right)^{\frac{\nu\wedge 2}{2}}+\alpha^{\frac{1\wedge(1+\nu)}{2}}\right)$$

The assumptions we made to state *Theorem 5* ensure that I has the rate $O_p(1)$ while II and III have the rate $O_p(N^{-1/2})$. Let us begin with I.

Proof of I:

$$\left\|\hat{M}_{\alpha}^{-1}-M_{\alpha}^{-1}\right\|\leq\left\|M_{\alpha}^{-1}\right\|\left\|\hat{M}_{\alpha}^{-1}\right\|\left\|\hat{M}_{\alpha}-M_{\alpha}\right\|$$

As M_{α} and $\hat{M}_{\alpha}$ are finite rank operators, their inverses are bounded. So we need to look at the convergence of the third term.

$$\left\|\hat{M}_{\alpha}-M_{\alpha}\right\|=\left\|[\widehat{T_X^{*}T}(\alpha I+\hat{T}^{*}\hat{T})^{-1}\widehat{T^{*}T_X}-\widehat{T_X^{*}T_X}]-[T_X^{*}T(\alpha I+T^{*}T)^{-1}T^{*}T_X-T_X^{*}T_X]\right\|$$

$$\leq \left\| \widehat{T_X^* T}(\alpha I + \hat{T}^* \hat{T})^{-1} \widehat{T^* T_X} - T_X^* T(\alpha I + T^* T)^{-1} T^* T_X \right\|$$

$$+ \left\| T_X^* T_X - \widehat{T_X^* T_X} \right\|$$

$$\leq \underbrace{\left\| \widehat{T_X^* T} - T_X^* T \right\| \left\| (\alpha I + T^* T)^{-1} T^* T_X \right\|}_{A}$$

$$+ \underbrace{\left\| T_X^* T[(\alpha I + \hat{T}^* \hat{T})^{-1} - (\alpha I + T^* T)^{-1}] T^* T_X \right\|}_{B}$$

$$+ \underbrace{\left\| T_X^* T(\alpha I + T^* T)^{-1} \right\| \left\| \widehat{T^* T_X} - T^* T_X \right\|}_{C} + \underbrace{\left\| T_X^* T_X - \widehat{T_X^* T_X} \right\|}_{D}$$

- The first term in A is of order $O(1/\sqrt{N})$ by Assumption 13 and the second term is $O(\alpha^{(\eta \wedge 2)/2})$ by Florens, Johannes, and Van Bellegem (2009)

- B can be decomposed as the following:

$$\|B\| \leq \left\| T_X^* T(\alpha I + \hat{T}^* \hat{T})^{-1} \right\| \left\| T^* T - \hat{T}^* \hat{T} \right\| \left\| (\alpha I + T^* T)^{-1} T^* T_X \right\|$$

  The first term is bounded. The second term is of order $O((Nh^{p+q+1})^{-1/2} + h^s)$ by Assumption 6 and the third term is of order $O(\alpha^{(\nu \wedge 2)/2})$ as before.

- The first of C is $O(\alpha^{(\nu \wedge 2)/2})$ and the second term is $O\left((Nh)^{-1/2} + h^s\right)$ by Assumption 12.

- Finally D is of order $O\left(N^{-1/2}\right)$ by Assumption 13.

Proof of II:

We can denote $\hat{e} = X_0 - \hat{T}(H, \varphi) + T_X \beta$ as $T(H, \varphi) = T_x \beta + X_0$. Then:

$$\|\hat{e}\| \leq \left\| \hat{T} - T \right\|$$

which is of order $O((Nh^{p+q+1})^{-1/2}+h^{2s})$ by Assumption 6. Then we can write II as:

$$[\hat{T_X^*}-\widehat{T_X^*T}(\alpha I+\hat{T}^*\hat{T})^{-1}\hat{T}^*]\hat{e}$$

$$=\left\{(\hat{T_X^*}-\widehat{T_X^*T}(\alpha I+\hat{T}^*\hat{T})^{-1}\hat{T}^*)-(T_X^*-T_X^*T(\alpha I+T^*T)^{-1}T^*)\right\}\hat{e}$$

$$+(T_X^*-T_X^*T(\alpha I+T^*T)^{-1}T^*)\hat{e}$$

The first part is of order

$$O_p\left(\left[\frac{1}{\sqrt{N}}\left(1+\frac{1}{\sqrt{\alpha}}\right)\left(\frac{1}{\sqrt{Nh^{p+q+1}}}+h^s\right)\right]+\frac{1}{\sqrt{\alpha}}\left(\frac{1}{Nh^{p+q+1}}+h^s\right)+\alpha^{\frac{\eta\wedge 2}{2}}\left(\frac{1}{Nh^{p+q+1}}+h^s\right)\right)$$

and the second part is of order $O(\alpha^{(\nu\wedge 2)/2}((Nh^{p+q+1})^{-1/2}+h^s))$.

Proof of III:

By Assumption 5, we can write:

$$\left\|[\hat{T_X^*}-\widehat{T_X^*T}(\alpha I+\hat{T}^*\hat{T})^{-1}\hat{T}^*]\hat{T}(H,\varphi)\right\|\leq\left\|\widehat{T_X^*T}\right\|(H,\varphi)$$

$$+\left\|\widehat{T_X^*T}(\alpha I+\hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\right\|\left\|(T^*T)^{\nu/2}-(\hat{T}^*\hat{T})^{\nu/2}\right\|\|g\|$$

$$+\left\|\widehat{T_X^*T}(\alpha I+\hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}(\hat{T}^*\hat{T})^{\nu/2}\right\|\|g\|$$

The first term is $O(N^{-1/2})$, the second is of order $O(\alpha^{1/2})$. Moreover by Engl, Hanke, and Neubauer (1996) $\left\|(T^*T)^{\nu/2}-(\hat{T}^*\hat{T})^{\nu/2}\right\|\leq\left\|T^*T-\hat{T}^*\hat{T}\right\|^{(\nu\wedge 2)/2}$. The rate of the last part is also given by Engl, Hanke, and Neubauer (1996) and is equal to $O(\alpha^{1\wedge(1+\nu)/2})$.

■

D Simulation Results

Below, we present the simulation results. Figures contain both the $\hat{H}_1$ and $\hat{\varphi}$ for different values α_H and α_φ

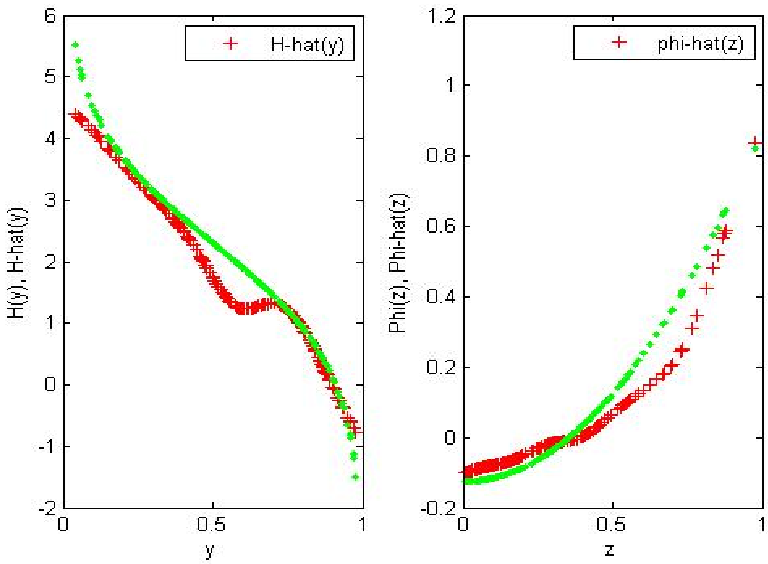

Figure 1: *Estimated functions for a sample of* 200. Pluses are the estimated value of the function at each point whereas the diamonds are the true value of the function at each point.α_H and α_φ are chosen by the data driven rule given in *Section 4*. $\alpha_H = 0.0307$ and $\alpha_\varphi = 0.307$

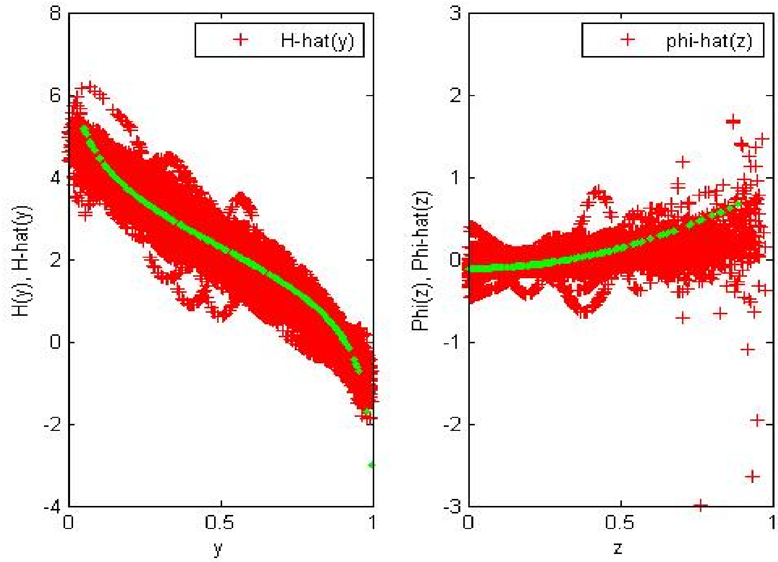

Figure 2: *Monte Carlo simulation.* Diamonds show the true values of the function at each point. Pluses are the estimated values of the function at each point and at each simulation. α_H and α_φ are chosen by the data driven rule at each simulation.

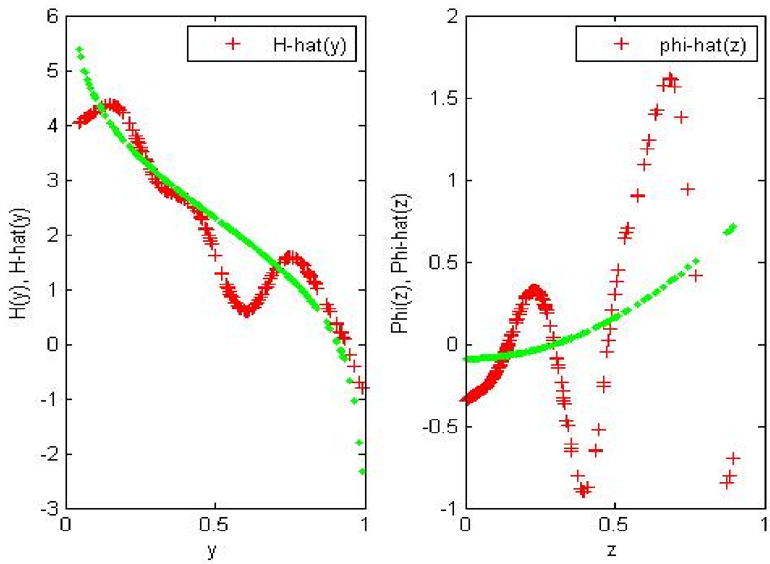

Figure 3: *Estimated functions for* $\alpha_H = \alpha_\varphi = 0.0106$. Pluses are the estimated value of the function at each point whereas the diamonds are the true value of the function at each point.

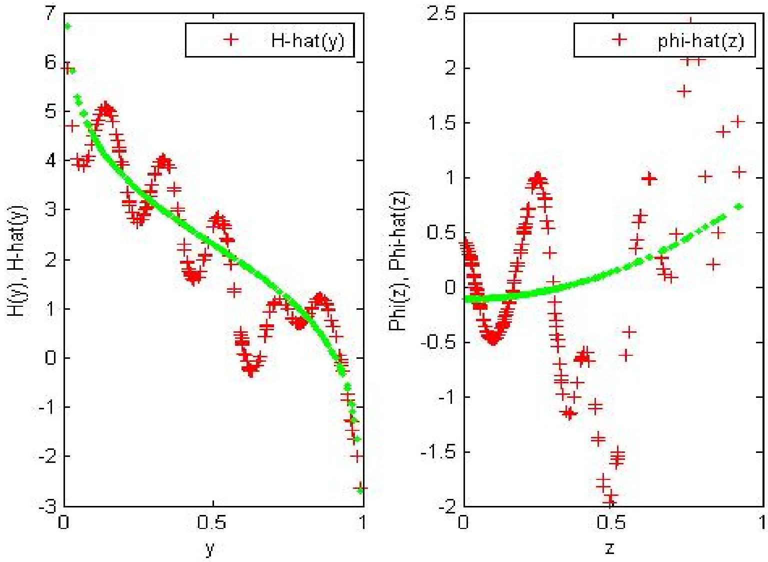

Figure 4: *Estimated functions for* $\alpha_H = \alpha_\varphi = 10^{-5}$. Pluses are the estimated value of the function at each point whereas the diamonds are the true value of the function at each point.

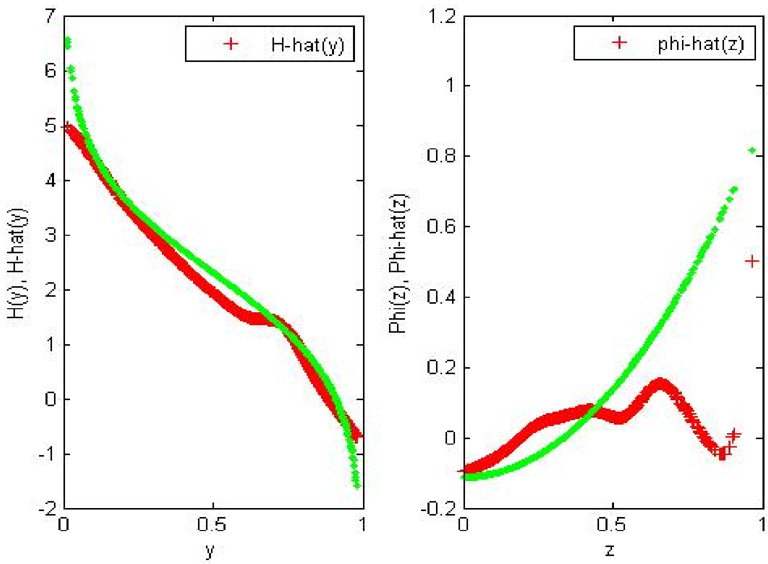

Figure 5: *Estimated functions with a sample size of* 1000. α_H and α_φ are the data driven optimal values for a sample size of 200. Pluses are the estimated value of the function at each point whereas the diamonds are the true value of the function at each point.

References

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

Andrews, D. W. (2011): "Examples of L^2-Complete and Boundedly-Complete Distributions," Cowles Foundation Discussion Papers 1801, Cowles Foundation for Research in Economics, Yale University.

Bass, F. M. (1969): "A New Product Growth for Model Consumer Durables," *Management Science*, 15, 215–227.

Bontemps, C., M. Simioni, and Y. Surry (2008): "Semiparametric hedonic price models: assessing the effects of agricultural nonpoint source pollution," *Journal of Applied Econometrics*, 23(6), 825–842.

Box, G. E. P., and D. R. Cox (1964): "An Analysis of Transformations," *Journal of the Royal Statistical Society*, 26(2), 211–252.

Breunig, C., and J. Johannes (2009): "On rate optimal local estimation in nonparametric instrumental regression," Discussion paper, Heidelberg University.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 77. Elsevier.

Chen, X., and M. Reiss (2007): "On rate optimality for ill-posed inverse problems in econometrics," CeMMAP working papers CWP20/07, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Chiappori, P.-A., I. Komunjer, and D. Kristensen (2011): "Nonparametric Identification and Estimation of Transformation Models," CAM Working Papers 2011-01, University of Copenhagen. Department of Economics. Centre for Applied Microeconometrics.

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

DHaultfoeuille, X. (2011): "On The Completeness Condition In Nonparametric Instrumental Problems," *Econometric Theory*, 27(03), 460–471.

Engl, H. W., M. Hanke, and A. Neubauer (1996): *Regularization of Inverse Problems*. Kluwer Academic Publications, Dordrecht.

Engle, R. F., C. W. J. Granger, J. Rice, and A. Weiss (1986): "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81(394), 310–20.

Feve, F., and J.-P. Florens (2010): "The Practice of Non Parametric Estimation by Solving Inverse Problems: The Example of Transformation Models," *Econometrics Journal*.

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., J. Johannes, and S. Van Bellegem (2009): "Instrumental Regression in Partially Linear Models," TSE Working Papers 10-167, Toulouse School of Economics (TSE).

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

Hall, P., and J. L. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 32, 2904–2929.

Horowitz, J. L. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103–137.

——— (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79(2), 347–397.

Hu, Y., and J.-L. Shiu (2011): "Nonparametric identification using instrumental variables: sufficient conditions for completeness," CeMMAP working papers CWP25/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 74. Elsevier.

Kaiser, U., and M. Song (2009): "Do media consumers really dislike advertising? An empirical assessment of the role of advertising in print media markets," *International Journal of Industrial Organization*, 27(2), 292–301.

Linton, O., S. Sperlich, and I. V. Keilegom (2008): "Estimation of a Semiparametric Transformation Model," *Annals of Statistics*, 36(2), 686–718.

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