Università degli Studi di Milano Bicocca
Dipartimento di Economia, Metodi Quantitativi e Strategie d'Impresa

Dottorato di Ricerca in Economia Politica
XXV Ciclo
(a.a. 2009-2010 - a.a. 2013-2014)

ESSAYS IN NONPARAMETRIC ESTIMATION WITH INSTRUMENTAL VARIABLES

Settore scientifico-disciplinare di afferenza: SECS-P/05

Tesi di Dottorato di
Samuele Centorrino
Matricola 727350

Relatore della tesi
Prof.ssa *Alessia Paccagnini*

Coordinatore del Dottorato
Prof.ssa *Giovanna Iannantuoni*

"Thoughts without contents are empty.
Opinions without concepts are blind."
Immanuel Kant

To my parents, Angela e Nando

Acknowledgments

This thesis would never have been achieved without the constant guidance, patience and support of Jean-Pierre Florens. I would like to thank Stéphane Bonhomme, Frédérique Fève, Frank Kleibergen, Pascal Lavergne, Adam McCloskey, Nour Meddahi, Blaise Melly, Jeffrey S. Racine, and Eric Renault for fruitful discussions and valuable comments.

My thanks go to Alessia Paccagnini for her supervision, and Giovanna Iannantuoni for being a very helpful and patient director of Graduate Studies.

I would also like to express my gratitude to Patrick Gagliardini, Peter Robinson, and Anna Simoni, for having accepted to be part of my thesis committee.

A huge thank goes to Maria, and to my family for their immense and unconditional support during this journey.

Financial support from the University of Milan Bicocca is gratefully acknowledged.

Abstract

This thesis deals with the nonparametric estimation of the regression function in additive separable models with endogeneity.

Endogeneity is, in this case, broadly defined. It can relate to reverse causality (the dependent variable can also affects the independent regressor) or to simultaneity (the error term contains information that can be related to the explanatory variable). Identification and estimation of the regression function is performed using the method of instrumental variables.

In this setting, the function of interest is known to be the solution of an inverse problem which is *ill-posed* and, therefore, it needs to be recovered using regularization techniques.

In the first chapter, this estimation problem is considered when the regularization is achieved using a penalization on the $\mathbb{L}^2-$norm of the function of interest (so-called Tikhonov regularization). We derive the properties of a leave-one-out cross validation criterion in order to choose the regularization parameter.

In the second chapter, coauthored with Jean-Pierre Florens, we extend this model to the case in which the dependent variable is not directly observed, but only a binary transformation of it. We show that identification can be obtained via the decomposition of the dependent variable on the space spanned by the instruments, when the residuals in this reduced form model are taken to have a known distribution. We finally show that, under these assumptions, the consistency properties of the estimator are preserved.

Finally, chapter three, coauthored with Frédérique Fève and Jean-Pierre Florens, performs a numerical study, in which the properties of several regularization techniques are investigated. In particular, we gather data-driven techniques for the sequential choice of the smoothing and the regularization parameters and we assess the validity of wild bootstrap in nonparametric instrumental regressions.

Contents

List of Figures

List of Tables

Introduction

The assessment of causality in economic phenomena is one of the crucial, albeit among the most challenging, tasks of a researcher.

Since Economics is the science of choices and decisions, it is essential to uncover the determinants of these decisions and their causes. The difficulty of this task steams from the fact that the same effect can have different causes. The job of the economist is, therefore, to provide a meaningful theory that can reasonably exclude the irrelevant ones.

The job of the econometrician is slightly different and, perhaps, somehow a little easier. Given an effect and a cause, we often ask ourselves what are the meaningful assumptions to be made in order to retrieve the structural relation between the two.

Sometimes an economic model is straightforward about the relation between two phenomena, especially when the cause involves natural facts that cannot be affected by economic decisions. However, in many interesting cases, it is impossible to distinguish the effect from its cause. A famous example is the one about the estimation of demand functions: a change in price affects the quantity demanded, although a shift in the quantity supplied also impacts the final price. Therefore, a very simple economic model, leaves the econometrician with a puzzling *egg-chicken* problem, and the feeling that something shall be done about it in order to effectively assess the impact on price changes on the quantity demanded.

There is a vast debate on the definition of exogeneity (in all its different nuances) and causality in econometrics (see, for instance Florens and Heckman, 2003; Klein, 1990; Pearl, 2000).

This thesis does not contribute directly to this debate, as its author yet lacks of enough experience to enter it. By contrast, it tries to give a set of conditions and tools, in a particular class of models, under which a researcher can carry nonparametric estimation, when assumptions about exogeneity are made. Nonparametric estimation is considered here because of its flexibility and the fact that many structural economic relations should be uncovered, at least in a first step, using

the information coming from the data and not from some arbitrary parametric model.

When exogeneity breaks down, the assessment of causality requires to separate the common causes underlying two phenomena from the true causal relation. These common causes are often unobserved by the econometrician, and therefore left into the error term. Thus, endogeneity is defined, in econometrics, as the failure of some type of independence condition between the cause and the unobserved component. In this particular case, the structural relation between the cause and the effect cannot be properly captured, as it is contaminated by the residuals.

Instrumental variables are standard tools to achieve identification and carry on estimation in econonometric models with endogeneity. The underlying concept behind instrumental variables is to remove the common causes from the econometric model in a way that the researcher is able to extract the, hopefully, true relation between the cause and the effect.

In the standard iid setting, when an additive separable specification is considered and when the researcher wants to estimate the structural relation nonparametrically, the function of interest is known to be solution of an *ill-posed* inverse problem (see, for instance Darolles et al., 2011*a*; Horowitz, 2011, and references therein). The illposedness arises from the fact that the mapping defining the function has a noncontinuous inverse and, therefore, the solution cannot be found unless this inverse mapping is transformed into a continuous one. This *regularization* of the mapping can be done in several ways and many of them are considered in this thesis. However, regularization boils down to the choice of a single constant parameter which slightly modifies the mapping. In practice, in the context of nonparametric estimation of instrumental variable regressions, we lack of data-driven methods to select this parameter, and applied researchers lack of guidance to apply them.

The contribution of the this thesis to this literature is thus threefold:

(i) It provides an optimal data-driven criterion for the selection of this regularization parameter under a very specific regularization scheme (so-called Tikhonov regularization).

(ii) Extend the framework of nonparametric instrumental regressions to the case in which the dependent variable is not directly observed, but only a binary transformation of it.

(iii) It provides a detailed explanation and gives practical tools to implement these regularization

methods, when the researcher wants to use nonparametric estimation with instrumental variables.

The exposition of the results of this research has privileged a consequent unfolding.

Chapter 1 discusses the results about the data-driven selection of the regularization parameter in nonpametric instrumental regressions and it proves its optimality. Chapter 2, coauthored with Jean-Pierre Florens, extends the nonparametric instrumental variable framework to binary response models. Finally, Chapter 3, coauthored with Frédérique Fève and Jean-Pierre Florens, presents the investigation about the small sample properties of various regularization schemes and show the validity of wild bootstrap. Although Chapter 1 has been the last one to be started, as it was inspired by some empirical observation when working on chapters 2 and 3, its results are used in the latter part of this work and are therefore presented first.

Chapter 1

On the Choice of the Regularization Parameter in Nonparametric Instrumental Regressions

Abstract

This chapter studies the implementation of the instrumental variable approach in nonparametric regression models with endogenous variables and presents the properties of a data driven criterion for the selection of the regularization parameter. This choice is deemed crucial and represents a major challenge in the application of nonparametric instrumental regressions. We propose a leave-one-out cross-validation criterion which leads to a rate optimal choice of the regularization constant, up to some regularity conditions on the regression function. The main results of this chapter are derived for the case where the ultimate object of interest is the regression function itself, and they are further broadened to the estimation of its derivatives of any order. In economics this extension is extremely relevant, as it provides a methodology to obtain a direct estimation of marginal effects. Extensive numerical simulations show that our cross-validation criterion outperforms available methodologies for different penalization schemes and smoothness properties of the function of interest. Using the 1995 wave of the U.K. Family Expenditure Survey, an illustration is presented about the estimation of the Engel curve for several goods. This application emphasizes the properties, the flexibility and the simplicity of cross-validation in this framework, irrespective of the nonparametric approach chosen to estimate the conditional mean functions.

1.1 Introduction

Econometricians and economists are often interested in causal relations between variables. These causal relations are usually modeled as functional dependencies. The response (or endogenous, dependent) variable is usually written as an unknown function of the predictors (or regressors, or exogenous, independent variables) and of an unobservable random error term, which, according to the setting under study, is supposed to satisfy some independence conditions with respect to the predictors. These independence conditions enable one to write down the unknown function as a (conditional) moment of the response, and, ultimately, they allow the researcher to make inference on it.

However, in certain cases, these conditions may fail to hold. The error term may, for instance, contain unobservable regressors that are likely to be correlated with the observed independent vari-

ables; or the causality structure between the response and the predictors is reversed - the dependent variable is somehow affecting the regressors. In econometrics, this problem is usually referred as endogeneity of the predictors - the dependent and the independent variables are simultaneously determined by the unobservables. This endogeneity issue does not allow one to write down the unknown function as a moment of the response variable, and it therefore requires to be properly taken into account for correct identification and inference.

Define the response variable Y, the predictors X and a random error U. This chapter deals with the following additively separable model:

$$Y = \varphi(X) + U, \tag{1.1.1}$$

with φ being a smooth function, when the standard mean independence condition fails to hold. That is,

$$\mathbb{E}\left(U|X = x\right) \neq 0.$$

We therefore consider identification and nonparametric estimation of the regression function, φ, using the method of instrumental variables.

In the nonparametric instrumental variable setting, φ is a solution of an *ill-posed* inverse problem. Hence, its estimation requires the implementation of a regularization scheme and the consequent choice of a regularization constant. The latter selection is crucial and it represents one of the main challenges for the fully nonparametric estimation of φ in (3.2.1a) when X is endogenous.

This chapter contributes to this literature by presenting the properties and the application of a very simple *leave-one-out* cross-validation criterion for the selection of the regularization parameter. While cross-validation has been extensively used in related frameworks, this chapter is the first one to discuss its properties in the setting of nonparametric instrumental regressions.

Another important theoretical contribution of the chapter is to extend the properties of the cross validated regularization parameter to the case where the main object of interest is not the regression function itself, but its derivatives of any order. This is very relevant for economic applications, as it allows one to obtain a direct estimation of marginal effects.

Instrumental variables are a standard approach in econometrics to identify and estimate func-

tional dependency in the presence of endogenous regressors. The underlying rationale for using instrumental variables is that, in order to uncover the true functional relation, we need a source of exogenous variation that should be informative about the phenomenon under study. Our *instruments*, defined here as W, must therefore retain some correlation with the endogenous predictors and satisfy the exogeneity condition with respect to the random component. In the separable model (3.2.1a), one has:

$$\mathbb{E}(U|W = w) = 0;$$

i.e., the error term in (3.2.1a) has mean 0 on the space spanned by W (see, e.g. Newey and Powell, 2003; Hall and Horowitz, 2005; Carrasco et al., 2007; Darolles et al., 2011*a*; Horowitz, 2011; Chen and Pouzo, 2012*a*, among others).

This assumption allows one to eliminate the noise term in (3.2.1a), by taking the expectation with respect to W. Thus, our object of interest, the function φ, is now implicitly defined by the equation:

$$\mathbb{E}(\varphi(X)|W) = r, \tag{1.1.2}$$

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

Nonetheless, estimation may represent an important additional layer of difficulty when considering models with instrumental variables. A parametric specification of the function of interest φ could be easily handled, for instance, with classical two stage least squares (TSLS) regressions. However, the latter specification imposes several restrictions on the shape of φ, which may or may not be justified by the economic theory.[1] Therefore, a parametric specification might not be appropriate for some empirical applications. More generally, the researcher would like to maintain some flexibility in the specification of the function φ. Consequently, in this chapter, we focus on the fully nonparametric estimation of the regression function (Hall and Horowitz, 2005; Darolles et al., 2011*a*).

As a specific example of application of the framework of nonparametric instrumental regressions, consider the estimation of the shape of the Engel curve for a given commodity (or group of commodities; see, e.g., Blundell et al., 2007; Horowitz, 2011). The Engel curve describes the expansion path for commodity demands as the household's budget increases. Therefore, to estimate its shape,

[1] See, for instance, Horowitz (2011) for an insightful discussion about the trade-off between parametric and nonparametric specifications.

it would be sufficient to regress the share of the household's budget spent for this given commodity, the response variable Y, over the total household's budget, the predictor X. However, the latter is likely to be jointly determined with individual demands, and therefore one ought to consider it as an endogenous regressor in the estimation of consumer expansion paths. Therefore, empirical studies which aim at obtaining meaningful results about the *structural* shape of the Engel curve shall take this endogeneity problem into account for identification and inference.

As discussed in Blundell et al. (2007), the allocation model of income to individual consumption goods and savings suggests exogenous sources of income provide a suitable instrumental variable for total expenditure, as they are likely to be related to the total household expenditure and not to be jointly determined with individual's budget shares. Consequently, they provide a source of exogenous variations that allows one to identify and estimate the shape of the Engel curve by using gross income as an instrument for total expenditure. However, nonlinearities in the total expenditure variable may be required to capture the observed microeconomic behavior in the estimation of the Engel curve (see also Hausman et al., 1991; Lewbel, 1991; Banks et al., 1997), so that a nonparametric specification of the latter seems appropriate.

In the framework of instrumental variables, flexibility comes at the cost of a more cumbersome estimation methodology. While it is straightforward to obtain a nonparametric estimator of r, the right hand side of equation (2.2.2), direct estimation of φ is not feasible as it requires one to disentangle φ from its conditional expectation with respect to W. Namely, equation (2.2.2) can be rewritten as:

$$\int \varphi(x) f(x|w) dx = r \tag{1.1.3}$$

which defines a Fredholm integral equation of the first kind (Kress, 1999), where $f(x|w)$ is the conditional distribution of X given W. The main issue in the estimation of this equation is that its solution may not exist or may not be a continuous function of r. In this sense, φ is a solution of a problem that is *ill-posed*.[2]

A *naïve* way to look at the *ill-posedness* of the inverse problem is to imagine the integral operator in equation (1.1.3) as an infinite dimensional matrix. This matrix is one-to-one and therefore in-

[2] In 1923, Hadamard postulated three requirements for problems in mathematical physics: a solution should exist, the solution should be unique, and the solution should depend continuously on the data. A problem satisfying all three requirements is called *well-posed*. Otherwise, it is called *ill-posed*.

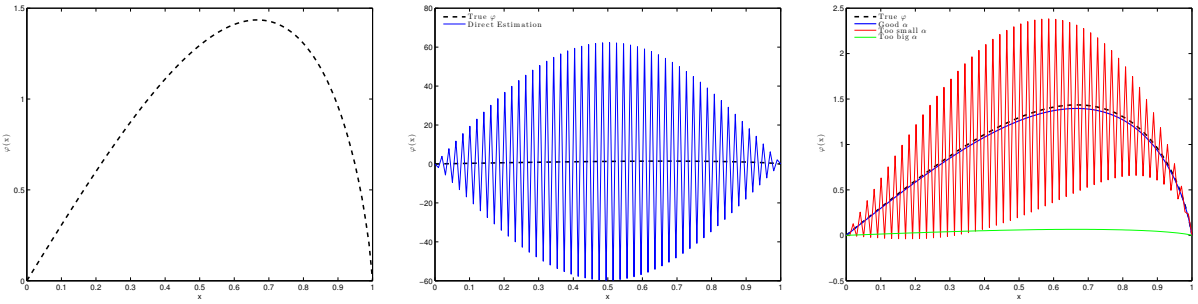
vertible, so that the solution φ is uniquely defined. However, its smallest eigenvalues are getting arbitrarily close to zero so that, in practice, the direct inversion leads to an explosive, non-continuous solution. Moreover, the fact that r is not observed and should be estimated introduces a further error which renders the *ill-posedness* of the problem even more severe.

The classical way to circumvent ill-posedness is to *regularize* the integral operator. Regularization, in this context, boils down to choosing a constant parameter which transforms the ill-posed into a well-posed inverse problem.

Therefore, estimating nonparametrically the shape of the function of interest requires, besides the usual selection issues related to the nonparametric estimation (e.g., selection of the smoothing parameters), also the choice of a regularization parameter. A sound criterion for choosing this tuning constant is extremely important, as an erroneous alternative will lead to misleading conclusions about the shape of the function of interest. Heuristically, the role of the regularization constant is to *smooth* the inverse mapping that is not continuous. Thus, its choice appears essential in applications, as *undersmoothing* leads one to obtain a solution that wiggle around the true function but does not give ultimately any guidance about its shape; *oversmoothing*, by contrast, shuts down the information coming from the data completely and delivers an almost constant solution.

Figure (1.1) 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. Define α to be our regularization parameter. A large value of α *oversmooths* the inverse mapping. The function obtained is the flat green line in Figure (1.1), 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.

The main aim of this chapter is therefore to propose and explore the properties of a criterion for a sound data driven selection of the regularization parameter when the so-called Tikhonov regularization is used to smooth the inverse mapping (Darolles et al., 2011*a*). The criterion advocated in this



(a) (b) (c)

Figure 1.1: The true function (a) and its numerical approximation by direct inversion of the operator (b) and using several values of the regularization parameter (c).

chapter is a simple *leave-one-out* cross-validation function. The intuition behind cross-validation is to choose the regularization constant that minimizes the prediction error for observation i, when the latter is not used to compute the estimator of φ. Tikhonov regularization is maintained in this chapter because of its simplicity. It can in fact be related in a straightforward way to the very simple linear framework or, more exactly, to the Ridge regression in finite dimensional parametric models.

Cross-validation (CV) has been already advocated as a viable solution to choose the regularization parameter in case of penalized Ridge regressions, and for ill-posed solutions of integral equations of the first kind (Wahba, 1977; Hansen, 1992; Groetsch, 1993; Vogel, 2002). However, in the former literature, the problem is by definition finite dimensional; while in the latter the integral operator is supposed to be known.

Golub et al. (1979) and Lukas (1993, 2006) discuss the application of Generalized Cross-Validation (GCV) to Ridge regressions and to the linear inverse problem in the statistical literature respectively. GCV is generally preferred to CV, as it does not require calculation of the estimator at each sample point and, therefore, it reduces computation time tremendously. However, GCV ignores the weight of each single data point in the prediction and the minimization of the objective criterion can be extremely ill-conditioned in presence of outliers.

Under the so-called Petrov-Galerkin regularization scheme, Marteau and Loubes (2012) discuss the properties of the adaptive selection of the regularization parameter when the conditional expectation operator in (1.1.3) is known. They prove an Oracle inequality for their minimization

criterion. Horowitz (2012) extends their framework when the conditional expectation operator is instead estimated, which is more relevant for econometrics. Recently, Breunig and Johannes (2011) have provided similar results for the estimation of linear functionals of the unknown function φ.

Fève and Florens (2010) discuss and prove the properties of a data driven selection of the regularization parameter under Tikhonov regularization. In order to obtain a rate optimal value of the parameter, they minimize the sum of squared residuals from the sample counterpart of equation (2.2.2), which is penalized in order to admit a minimum. This work shows that their criterion generally regularizes the function too much, therefore inducing a larger regularization bias in finite samples. Furthermore, when the function of interest is not smooth enough (in a sense that will be made more precise below), their criterion may not have a solution.

To the best of our knowledge, this chapter is the first that explores the properties of *leave-one-out* cross-validation procedure for the choice of the regularization parameter in nonparametric econometric problems. That is, when the integral operator that defines the inverse problem is estimated and not observed. Although we limit the presentation of results to the framework of nonparametric instrumental regressions, we believe that the criterion proposed here may be extended to other nonparametric problems in econometrics, where the function of interest is defined as the solution of an *ill-posed* inverse problem. For instance, to the estimation of structural quantile effects (Gagliardini and Scaillet, 2012); to the problem of density deconvolution (Carrasco and Florens, 2011); or to the estimation of the spectral density (Huang et al., 2011). We provide bounds in probability for the cross-validation function and show that the minimization of these bounds delivers a regularization parameter that goes to zero at a rate that is optimal in the Mean Squared Error sense. Our proofs use results about the relationship between the spectrum and the diagonal elements of a positive bounded operator.

A further contribution of this work is to extend the cross-validation criterion to the estimation of the derivatives of the regression function. This extension is interesting in several respects. From the theoretical standpoint, the regression function can be written as the integral of its derivative of any order. The integral operator smooths further the solution of our ill-posed inverse problem and allows one to obtain an estimate which is less oscillating. From the applied point of view, derivatives have a straightforward interpretation in many economic models as marginal effects.

Hence, the ultimate goal of the researcher may be to extract these marginal effects from the data. Consequently, we extend the theoretical results to obtain a data-driven value of the regularization parameter for direct estimation of any derivative of the regression function.

The chapter is structured as follows. The next session describes the main assumptions that are necessary to define our nonparametric estimator in this instrumental variable framework. We describe in details the estimation method and the choice of alpha in section (1.3). In particular, we discuss the selection of the regularization parameter with respect to the smoothness properties of both the regression function and the joint distribution of the endogenous variable and the instruments (so-called *source condition*). Moreover, we establish some results about the relationship between the choice of the smoothing parameter for the nonparametric estimation and the convergence bounds for the regularization constant. Section (1.3) also contains our main proofs about the properties of the cross-validation criterion for the direct estimation of φ. Section (1.4) presents the extension to the estimation of derivatives.

The chapter is concluded by an extensive simulation study in which we show that our cross-validation criterion seems to behave well in finite sample, and for different smoothness properties of the function φ and of the joint distribution of X and W. Finally, an empirical application to the estimation of the Engel curve for food, fuel and leisure in a sample of UK households shows the practical usefulness of our data-driven procedure.

1.2 The main framework

Let (Y, X, W) a random vector in $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$, such that:

$$Y = \varphi(X) + U, \quad \text{with} \quad \mathbb{E}(U|W) = 0. \tag{1.2.1}$$

For simplicity, the assumption that W and X are defined on the unit hypercube of dimension $p+q$ is maintained. Suppose further that $\varphi \in \mathbb{L}^2_X$, the space of square integrable functions of X. Define T, the conditional expectation operator which maps $\mathbb{L}^2_X$ into $\mathbb{L}^2_W$, and its adjoint T^*, which maps $\mathbb{L}^2_W$ into $\mathbb{L}^2_X$. Further denote by $\{\zeta_i, \psi_i, i \geq 0\}$, two orthonormal sequences in $\mathbb{L}^2_X$ and $\mathbb{L}^2_W$, respectively. In the following, Y is supposed to be observed, although the results of this chapter

also hold when Y is latent and the researcher observes $\tilde{Y} = \mathbb{1}(Y > 0)$, a binary transformation of it (see, Chapter 2).

Our framework needs the following high level assumption.

Assumption 1. *The joint distribution of the instruments W and the endogenous variable X is dominated by the product of the marginal distributions and its density, $f_{X,W}(x,w)$, is square integrable with respect to the product of the marginals.*

Notice that this assumption implies that T and T^* are Hilbert–Schmidt operators. This is a sufficient condition for compactness of T, T^* and TT^* (Carrasco et al., 2007). Moreover it implies the following (see, e.g. Kress, 1999; Conway, 2000).

Proposition 1.2.1. *There exists a singular value decomposition (SVD). That is, there is a non-increasing sequence of nonnegative numbers $\{\lambda_i, i \geq 0\}$, such that:*

(i) $T\zeta_i = \lambda_i \psi_i$.

(ii) $T^*\psi_i = \lambda_i \zeta_i$.

The existence of a SVD implies that the λ_i's are the eigenvalues of the operators T and T^* and ζ_i and ψ_i the corresponding eigenfunctions. Therefore, for any pair of functions $g \in \mathbb{L}_X^2$ and $m \in \mathbb{L}_W^2$, one can write:

$$\begin{aligned} (Tg)(w) &= \sum_{i=1}^{\infty} \lambda_i \langle g, \zeta_i \rangle \psi_i, \\ (T^*h)(x) &= \sum_{i=1}^{\infty} \lambda_i \langle m, \psi_i \rangle \zeta_i. \end{aligned}$$

Using operator's notations, equation (2.2.2) can be rewritten as follows:

$$T\varphi = r \tag{1.2.2}$$

The *ill-posedness* of the inverse problem arises because of the compactness of T and T^*, $\lambda_i \to 0$ as

$i \to \infty$ and therefore the inversion of the operator T would lead to the noncontinuous solution:

$$\varphi = T^{-1}r = \sum_{i=1}^{\infty} \frac{\langle r, \psi_i \rangle}{\lambda_i} \zeta_i.$$

As stressed in Darolles et al. (2011*a*), Assumption (1) *is not* a simplifying assumption but describes a realistic framework. The continuous spectrum of the operator depends on the joint distribution and it cannot be bounded from below by a strictly positive quantity. The following example clarifies the matter.

Example 1 (The Normal Case)**.** Suppose that $(X, W) \in \mathbb{R}^2$ is jointly normal with mean 0 and variance matrix given by: $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, with $|\rho| < 1$. Then the conditional distribution of X given $W = w$ is normal with mean equal to ρw and variance $1-\rho^2$. Therefore, the eigenvectors associated to the operator T are Hermite polynomials and its eigenvalues are given by $(\sqrt{\rho^2})^i$. Notice that, as $i \to \infty$, the eigenvalues are converging to 0, which causes the *ill-posedness* of the problem. ■

Finally assume that all other necessary identification conditions are satisfied (Darolles et al., 2011*a*; D'Haultfoeuille, 2011). In particular, the following is supposed to hold throughout:

$$T\varphi \overset{a.s.}{=} 0 \quad \Rightarrow \quad \varphi \overset{a.s.}{=} 0, \quad \forall \varphi \in \mathbb{L}_x^2.$$

This condition is related to the concept of completeness in statistics and, consequently, it is normally referred as *completeness condition*. Heuristically, it can be interpreted as a generalization of a rank condition for linear models with endogeneity. In particular, it implies that every non-constant and square integrable function of X is correlated with some square integrable function of W. Although recent work by Canay et al. (2013) has shown that this condition is not directly testable, Andrews (2011) and Chen et al. (2013) have established some genericity results, which allow one to claim that the completeness condition is normally satisfied by most distributions in a large class.

To cope with the noncontinuity of the inverse problem, this chapter follows the framework of Darolles et al. (2011*a*) and considers φ as the solution of the following penalized criterion:

$$\varphi^{\alpha} = \underset{\varphi \in \mathbb{L}_X^2}{\arg\min} \|T\varphi - r\|^2 + \alpha \|\varphi\|^2, \tag{1.2.3}$$

where α is called the regularization (or penalization) parameter. Therefore:

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

The idea behind Tikhonov regularization is to control via α the rate of the decay of the eigenvalues of T to 0. This introduces a regularization bias which converges to 0 with α. The rate of decrease to 0 of this bias depends on two main factors: the speed of decay of the λ_i's to 0; and the smoothness of the function φ. In particular, the former is related to the properties of the joint density of the vector (X, W) and determines how severe the inverse problem is.

Following Darolles et al. (2011*a*), these features are summarized in a single parameter $\beta > 0$.

Assumption 2 (Source condition)**.** *For some real $\beta > 0$, and a pair of functions $g \in \mathbb{L}_X^2$ and $m \in \mathbb{L}_W^2$, one has:*

$$\sum_{i=1}^{\infty} \frac{\langle g, \zeta_i \rangle^2}{\lambda_i^{2\beta}} < \infty, \qquad \textit{and} \qquad \sum_{i=1}^{\infty} \frac{\langle m, \psi_i \rangle^2}{\lambda_i^{2\beta}} < \infty.$$

An equivalent way of stating this assumption is to say that, for a given $v \in \mathbb{L}_x^2$,

$$\varphi = (T^{*}T)^{\frac{\beta}{2}} v,$$

which implies:

$$\varphi \in \mathcal{R}\left((T^{*}T)^{\frac{\beta}{2}}\right).$$

This notation clearly links the properties of the function φ with the ones of the joint distribution of (X, W), through the conditional expectation operator T.

Notice that the source condition in (2) may not hold when the eigenvalues have an exponential rate of decay, as in the Gaussian case presented in example (1). In fact, it may be impossible to find a strictly positive value of β which satisfies (2), unless the function of interest is sufficiently smooth. That is, the Fourier coefficients of the function φ should decay sufficiently fast to zero. In this case the inverse problem is said to be *severely ill-posed* and it ought to be distinguished from the *mildly ill-posed* case, where eigenvalues have a polynomial rate of decay. That is, when $\lambda_i \approx i^{-b}$, for $b > 0$. Separating these two cases is also essential because they lead to different rates of convergence of the regularized estimator (see also Chen and Reiss, 2011). In fact, while in the

mildly ill-posed case, the estimator has a polynomial rate of convergence, in the severely ill-posed case, rates of convergence are polynomial in the logarithm of the sample size. An important remark about the results presented in this chapter is that they do not hold for the *severely ill-posed* case.

Nonetheless, if assumption (2) is satisfied, one obtains that the rate of convergence of the regularization bias as following:

$$\|\varphi^{\alpha} - \varphi\|^2 = O_p\left(\alpha^{\min(\beta,2)}\right).$$

The term $\min(\beta, 2)$ arises because Tikhonov regularization cannot take advantage of an order of regularity higher than 2. This is related to the so-called *qualification* of a regularization method (see Engl et al., 2000). It is possible to increase the qualification of Tikhonov regularization, by considering an iterative approach (Fève and Florens, 2010), i.e.:

$$\begin{aligned}
\varphi^{\alpha}_{(1)} &= (\alpha I + T^*T)^{-1} T^*r, \\
&\vdots \\
\varphi^{\alpha}_{(k)} &= (\alpha I + T^*T)^{-1} \left(T^*r + \alpha\varphi^{\alpha}_{(k-1)}\right), \\
&\vdots
\end{aligned}$$

This iterative method allows one to exploit higher orders of regularity of the function φ. In fact:

$$\|\varphi^{\alpha}_{(k)} - \varphi\|^2 = O_p\left(\alpha^{\min(\beta,2k)}\right), \qquad \forall k \geq 1. \tag{1.2.4}$$

In the following, $\varphi^{\alpha}_{(1)} = \varphi^{\alpha}$, and it is referred to as the non-iterated Tikhonov solution of (1.2.3).

1.3 Nonparametric estimation and the choice of α

Suppose we observe $\{(y_i, x_i, w_i), i = 1, \ldots, N\}$, an iid realization of the random variables (Y, X, W).[3]

For simplicity of exposition, only the local constant nonparametric estimation of the function φ is analyzed here. Consider the class of continuous bounded kernels K_h of order $\rho \geq 2$ with bandwidth

[3] As usual, this assumption could be relaxed to allow for stationary mixing time series (Hansen, 2008).

parameter h.[4] For simplicity, at least in the exposition of the theoretical results, we assume that the same bandwidth h_N is used for both X and W. The estimation of φ consists of 3 main steps:

(i) Estimate r, the conditional expectation of Y given W. Note that this gives also an estimator of the conditional expectation operator T, which corresponds to the matrix of kernel weights (Fève and Florens, 2010). This can be achieved using the classical Nadaraya-Watson kernel estimator, i.e.:

$$\hat{r} = \frac{\sum_{i=1}^{N} y_i K_{h_N}(w_i - w)}{\sum_{i=1}^{N} K_{h_N}(w_i - w)} = \hat{T} y.$$

(ii) In the same way, an estimator of the operator T^* is obtained as the conditional expectation of $\hat{r}$ given X, i.e.:

$$\hat{T}^* \hat{r} = \frac{\sum_{i=1}^{N} \hat{r}_i K_{h_N}(x_i - x)}{\sum_{i=1}^{N} K_{h_N}(x_i - x)}.$$

(iii) Finally, for a given sample value of the parameter α, say α_N, the Tikhonov regularized estimator of φ is retrieved as:

$$\hat{\varphi}^{\alpha_N} = \left(\alpha_N I + \hat{T}^* \hat{T}\right)^{-1} \hat{T}^* \hat{r}.$$

The following theorem contains the rate of convergence in MSE for the estimator $\hat{\varphi}^{\alpha_N}$.

Theorem 1.3.1 (Darolles et al. 2011*a*)**.** *Under assumptions (1) and (2), and the convergence of the regularization bias given in (1.2.4):*

$$\|\hat{\varphi}^{\alpha_N} - \varphi\|^2 = O_P\left[\frac{1}{\alpha^2}\left(\frac{1}{N} + h_N^{2\rho}\right) + \left(\frac{1}{N h_N^{p+q}} + h_N^{2\rho}\right)\alpha_N^{\min(\beta-1,0)} + \alpha_N^{\min(\beta,2)}\right]. \tag{1.3.1}$$

■

The proof of this result is not reported here and we refer interested readers to Darolles et al. (2011*a*) for details. Notice that the upper bound in this theorem is given by the sum of three terms. The

[4] For a more general theoretical presentation, see Darolles et al. (2011*a*).

first component is a variance term that explodes, for a fixed sample size N, when the value of α approaches zero. The last component is the usual bias term which disappears when α gets smaller. These two terms need to be balanced with the right choice of the regularization parameter. Finally, the term in the middle is related to the estimation of the rhs of equation (1.2.2). As a matter of fact, when $\beta \geq 1$, it simply gives the upper bound for the nonparametric estimation of r. On the contrary, when $\beta < 1$, the nonparametric estimation error is multiplied by a component that goes to infinity with α. The latter term corresponds to a penalty that slows down the rate of convergence when our inverse problem becomes more ill-posed or the function φ is relatively less smooth.[5]

Darolles et al. (2011*a*) discuss the assumptions that make this upper bound for the MSE converging to 0, upon some premises on the convergence of the bandwidth parameter to 0 as the sample size grows. Namely, they suppose that the bandwidth can be chosen to be bounded in probability by $N^{-1/2\rho}$, to exploit the parametric rate of convergence of the first term in (1.3.1). They discuss the choice of the regularization parameter, given this particular bandwidth selection.

Here, the choice of the bandwidth is instead supposed to be a function of the dimension of the endogenous variable, p, the dimension of the instrument, q, and the order of the kernel ρ, i.e.:

$$h_N^{2\rho} \approx N^{-\gamma(p,q,\rho)}, \qquad \text{with} \qquad 0 < \gamma(p,q,\rho) \leq 1.$$

where $\gamma(\cdot)$ is a real function. For instance, if the bandwidth is chosen such that the squared bias and the variance of the nonparametric regression converge at the same rate, one has:

$$\gamma(p,q,\rho) = \frac{2\rho}{2\rho + p + q}.$$

In the following, for simplicity, define $\gamma \equiv \gamma(p,q,\rho)$. Heuristically, α_N has to be chosen to converge to 0 at some rate, which depends on the sample size. When $\beta \geq 1$, the result is straightforward, as the middle term in the decomposition does not depend on α. Otherwise, the rate of convergence depends on the choice of the bandwidth parameter, i.e. on the choice of γ.

The optimal rate of convergence for α_N, which makes the MSE in (1.3.1) asymptotically 0 can

[5] A small value of β could also be an indication of weak instruments. This point is not explicitly discussed in this chapter and it is left for further research.

therefore be expressed in terms of β and γ.

Corollary 1.3.2 (Convergence of the upper bound to 0 and rate optimal α_N)**.** *The rate optimal value of α_N, for which (1.3.1) $\overset{a.s.}{\to} 0$, is such that:*

(i) If $\beta \geq 1$ and $0 < \gamma \leq 1$, so that $N^{\gamma}\alpha_N^2 \to \infty$, and $Nh_N^{p+q} \to \infty$, then:

$$\alpha_N \approx N^{-\frac{\gamma}{\min(\beta,2)+2}}.$$

(ii) If $\beta < 1$ and

$$\gamma \leq \frac{2\rho}{2\rho + p + q},$$

in such a way that $N^{\gamma}\alpha_N^2 \to \infty$, and $Nh_N^{p+q} \to \infty$, then:

$$\alpha_N \approx N^{-\frac{\gamma}{\beta+2}}.$$

(iii) If $\beta < 1$ and

$$\gamma \in \left(\frac{2\rho}{2\rho + p + q}, \frac{2\rho(\beta+2)}{(p+q)(\beta+2) + 2\rho} \right),$$

in such a way that $N^{\gamma}\alpha_N^2 \to \infty$, and $Nh_N^{p+q} \to \infty$, then:

$$\alpha_N \approx N^{-\frac{\gamma}{\beta+2}}.$$

Otherwise, if:

$$\gamma \in \left[\frac{2\rho(\beta+2)}{(p+q)(\beta+2) + 2\rho}, \frac{2\rho}{p+q} \right),$$

in such a way that $Nh_N^{p+q}\alpha_N^{1-\beta} \to \infty$, then:

$$\alpha_N \approx N^{-1+\frac{p+q}{2\rho}\gamma}.$$

Proof. See the Appendix. ■

Notice, in particular, that, when $\beta \geq 1$, the MSE converges to 0, independently of the choice of the bandwidth. Nonetheless, it would be necessary to choose the bandwidth parameter in such a

way to balance the variance and the squared bias of the nonparametric estimator. Therefore:

$$\gamma = \frac{2\rho}{2\rho + p + q}. \tag{1.3.2}$$

On the one hand, this generally slows down the convergence of α to 0, by a factor which is proportional to γ. On the other hand, following the arguments in Darolles et al. (2011*a*), with $\gamma = 1$, the variance term in α converges faster to 0. However, this generates higher variance in the nonparametric estimation (second term of the upper bound in 1.3.1). Moreover, it requires additional constraints on the value of ρ. In fact, in order to avoid the variance term of the nonparametric estimation to diverge, it is necessary to assume, with $\gamma = 1$:

$$\rho > \frac{p+q}{2}. \tag{1.3.3}$$

This constraint hardly matters in practice when the dimensions of the endogenous variable and the instruments are small. For instance, when p and q are both equal to 1. Nevertheless, when the researcher has the possibility to use more instruments, she needs to employ higher order kernels, that are seldom used in practice. A different approach would be to use local polynomials estimation, with the order of the polynomial that increases with the number of instruments used. A similar reasoning applies if the value of γ is chosen too small. In this case, the squared bias in the nonparametric estimation is going to play the role of further slowing down the convergence of (1.3.1) to 0.

When $\beta < 1$, the choice of the bandwidths enters directly the convergence to 0 of the regularization parameter. The case $\beta < 1$ arises for example when the instruments are not very strong; but also when the function of interest is not sufficiently smooth or when the inverse problem is more severely *ill-posed*. As a matter of fact, for given smoothness characteristics of the function of interest, if the decay of the eigenvalues of T is faster, a smaller β is implied by the source condition given in Assumption (2). If γ is taken equal to 1, point *iii* of Corollary (1.3.2) shows again that one needs condition (1.3.3) in order to obtain a value of α that does not diverge with the sample size. The optimal selection of the bandwidth for nonparametric regressions instead guarantees the squared bias and the variance to be balanced and appears to be, in this case too, the most reasonable

choice.

A last important remark about the rate of convergence is related to the dimension of the instrument W. In standard nonparametric regression, the larger the dimension of the conditioning variable, the slower the rate of convergence of the estimator (so-called *curse of dimensionality*). In the instrumental variable setting, this seems a contradictory result: the more instruments added, the more precise should be the estimation of the function of interest φ. Hence, the result of Theorem (1.3.1) is designed in such a way that the dimension of the instrument does not matter for the speed of convergence of the estimator when the bandwidth is chosen proportional to N^{-1}. However, Corollary (1.3.2) shows that the dimension of W matters independently of the choice of the bandwidth. If γ is chosen equal to 1, in order to exploit the parametric rate of convergence of the first term in (1.3.1) and for a given dimension of the endogenous variable X, constraint (1.3.3) binds the number of instruments that can be used for a given order of the kernel. In the same way, an optimal choice of h, in the sense of nonparametric regressions, takes into account the dimension of W and deteriorates the rate of convergence of $\hat{\varphi}^{\alpha}$ toward its true value. The latter approach, while it has clear disadvantages in terms of rate of convergence, still ensures that the estimator does not diverge when more instruments are used for inference. Furthermore, equation (1.2.2) defines the function φ with respect to the conditional expectation of the dependent variable Y given W, defined as r. Heuristically, the more precise the estimation of r, the more precise the estimation of φ.

In the following, it is therefore assumed that the bandwidth is chosen by fixing γ as in (1.3.2). Methods like cross-validation or the improved Akaike Information Criterion of Hurvich et al. (1998) are known to deliver such optimal selection (see, e.g., Härdle and Marron, 1985; Li and Racine, 2007).

Upon the choice of the bandwidth parameter, the main objective of this work is to devise a method which delivers a rate optimal value of α_N and that works reasonably well in practice, i.e. it adapts to the characteristics of the data at hand. This chapter considers criteria of the form:

$$P(\alpha_N)\|\hat{T}\hat{\varphi}^{\alpha_N} - \hat{r}\|^2, \tag{1.3.4}$$

where $P(\alpha_N)$ is a penalization function. These criteria select α_N as the minimizer of the sum of

squared residuals in (1.2.2).

Fève and Florens (2010) propose a data-driven method for the choice of α_N which is based on the minimization of the following criterion:

$$SSR(\alpha_N) = \frac{1}{\alpha_N}\|\hat{T}\hat{\varphi}_{(2)}^{\alpha_N} - \hat{r}\|^2, \tag{1.3.5}$$

where $\hat{\varphi}_{(2)}^{\alpha_N}$ is twice iterated Tikhonov estimator, i.e.:

$$\hat{\varphi}_{(2)}^{\alpha_N} = \left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\left(\hat{T}^*\hat{r} + \alpha_N\hat{\varphi}_{(1)}^{\alpha_N}\right) = \left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\left[I + \alpha_N\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\right]\hat{T}^*\hat{r}.$$

This criterion belongs to family (1.3.4), where $P(\alpha_N) = 1/\alpha_N$. Although, in their framework, estimation is carried on using a simple non-iterated Tikhonov approach, the twice iterated Tikhonov serves the scope of increasing the qualification and, therefore, reduces the regularization bias. Fève and Florens (2010) prove, in the case of transformation models, that this criterion produces a choice of α_N which is rate optimal.

In the case of instrumental variable regressions, the following result can be proved.

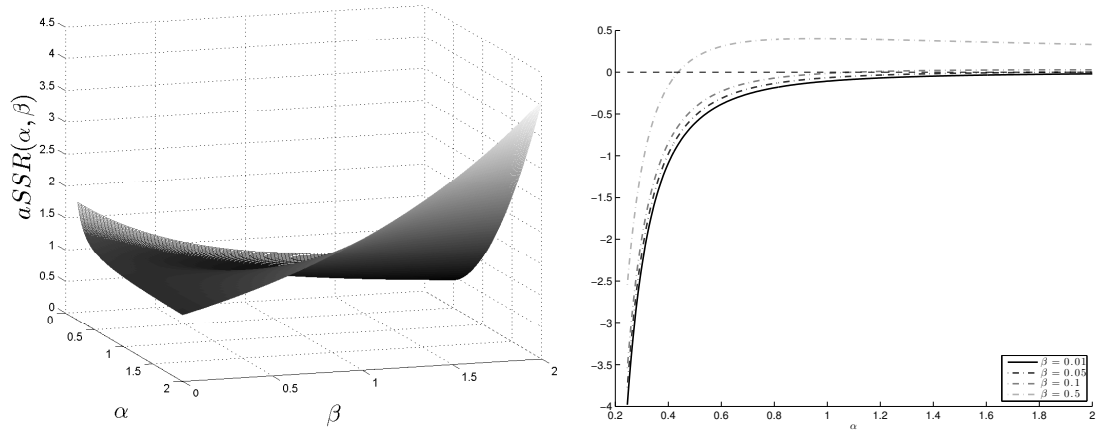
Lemma 1.3.3. *The $SSR(\alpha_N)$ criterion in (1.3.5) is bounded in probability by:*

$$aSSR(\alpha_N, \beta) = \frac{1}{\alpha_N}\left[\frac{1}{\alpha_N}\left(\frac{1}{N} + h_N^{2\rho}\right) + \left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right)\left(1 + \alpha_N^{\min(\beta,1)}\right) + \alpha_N^{\min(\beta+1,4)}\right].$$

Proof. See the Appendix. ■

This criterion has the same speed of convergence as the original MSE in (1.3.1). Therefore, upon the optimal choice of the bandwidth, theoretically, α is selected in such a way that the variance and the bias term converge at the same speed. However, despite this optimality result, it is impossible using this criterion to balance the two terms in the asymptotic upper bound when β becomes smaller. This is due to the fact that the regularization bias converges to 0 too slowly (see, also Engl et al., 2000, for a discussion). The heuristic explanation is related to the fact that the regularization bias α^β stays roughly constant for any value of α. While the variance term gets very large when the α is close to 0 and, for a fixed sample size N, decays to zero only when α grows larger. The minimization of this function thus leads to choose a parameter α which only affects

the variance term. That is, a very large value of it.



(a) (b)

Figure 1.2: A 3 dimensional plot of $aSSR(\alpha_N, \beta)$ (left), and its derivative wrt α_N for several values of β (right).

Therefore, for $\beta < 1$, the SSR criterion may lead to *over-regularize* the solution of the inverse problem, i.e. choose a large value of α_N. Moreover, when β gets sufficiently close to 0, the only solution is obtained for $\alpha_N \to \infty$. Figure (1.2) graphically illustrates the issue. On the left panel, the function $aSSR(\cdot, \cdot)$ is plotted for $N = 1000$, $\rho = 2$, $p = 1$, $q = 2$, and for a reasonable range of values for the two parameters α_N and β, with γ as in (1.3.2).

It can be noticed that, when β is smaller than a certain threshold, the function is strictly decreasing to 0 as $\alpha_N \to \infty$. On the right panel, the derivative of the function $aSSR(\cdot, \cdot)$ with respect to α_N is plotted for several values of β. As it can be seen, the derivative converges to 0 as α_N grows, but it never crosses the 0 line.

A possible way to correct for this numerical problem is to modify the penalization term $P(\alpha_N)$, in such a way that the variance term does not converge too fast to zero as α increases. However, this solution does not seem to be practicable, as it requires some previous knowledge of the parameter β.

To overcome the deficiencies of available methods, this chapter discusses a *leave-one-out* procedure for the selection of the regularization parameter. Define the cross-validation function:

$$CV(\alpha_N) = \|\hat{T}\hat{\varphi}_{(-i)}^{\alpha_N} - \hat{r}\|^2, \tag{1.3.6}$$

where $\hat{\varphi}_{(-i)}^{\alpha_N}$ is the non iterated Tikhonov estimator of φ that has been obtained by removing the i^{th} observation from the sample. The heuristic idea behind the choice of this function is similar to the one exploited in the selection of the smoothing parameter by cross-validation in nonparametric regressions. One is looking for the value of α_N that minimizes the prediction error for the observation i, when this observation is not used to compute the estimator of φ. The optimal α_N is therefore obtained as:

$$\alpha_N^{CV} = \arg\min_{\alpha>0} CV(\alpha_N).$$

The following result can be proven.

Theorem 1.3.4. *The $CV(\alpha_N)$ criterion is bounded in probability by:*

$$aCV(\alpha_N, \beta) = \left(\frac{\alpha_N + 1}{\alpha_N}\right)^2 \left[\frac{1}{\alpha_N}\left(\frac{1}{N} + h^{2\rho}\right) + \alpha_N^{\min(\beta+1,2)} + \left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\right].$$

Proof. See the Appendix. ■

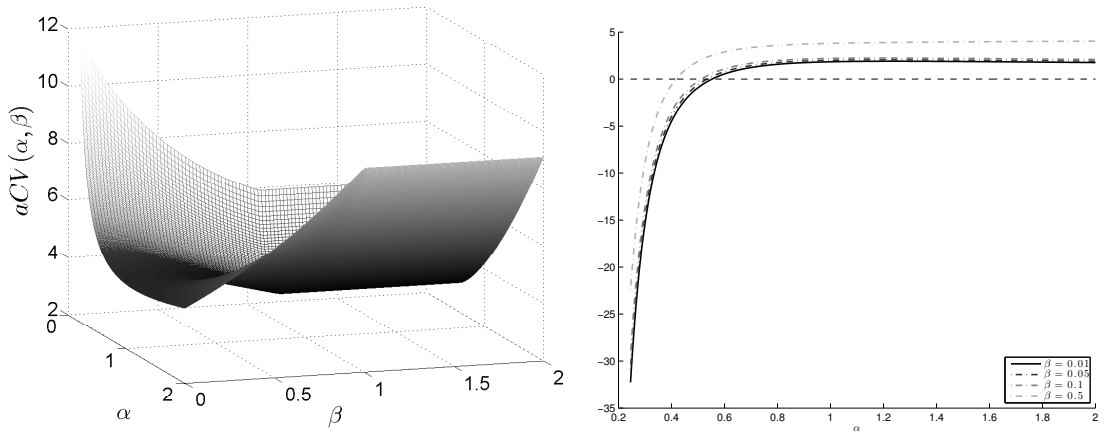
The main argument behind this result is that the CV function can be written, in finite samples, as the sum of squared residuals, where each term of the sum is weighted by the corresponding diagonal element of a properly defined matrix. When the sample size converges to infinity, this matrix is more properly defined as on operator, so that the main goal of the proof is to bound its diagonal terms. This explains the penalizing term in α that is multiplying the upper bound for the sum of squared residuals.

An example about the behavior of this criterion function is reported in figure (1.3). Consider, as before, a case where $N = 1000$, $\rho = 2$, $p = 1$, $q = 2$, and the bandwidth is chosen such that:

$$\gamma = \frac{2\rho}{p + q + 2\rho}.$$

As it is visible from the figure, the CV function attains a minimum even for very small values of β.

It is interesting to notice that, asymptotically, the CV criterion also belongs to the family (1.3.4).



(a) (b)

Figure 1.3: A 3 dimensional plot of $aCV(\alpha_N, \beta)$ (left), and its derivative wrt α_N for several values of β (right).

The penalizing factor is given by:

$$P(\alpha_N) = \left(\frac{\alpha_N + 1}{\alpha_N}\right)^2 = 1 + \frac{1}{\alpha_N} + \frac{1}{\alpha_N^2},$$

which contains the term $1/\alpha_N$. However, it has also two additional components: a constant and a quadratic term. When α_N approaches 0 too fast, then the quadratic term increases the value of the cross-validation function. By contrast, when α_N approaches infinity too fast, the constant term is going to increase the weight of the residual sum of squares. Therefore, the cross-validation method is similar in spirit to the minimization of the sum of squared residuals proposed in Fève and Florens (2010). However, it is not undermined when β gets too close to 0.

This section is concluded with the following result about the rate of convergence of the α_N parameter chosen using our cross-validation procedure.

Corollary 1.3.5. *For an optimal choice of the smoothing parameter h, the minimization of the cross-validation function (1.3.6) leads to a choice of the regularization parameter α_N, such that:*

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

Proof. The value of α_N is chosen, such that:

$$\frac{1}{\alpha_N}\left(\frac{1}{N}+h^{2\rho}\right)\approx\alpha_N^{\min(\beta+1,2)}.$$

Since the bandwidth is proportional to $N^{-\frac{1}{p+q+2\rho}}$, one has that:

$$\frac{1}{\alpha_N}\left(\frac{1}{N}+h^{2\rho}\right)\approx\frac{1}{\alpha_N}N^{-\gamma},$$

and the result easily follows. ■

The cross-validation criterion leads to a choice of the regularization parameter similar to the one achieved using the discrepancy principle of Morozov (1967).[6] The discrepancy principle consists in selecting the value of α, such that:

$$\|\hat{T}\hat{\varphi}^{\alpha_N}-\hat{r}\|\leq\tau\delta,$$

where τ is a positive constant, and δ represents some observational error. This error is related to the approximation of the right hand side of equation (1.2.2) (see, e.g. Engl et al., 2000; Mathé and Tautenhahn, 2011; Blanchard and Mathé, 2012). In our case, δ could be approximated by the nonparametric estimation error in r, i.e. $N^{-\gamma}$. However, the open question remains about the choice of the tuning constant τ.

The cross-validation criterion eliminates this further need and achieves the same order of convergence. The choice of α is rate optimal, following the results of Darolles et al. (2011*a*), only when $\beta\leq 1$. Notice that this is not a serious flaw, when the sample has moderate size. However, as the sample size grows, and the regularity of the function of interest is greater than 1, it would lead to *under-regularize* the solution of the inverse problem, i.e. choosing a value of the regularization parameter which decays to 0 faster than the optimal one. This is a known feature of *leave-one-out* methods, for instance, in the case of the selection of the smoothing parameter in standard nonparametric regressions (Li and Racine, 2007).

For higher values of β, it would be feasible to achieve the optimal rate of using the same idea as

[6] A similar rate of convergence is achieved by all so-called *heuristic* methods that selects the regularization parameter as the minimizer of the prediction error. Interested readers are referred to Ch.4 and 5 of Engl et al. (2000) for a discussion on this topic.

in the SSR method of Fève and Florens (2010). That is, we could increase the qualification of the regularization procedure with an iterated Tikhonov approach. However, the underlying idea of cross-validation function is to choose the value of the tuning constant that minimizes the prediction error. Therefore, if an iterative approach is used in the CV criterion to find such a value of alpha, it also needs to be used for estimation.

An alternative approach would be to consider the properties of the CV criterion for the penalization of the function in Hilbert scales, i.e., the penalization of the derivatives of the function, instead of the function itself (Florens et al., 2011). This last point is discussed in the next section.

1.4 A more general approach to the Regularization in Hilbert Scale

Following the result in the previous section, it can be shown that the cross-validation procedure of this chapter has a broader scope of application, beyond the standard $\mathbb{L}^2$ penalization of the function of interest. Introduce the additional assumption that $\varphi \in \mathbb{C}^u$, i.e. φ has at least u continuous derivatives, with $u \geq 0$. Then, the function of interest can be approximated by the integral of its derivative of any order.

Define $\{L^s, s \in \mathbb{R}, s \geq 0\}$, the unbounded, self-adjoint and strictly positive family of operators, with the convention that $L^0 = L^s L^{-s} = I$, the identity operator. For each value of s, their domain is such that:

$$\mathcal{D}(L^s) = \left\{ \varphi \in \mathbb{C}^s \quad : \quad \varphi^{(s)} \in \mathbb{L}^2_X \quad , \quad \varphi(0) = \varphi'(0) = \cdots = \varphi(0)^{(s-1)} = 0 \right\}.$$

When $s \geq 0$, this domain is called the Hilbert Scale induced by L^s (see Engl et al., 2000; Krein and Petunin, 1966). Note that these spaces are densely and continuously embedded into each other, i.e. for any $t > s$, $\mathcal{D}(L^t) \subset \mathcal{D}(L^s)$. The boundary conditions imposed on the first $s-1$ derivatives ensure that the operator L^s has a bounded inverse L^{-s}.

By means of the definition of the operator L^s, φ can be now defined as the solution of:

$$\min_{\varphi^{(s)} \in \mathcal{D}(L^s)} \|T\varphi - r\|^2 + \alpha \|L^s \varphi\|^2,$$

which gives:

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

where $B = TL^{-s}$ and $\varphi^{(s),\alpha}$ is the regularized version of φ in the norm induced by the Hilbert scale L^s. A detailed explanation on how to approximate L^s, at least when s is equal to 1, is given in Chapter 3 of this manuscript and in Florens and Racine (2012).[7] This section explores the extension of the CV selection criterion of Theorem (1.3.4) to this more general case.

Assumptions stated in section (1.2) are maintained here. In particular, the operator T is assumed to be one to one and the solution φ exists.[8] However, some further assumptions are needed that link the operator T with the Hilbert scale induced by L^s (see also Carrasco et al., 2013; Engl et al., 2000; Florens et al., 2011). Denote by $\|x\|_s = \|L^s x\|$ and $\langle x, y\rangle_s = \langle L^s x, L^s y\rangle$, the norm and the inner product induced by the operator L^s, respectively.

Assumption 3. *The operator T satisfies the following inequality:*

$$\underline{m}\|g\|_{-a} \leq \|Tg\| \leq \overline{m}\|g\|_{-a},$$

for any $g \in \mathcal{D}(L^s)$, $a > 0$ and $0 < \underline{m} < \overline{m} < \infty$.

The scalar a measures the degree of *ill-posedness* of the inverse problem through the properties of the operator T, i.e. the joint distribution of (X, W). Then for B defined as above, $|\nu| \leq 1$ and $s \geq 1$, Assumption (3) implies the following inequality (see Engl et al., 2000, Corollary 8.22, p. 214):

$$\underline{c}(\nu)\|g\|_{-\nu(a+s)} \leq \| \left(B^*B\right)^{\nu/2} g\| \leq \overline{c}(\nu)\|g\|_{-\nu(a+s)}, \tag{1.4.1}$$

[7] Notice that, in practice, L is defined to be the first order differential operator, which is generally not self-adjoint. To obtain a self-adjoint construction of it, it is possible to define it as $L\varphi = \sqrt{-\varphi^{(2)}}$ (see also Carrasco et al., 2013).

[8] See Florens et al. (2011) for the non identified case of Tikhonov regularization in Sobolev norm.

for any $g \in \mathcal{D}\left((B^*B)^{\nu/2}\right)$ with $\underline{c}(\nu) = \min\{\underline{m}^\nu, \overline{m}^\nu\}$ and $\overline{c}(\nu) = \max\{\underline{m}^\nu, \overline{m}^\nu\}$.

Note that inequality (1.4.1) also entails that:

$$\mathcal{D}\left((B^*B)^{\nu/2}\right) = \mathcal{D}\left(L^{\nu(a+s)}\right). \tag{1.4.2}$$

Furthermore, Assumption (3), together with the fact that $\varphi \in \mathcal{D}(L^u)$ implies the source condition (2), with $\beta = u/a$ (Carrasco et al., 2013; Florens et al., 2011). Heuristically, this can be explained by the fact that the source condition summarizes the *ill-posedness* of the inverse problem, which is determined by the regularity of the function φ, i.e. its number of continuous derivatives, and the properties of the conditional expectation operator T. Formally, for any value of s and u, $L^s\varphi \in \mathcal{D}(L^{u-s})$, that by (1.4.2) implies $\varphi \in \mathcal{D}\left((B^*B)^{\frac{u-s}{2(a+s)}}\right)$. Therefore, there exists a vector $v \in \mathbb{L}_X^2$, such that:

$$L^s\varphi = (B^*B)^{\frac{u-s}{2(a+s)}}\, v.$$

For $s = 0$, this leads to:

$$\varphi = (T^*T)^{\frac{u}{2a}}\, v = (T^*T)^{\frac{\beta}{2}}\, v, \quad \text{with} \quad \beta = \frac{u}{a},$$

which is the source condition, as stated above (see also Carrasco et al., 2007, 2013).

Before presenting the main result of this section, similarly to the baseline case, we need to obtain the order of the regularization parameter that minimizes the upper bound for the mean squared error, when the s^{th} derivative of the function is penalized. Notice that the upper bound is on the s^{th} derivative of φ, rather than on φ itself. As a matter of fact, $\varphi^{(s)}$ is the direct solution of the inverse problem, and the regularization parameter has to be chosen to target this estimator. Consequently, the following theorem establishes an upper bound for $\hat{\varphi}^{(s),\alpha_N}$ and derives the optimal order for α_N.

Theorem 1.4.1. *Suppose that φ is u times differentiable, and that assumptions (1), (2), and (3) hold. Suppose further the s^{th} derivative of φ is estimated, where $s \leq u \leq a + 2s$. Then:*

$$\|\hat{\varphi}^{(s),\alpha_N} - \varphi^{(s)}\|^2 = O_P\left[\frac{1}{\alpha^{\frac{2a+s}{a+s}}}\left(\frac{1}{N} + h_N^{2\rho}\right) + \left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right)\alpha_N^{\frac{u}{a+s}-1}\|\varphi\|_u^2 + \alpha_N^{\frac{u-s}{a+s}}\|\varphi\|_u^2\right].$$

Proof. See the Appendix. ■

A generalization of this theorem can be found in Florens et al. (2011), Johannes et al. (2011) and Carrasco et al. (2013). An essential implication of this result is that the penalization of regression derivatives increases the *qualification* of the Tikhonov regularization, upon the assumption that T is one-to-one. It therefore allows one to exploit higher order of smoothness of the function φ.

It is straightforward to notice that, when $s = 0$, the result of this theorem is equivalent to Theorem (1.3.1) above. However, in this case, the role of the parameter β, that defines the source condition, is more clearly decomposed into its two main determinants: the smoothness of the regression function, given by u; and the *ill-posedness* of the inverse problem, expressed by a. An increase in u, everything else held constant, leads to a faster convergence of the regularization bias. Furthermore, it also impacts the middle term, which is related to the nonparametric estimation error, that also converges faster to 0. By contrast, an increase in a, everything else held constant, impacts both the bias and the variance term. In particular, it reduces the speed of convergence of the former, and it also accelerates the explosion of the latter, when α tends to 0.

If we suppose that the bandwidth has been chosen so that $h_N^{2\rho} \approx N^{-\gamma}$, and,

$$\alpha^{\frac{2a+s}{a+s}} N^{\gamma} \to \infty,$$

then the variance term dominates the middle term for every $a > 0$ and therefore:

$$\alpha \approx \left(\frac{N^{-\gamma}}{\|\varphi\|_u^2} \right)^{\frac{a+s}{2a+u}}. \tag{1.4.3}$$

We can finally state the main result of this section.

Theorem 1.4.2. *Suppose that φ is u times differentiable, and that assumption (3) holds. Suppose further that φ is estimated by penalization of its s^{th} derivative, where $s \leq u \leq a + 2s$. Then, the cross-validation criterion (1.3.6) is bounded by the following function:*

$$\begin{aligned} aCV(\alpha, u, s, a) = \left(\frac{\alpha + \|B\|}{\alpha} \right)^2 \Bigg[& \alpha^{-\frac{a}{a+s}} \left(\frac{1}{N} + h^{2\rho} \right) \\ & + \alpha^{\frac{u}{a+s}} \|\varphi\|_u^2 \left(\frac{1}{Nh^{p+q}} + h^{2\rho} \right) + \alpha^{\frac{a+u}{a+s}} \|\varphi\|_u^2 + \left(\frac{1}{Nh^{p+q}} + h^{2\rho} \right) \Bigg]. \end{aligned}$$

Proof. See the Appendix. ■

For $s = 0$, the result of Theorem (1.4.2) is just a generalization of Theorem (1.3.4). This trivially implies that the CV α is chosen in such a way that:

$$\alpha^{CV} \approx \left(\frac{N^{-\gamma}}{\|\varphi\|_u^2} \right)^{\frac{a+s}{2a+u}},$$

which is of the same order as the optimal regularization parameter in equation (1.4.3). Again, this selection of the optimal parameter attains the same rate as the discrepancy principle of Morozov (see Engl et al., 2000). Furthermore, as already outlined above, since the qualification of the Tikhonov regularization increases when taking higher order derivatives, depending on the value of s, the criterion delivers rate optimal results also when $\beta > 1$.

1.5 A Numerical Illustration

In order to illustrate the small sample properties of the cross-validation procedure and to compare it to existing methods, we run two separate simulation schemes. The first setup is similar to the one employed in Hall and Horowitz (2005). The second is a variant of the setup used in Darolles et al. (2011*a*), where we introduce heteroskedasticity in the residuals.

In both simulation schemes, we employ second order Gaussian kernels. The conditional expectation operators T and T^* are estimated as the matrix of kernel weights from the local constant nonparametric regressions of Y on W, and of $\hat{r} = \hat{\mathbb{E}}(Y|W)$ on X (see also, Fève and Florens, 2010 and Chapter 3). Bandwidths are selected using least square cross-validation.[9]

In order to assess the performance of the two criteria, results are compared to those obtained with an *optimal* α. This optimal value is defined as the minimizer of the following mean squared error (MSE) function:

$$\alpha^{OPT} = \arg\min_{\alpha>0} \|\hat{\varphi}^\alpha - \varphi\|^2 .$$

Notice that this criterion produces the optimal value of α, given the estimation error. That is, for a given value of the smoothing parameters.

[9] Codes are available from the author upon request.

Furthermore, for the case of direct penalization of the function, we also run a comparison with the Generalized Cross-Validation (GCV) criterion of Golub et al. (1979). The minimizer of this criterion is defined as:

$$\alpha^{GCV} = \arg\min_{\alpha>0} \left\| \frac{\hat{T}\hat{\varphi}^{\alpha} - \hat{r}}{trace\left(\alpha * (\alpha I + \hat{T}^{*}\hat{T})\right)^{-1}} \right\|^{2}.$$

The properties of this selection criterion for noisy integral equations of the first kind have been established by Lukas (1993), although, to the best of our knowledge, they have not been extended to the case where the operator is estimated.

1.5.1 Setup 1

Samples of size $N = 1000$ are generated from the model:

$$f_{XW}(x,w) = 2C_f \sum_{i=1}^{\infty} (-1)^{i+1} i^{-b/2} \sin(i\pi x) \sin(i\pi w),$$
$$\varphi(x) = \sqrt{2} \sum_{i=1}^{\infty} (-1)^{i+1} i^{-a} \sin(i\pi x),$$
$$Y = \mathbb{E}\left(\varphi(X)|W = w\right) + V,$$

where C_f is a normalizing constant and $V \sim N(0, 0.1^2)$. The slice sampling method of Neal (2003) is used in order to simulate values of X and W from the joint pdf f_{XW}. The infinite series were truncated at $j = 100$ for computational purposes.

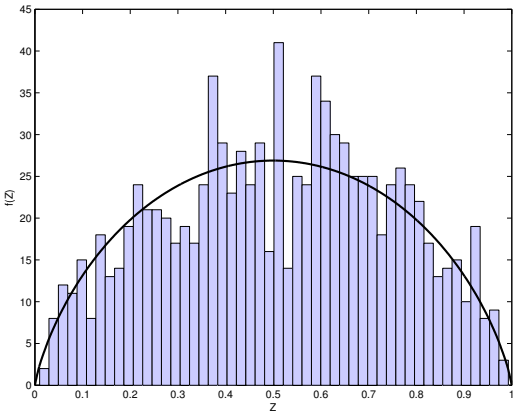


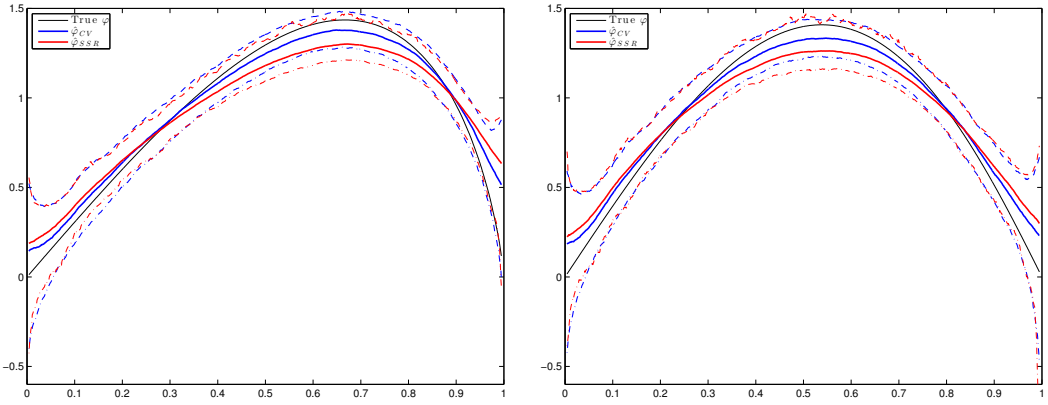
Figure 1.4: Marginal density of X and W, with one draw using slice sampling.

Notice that the values of a and b respectively control the smoothness of the function φ, through its Fourier coefficients, and the decay of the eigenvalues λ_i. The *source condition* can therefore be expressed in terms of the parameters a and b. As a matter of fact, the following inequality holds:

$$\beta < \frac{1}{b}\left(a - \frac{1}{2}\right)$$

with $a > 1/2$ and $b > 1$ (see Hall and Horowitz, 2005; Darolles et al., 2011*a*).[10]

Two different simulation schemes are run. In the former, a and b are taken equal to 2. In the latter, $a = 4$ and $b = 2$. In both cases, X and W have the same marginal distribution, which is depicted in figure (1.4). Note that in the former numerical study $\beta < 0.75$, while in the latter $\beta < 1.75$. 1000 paths of the endogenous variable X, the instrument W and the error V are simulated.



(a) $a = b = 2$ (b) $a = 4$, $b = 2$

Figure 1.5: Estimation of the function φ using the CV and the SSR criterion respectively, with penalization of the function.

Results of the numerical study are reported in Figure (1.5). The kernel Tikhonov estimator that uses the CV function to compute the data-driven value of α (blue line) is plotted against the same estimator that uses instead the SSR function of Fève and Florens (2010) (red line), and the true function φ (black line). It is evident from the figure that φ^{CV} estimator outperforms the φ^{SSR} estimator in terms of fitting. This implies a lower bias and a higher variance of the former estimator. The simulated pointwise 90% confidence intervals for the two estimators are also plotted. It is clear from the figures that our CV criterion guarantees a better coverage of the true

[10] In Hall and Horowitz (2005) the additional condition $a - 1/2 \leq b < 2a$ is imposed. However, this condition is necessary to prove minimax rate for the kernel Tikhonov estimator, which is not the goal of the present chapter.

function φ.

		Mean	Median	St.Dev	Min	Max
$a=2$	α^{CV}	0.01702	0.01661	0.00350	0.00819	0.03570
	α^{SSR}	0.04938	0.05133	0.02494	0.00003	0.13456
	α^{GCV}	0.00011	0.00011	0.00001	0.00007	0.00017
	α^{OPT}	0.01995	0.01936	0.00611	0.00632	0.04876
$a=4$	α^{CV}	0.01837	0.01785	0.00359	0.01015	0.03858
	α^{SSR}	0.04468	0.04705	0.02483	0.00003	0.12602
	α^{GCV}	0.00011	0.00011	0.00001	0.00007	0.00017
	α^{OPT}	0.02030	0.02005	0.00556	0.00645	0.04512

Table 1.1: Summary statistics for the regularization parameter, with penalization of the function.

Another comparison between the two vectors of alphas is reported in table (1.1). Summary statistics for the vector of α^{CV}, α^{SSR}, α^{GCV} and α^{OPT} are listed. Beside the evident fact that α^{CV} has a lower mean than α^{SSR}, its variance is also significantly smaller. Therefore, the regularization parameter chosen using the CV criterion is less sensitive to sample selection. Also, the average value of α^{CV} is closer to the average value of the optimal α, and their distributions overlaps. By contrast, the distribution of α^{SSR} is clearly shifted to the right, compared to the one of α^{OPT}. The α^{GCV} instead selects a too small value of the regularization parameter and its distribution remains very far from α^{OPT}. Notice that this result is consistent with the existing literature in statistical inverse problem (see Lukas, 1993, 2006).

Finally, in Figure (1.6), we plot the shape of the objective functions for CV, blue line, and SSR, red line, both for the case $a = b = 2$ (left panel) and the case $a = 4$ and $b = 2$ (right panel). The figure on the right panel confirms our theoretical intuition for the SSR criterion. Although the function admits in this case a minimum, it tends already to be very flat for a wide range of value around that point. By contrast, the CV function seems to behave better as the local minimum is isolated from the rest of the points (the function *spikes* for values below and above the minimum).

An equivalent comparative simulations exercise can be carried on in the case of the penalization by derivatives. In particular, following the notations in the previous section, $s = 1$, so that penalization is on the first derivative of the function, i.e. $B = TL^{-1}$. The framework is slightly different than in the baseline case. For the estimation of the conditional expectation operator T, one proceeds as before by regressing the dependent variable Y, on the instrument W. The integral operator

L^{-1} is approximated using the trapezoidal rule.[11] The main challenge in this case is to obtain the adjoint operator B^*. Define a function λ, such that, $\lambda^{'} \in \mathbb{L}_w^2$; f_X and S_X, the pdf and the survivor function of X, respectively; f_W, the pdf of W; and, finally,

$$S(u,w) = -\frac{\partial}{\partial w}\mathbb{P}\left(X \geq u, W \geq w\right).$$

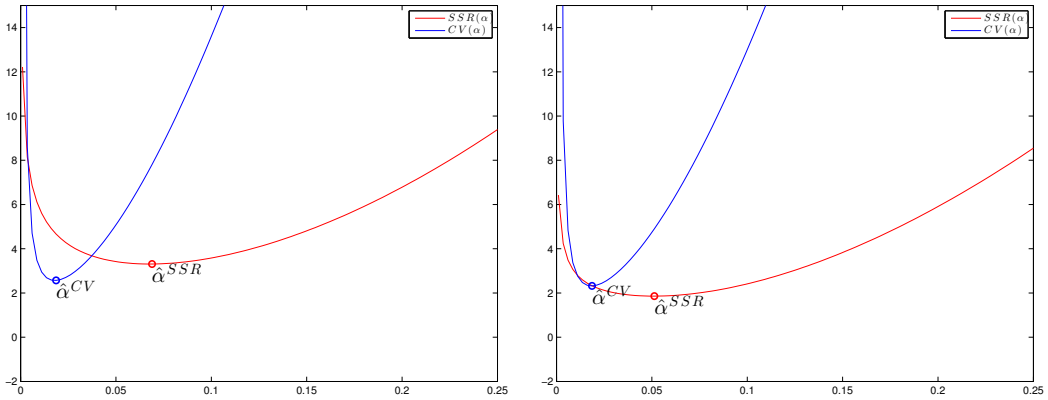
Then Florens and Racine (2012) show, in the case of Landweber-Fridman regularization, that the adjoint operator, B^*, is such that:

$$\left(B^*\lambda\right)(u) = \frac{1}{f_X(u)}\int \lambda(w)\left(S(u,w) - S_X(u)f_W(w)\right)dw.$$

Also, the function φ is restricted to have mean 0 in order to be identified. As a matter of fact, the first order differential operator is one-to-one only if it is restricted to this specific subset of functions. This is extremely important for the implementation of the Landweber-Fridman regularization, as the function of interest needs to be recentered at each iteration, in order to obtain a convergent scheme.

In the application to Tikhonov regularization, the estimation is extremely simplified. Notice that the identifying sample moment restriction for the estimation of φ is written as:

$$\hat{B}^*\hat{B}\varphi^{'} = \hat{B}^*\hat{r}.$$



(a) $a = b = 2$ (b) $a = 4$, $b = 2$

Figure 1.6: Objective functions: CV, blue line, and SSR, red line

[11] For a detailed description of the implementation the reader is referred to Florens and Racine (2012) and to chapter 3.

Therefore, *a fortiori*, the mean of the function φ is restricted to be equal the mean of Y (up to the regularization bias induced by the estimation). Thus, one can obtain B^* simply as:

$$\left(B^*\lambda\right)(u) = \frac{1}{f_X(u)} \int \lambda(w) S(u,w) dw.$$

This can be approximated by the matrix of survivor weights of X, multiplied by the inverse of a suitable nonparametric estimator of the pdf of X. Denote by $K_h(\cdot)$ a positive and symmetric kernel with (possibly) unbounded support, and define:

$$\mathcal{K}_h(x) = \int_{-\infty}^{x} K_h(u) du.$$

For each possible realization of the random variable X. The survivor matrix of weights is defined, for a sample of size N, as:

$$\hat{S}_x = \left[1 - \mathcal{K}_h\left(\frac{x - x_i}{h_x}\right)\right]_{i=1}^{N},$$

where the bandwidth h_x is chosen, in our case, using maximum likelihood cross-validation, and:

$$\hat{B}^* = diag\left(\hat{f}_x^{-1}\right)\hat{S}_x,$$

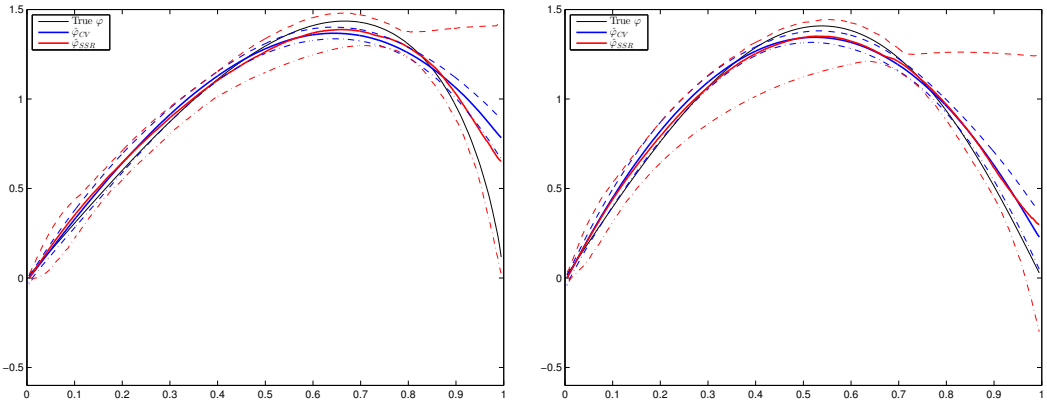
where $diag\left(\hat{f}_x^{-1}\right)$ is a diagonal matrix, whose elements are the inverse of the estimated density at each sample point. Hence, the Tikhonov regularized estimator with penalized first derivative is defined as:

$$\hat{\varphi}^{\alpha} = L^{-1}\hat{\varphi}'^{\alpha} = L^{-1}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\hat{r}.$$

The SSR criterion of Fève and Florens (2010) has been extended to this case by Fève and Florens (2013). They generalize the SSR criterion by taking as penalizing term the squared norm of the estimator $\hat{\varphi}^{\alpha}_{(2)}$. That is:

$$SSR(\alpha) = \|\hat{\varphi}^{\alpha}_{(2)}\|^2 \|\hat{T}\hat{\varphi}^{\alpha}_{(2)} - \hat{r}\|^2.$$

The implementation of the CV criterion remains instead unchanged. Results of this numerical simulations are reported in figure (1.7), both for the case where $a = b = 2$ (left panel), and for the case $a = 4$ and $b = 2$ (right panel).



(a) $a = b = 2$ (b) $a = 4$, $b = 2$

Figure 1.7: Estimation of the function φ using the CV and the SSR criterion respectively, with penalization of the first derivative of the function.

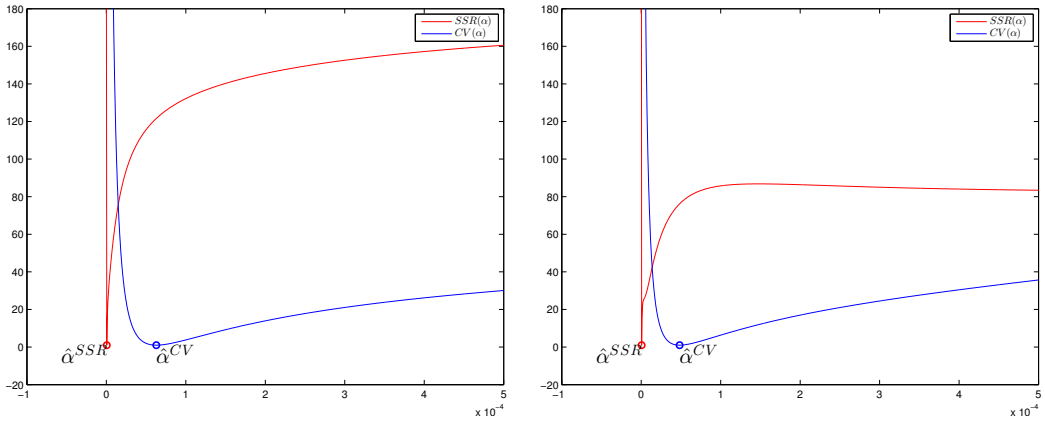
In this case the difference between the two estimators is not evident. The CV estimator has a lower variance and a larger bias compared to the SSR estimator. However the latter often seems to *under-regularize* with respect to the optimal solution, although its median value is very close to the true regression function. Accordingly, the 90% simulated confidence interval of the SSR estimator covers the true function much better than the corresponding confidence interval for the CV estimator.

		Mean	Median	St.Dev	Min	Max
$a = 2$	α^{CV}	0.0000576	0.0000550	0.0000225	0.0000101	0.0001515
	α^{SSR}	0.0013870	0.0000034	0.0041475	0.0000000	0.0580061
	α^{OPT}	0.0000025	0.0000009	0.0000049	0.0000000	0.0000391
$a = 4$	α^{CV}	0.0000464	0.0000445	0.0000183	0.0000060	0.0001134
	α^{SSR}	0.0021059	0.0000049	0.0050825	0.0000000	0.0340454
	α^{OPT}	0.0000056	0.0000037	0.0000069	0.0000000	0.0000495

Table 1.2: Summary statistics for the regularization parameter, with penalization of the first derivative of the function.

Table (1.2) reports the summary statistics for the two vectors of alphas. The comparison between those confirms our intuition. The α^{CV} has a smaller mean than α^{SSR}, although its median is substantially larger. This suggests again that α^{CV} is much more robust to sample selection. For some particular sample, in fact, the SSR criterion tends to pick values of α that are far away from the mean. These comparative results have to be interpreted with care, as the properties of the SSR criterion are not well established in this case. However, α^{CV} performs well also in comparison to α^{OPT}, despite the fact that its distribution is slightly shifted to the right.

Finally, in Figure (1.8), we plot the shape of the objective functions for CV, blue line, and SSR, red line, both for the case $a = b = 2$ (left panel) and the case $a = 4$ and $b = 2$ (right panel). While the form of the CV function is similar to the baseline case, the SSR criterion displays a very peculiar shape. The curve has a kink which identifies the local minimum.[12]



(a) $a = b = 2$ (b) $a = 4$, $b = 2$

Figure 1.8: Objective functions: CV, blue line, and SSR, red line

1.5.2 Setup 2

We simulate samples of size $N = 1000$ from the following data generating process:

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

$$X = 0.1W_1 + 0.1W_2 + V$$

$$U = -0.5V + \varepsilon$$

where:

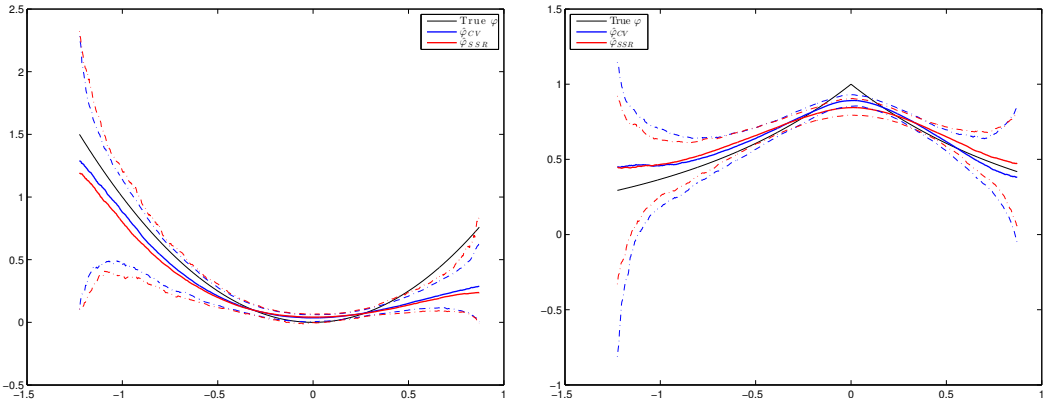
$$\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \right)$$

$$V|w_1, w_2 \sim \mathcal{N}\left(0, (0.27\exp(-0.1w_1 - 0.1w_2))^2\right)$$

$$\varepsilon \sim \mathcal{N}\left(0, (0.05)^2\right)$$

[12] This behavior may be caused by some numerical error in evaluating the criterion. However, even taking a finer grid around the minimum does not modify the shape of the objective function.

We choose two specification for the regression function: $\varphi(x) = x^2$, which corresponds to a high order of regularity; and $\varphi(x) = \exp(-|x|)$, which corresponds to a function that is not very regular and has a kink at zero. Moreover, since the latter is not everywhere differentiable, in this setup we only consider the estimation of φ by direct penalization.



(a) $a = b = 2$ (b) $a = 4$, $b = 2$

Figure 1.9: Estimation of the function φ using the CV and the SSR criterion respectively, with penalization of the function.

The main result of the estimation is plotted in Figure (1.9). The results are widely consistent with those obtained above. The CV estimator confirms to approximate the true function better than the SSR estimator.

Finally, Table (1.3) outlines the comparison between the distribution of α^{CV}, α^{SSR} and α^{GCV} and the distribution of the optimal α. The results reported are largely consistent with those outlined above. The GCV criterion always picks values of α that is too small, independently from the smoothness properties of the function under study.

		Mean	Median	St.Dev	Min	Max
$\varphi(x) = x^2$	α^{CV}	0.03707	0.02367	0.03924	0.00007	0.46136
	α^{SSR}	0.05930	0.03080	0.07141	0.00003	0.47414
	α^{GCV}	0.00026	0.00025	0.00007	0.00007	0.00045
	α^{OPT}	0.00997	0.00773	0.00770	0.00007	0.06359
$\varphi(x) = \exp(-\lvert x\rvert)$	α^{CV}	0.01108	0.01031	0.00393	0.00366	0.03676
	α^{SSR}	0.03216	0.02816	0.01989	0.00003	0.11374
	α^{GCV}	0.00007	0.00007	0.00000	0.00007	0.00011
	α^{OPT}	0.00771	0.00714	0.00402	0.00017	0.02692

Table 1.3: Summary statistics for the regularization parameter, with penalization of the function.

Furthermore, the comparison between the CV and the SSR criteria sheds more light on some of

our theoretical results. When the function is very regular, i.e. $\varphi(x) = x^2$, the SSR approaches the properties of CV, although α^{CV} still maintains a lower mean and standard deviation. Both estimators take values greater than α^{OPT}. However, when the function is not very regular, i.e. $\varphi(x) = \exp(-|x|)$, then α^{CV} clearly improves upon α^{SSR} and its summary statistics are also much closer to those of α^{OPT}. This is consistent with the fact that, in finite samples and for a low order of regularity β, the SSR criterion selects a large value of the regularization parameter.

1.6 An Empirical Application: Estimation of the Engel Curve

The estimation of the Engel Curve has been used by many authors as a motivating example for studying the properties of nonparametric instrumental regressions and the data driven choice of the regularization parameter (see, e.g., Blundell et al., 2007; Horowitz, 2011, 2012).

As it has already been pointed out in the introduction, the estimation of the Engel curve boils down to find the structural relation between the total household expenditure and the budget share allocated to a given commodity. As total expenditure is likely to be jointly determined with the its share for individual commodities, the explanatory variable in this problem is endogenous. However, it can be instrumented by the gross household income.

In this section, the separable model presented in (3.2.1a) is used to estimate the structural shape of the Engel curve, where Y is the budget share for each individual commodity; X is the natural logarithm of total expenditure; and W is the natural logarithm of gross total income. That is:

$$Y = \varphi(X) + U, \tag{1.6.1}$$

$$\mathbb{E}(U|W) = 0. \tag{1.6.2}$$

This example seems particularly suited to discuss the properties and the implementation of non-parametric instrumental regressions for several reasons. First, it restricts the analysis to the very simple case of a single instrument and a single endogenous variable. Second, both the former and the latter are continuously distributed and, therefore, satisfy the identification conditions. Finally,

economic theory can provide guidance about the shape of the curve, depending on the type of good under consideration, which allows the researcher to verify the consistency of the results obtained.

As the studies cited above, the present chapter focuses on the estimation of the Engel curve using data from the 1995 wave of UK Family Expenditure Survey. The database contains 1655 observations about households consisting of married couples with an employed head-of-household between the ages of 20 and 55 years.[13] This chapter focuses on the estimation of the Engel curve for three categories of nondurables and services: food, fuel, and leisure. Table (1.4) reports some summary statistics for these data.

	Mean	Median	St.Dev	Min	Max
Budget share food	0.2074	0.1959	0.0971	0.0014	0.6867
Budget share fuel	0.0651	0.0588	0.0373	0.0000	0.3831
Budget share leisure	0.1297	0.0822	0.1343	0.0000	0.8872
Log Total Expenditure	5.4215	5.4019	0.4494	3.6090	7.4287
Log Gross Income	5.8581	5.8568	0.5381	2.1972	8.0893

Table 1.4: Summary statistics UK Family Expenditure Survey.

In order to show the flexibility of the approach of this chapter, the application is presented under several nonparametric models for the estimation of the conditional expectation operators. In particular, both local constant and local linear kernels and cubic B-spline are analyzed here. Moreover, the direct estimation of the first derivative of the curve is also considered using local constant kernels. For each estimator, the smoothing parameters, i.e. either the bandwidths or the number of knots, are computed using least square cross-validation (Li and Racine, 2007). Bootstrap confidence intervals are obtained using the methodology presented in Chapter 3.

For comparison, we consider both the simple estimation of model (1.6.1) by Two Stage Least Squares (TSLS) and the nonparametric regression of Y on X. The former serves as a benchmark for a specification that overlooks nonlinearites completely. The TSLS model is in fact defined as

[13] Hoderlein and Holzmann (2011) point out a drawback of this model. Its additive separable structure may not capture unobserved preference heterogeneity in the population. Therefore it may impose restrictions on the structural shape of the Engel curve that cannot be justified by the economic theory. This suggests using this model specification with care in empirical applications.

follows:

$$Y = \zeta_0 + \zeta_1 X + U,$$

$$X = \delta_0 + \delta_1 W + V,$$

which imposes linear restrictions, not only on the regression function φ, but also on the auxiliary regression model that relates X to W. By contrast, notice that the nonparametric model, as defined in (1.6.1) is completely silent about the functional form relating the endogenous variable to the instruments. Therefore, although our results may not be too far from the fully linear specification of the $TSLS$ model, the two models would still not be equivalent.

The direct nonparametric regression of Y on X is instead used to compare our main results with a specification that accounts for a possibly nonlinear regression function, but it does not consider endogeneity. Furthermore, in the spirit of Blundell and Horowitz (2007), if the function obtained with the simple nonparametric regression - under the assumption of exogeneity - is fully contained inside the confidence bands of the nonparametric estimator under endogeneity, it is possible to conclude that the explanatory variable is indeed exogenous.[14]

Table (1.5) reports the estimators of the fully parametric TSLS model (standard errors in brackets). All coefficients in the three specifications are largely significant.

	Food	Fuel	Leisure
Intercept	0.5693	0.2668	-0.6243
	(0.0501)	(0.0186)	(0.0697)
Log Expenditure	-0.0668	-0.0372	0.1391
	(0.0092)	(0.0034)	(0.0129)

Table 1.5: Results of TSLS regressions. Standard Errors in brackets.

In order to compare this outcome with the nonparametric regression model, we proceed with a graphical analysis. Figures (3.18), (3.19) and (3.20) overlay the results of the parametric TSLS with the nonparametric IV estimator, and the nonparametric regression that assumes exogeneity, for food, fuel and leisure respectively. The black line is the nonparametric IV estimator; the dashed black lines are the 95% confidence intervals obtained using wild bootstrap; the blue line gives the nonparametric regression line of Y on Z; and, finally, the red line corresponds to the fitted TSLS

[14]Programming has been conducted in MatLab and codes are available from the author upon request.

values.

Results are similar to those obtained in related papers (see Blundell et al., 2007; Hoderlein and Holzmann, 2011). It is particularly interesting to notice that the shape of the Engel curve for the three goods and services considered is extremely different. Food is a necessity good, so that the Engel curve is downward sloping, i.e., the share of total expenditure devoted to food becomes less important as total expenditure increases. Fuel seems to have an irregular pattern as its relative weight on total expenditure is initially decreasing and then roughly constant towards higher total expenditure. Finally, leisure is, as expected, a luxury service as the Engel curve is nondecreasing in total expenditure.

Moreover, while all specifications do not look too different from the fully linear TSLS model, *locally*, the marginal effect ought not to be the same as the slope of the two estimators is different. Especially the nonparametric estimator of the curve for food seems to indicate that the linear specification could be supported by the data. However, the fitted TSLS values are never fully contained in the 95% bootstrap confidence bands.

Another important aspect to notice is that, in general, the nonparametric IV estimator has a different slope compared to the direct nonparametric regression of Y on X. As a matter of fact, the simple curve obtained from the nonparametric regression of the share of expenditure on food, fuel and leisure and total expenditure is often not included in the 95% bootstrap confidence interval. Only the specification for leisure seems to point towards the full exogeneity of the predictor. This could be due to expenditure for leisure not being systematically planned by the household.

However, for the scope of the present chapter, a more crucial result is that nonparametric instrumental regressions with data-driven choice of the regularization parameter yield systematically consistent results.

A final assessment of the performance of this estimator is reported in figure (1.13), (1.14) and (1.15). For food, fuel and leisure, these figures report, on the right panel, the direct estimator of the first derivative of the Engel curve, obtained using local constant kernels; and on the left panel, the estimator of the shape of the Engel curve, obtained as the integral of its first derivative. The nonparametric estimator of the derivative of the regression function when X is treated as

exogenous is also reported for completeness.[15]

Results are consistent with those previously discussed. The estimators of each derivative are roughly constant, which indicates the Engel curves to be linearly decreasing (increasing). However, the estimation for leisure goods seems to suggest an increasing marginal effect, and therefore a quadratic shape of the Engel curve. Therefore, as the log of total expenditure increases, households would devote an growing share of it to leisure goods. Moreover, this increasing effect is larger towards higher expenditures.

The estimation of marginal effects also allows for a further comparison between the fully linear model and the nonparametric instrumental variable specification. Although, the TSLS estimator may seem reasonably close to nonparametric IV, the estimation of marginal effects shows clearly that this is not the case. The nonparametric estimator gives, in general, richer information, while the linear model either under- or over-estimate marginal effects, especially in areas where we have more information coming from the data.

1.7 Conclusions

Leave-one-out cross-validation (CV) is often used and advocated as a simple data driven criterion to choose tuning parameters in nonparametric models. However, this chapter is the first one to provide theoretical results about the properties of the regularization parameter chosen by CV in nonparametric instrumental regressions, when the Tikhonov scheme is used in order to estimate the function of interest.

The chapter explores first the case where the $\mathbb{L}^2$ penalization is directly on the regression function. It is shown that the cross-validation criterion is bounded in probability. This bound delivers a regularization constant which possesses an optimal rate of convergence to zero, depending on the value of the regularity index β. Namely, we show that, when β is higher than 1, the CV criterion tends to under-regularize with respect to an optimally chosen regularization parameter.

[15] As already pointed out in related work (Florens and Racine, 2012), the two are not directly comparable. As a matter of fact, in standard nonparametric regression, the estimation of the nonparametric derivative is *self-consistent*, i.e. it is obtained as derivative of the conditional mean estimator. By contrast, in the penalized approach studied in this chapter, one obtains directly the estimator of the derivative, and the regression curve is computed as the integral of the latter.

Consequently, we have explored ways to improve this result, so that the properties of our data-driven α could be extended to cases in which $\beta > 1$.

A possible way to achieve this goal is to consider an iterated Tikhonov approach. Iterating the Tikhonov regularization allows one to increase its *qualification* and to take advantage of higher degrees of smoothness. However, in our case, this solution does not seem viable. As a matter of fact, the CV criterion finds the regularization parameter that minimizes the prediction error of our estimator. Therefore, if an iterated approach is used to determine its minimum, it must also be used for estimation.

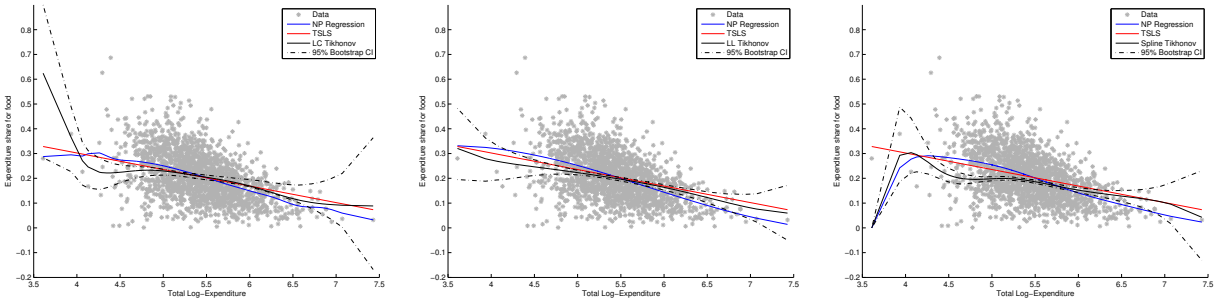
An alternative way is to impose the penalization on the derivatives of the function of interest, rather than on the function itself. Penalizing derivatives of the function is an indirect way of increasing the qualification of the Tikhonov approach, as any derivative is necessarily less smooth than its corresponding function. Moreover, this approach seems especially relevant for empirical studies, as marginal effects may be the main object of interest for the researcher. Therefore, the second part of the chapter establishes similar results for the estimation of derivatives. We show that in this second case the rate of convergence of the tuning constant chosen by cross-validation is equivalent to the optimal one.

An extensive simulation study shows that CV generally outperforms existing criteria for the selection of the regularization parameter. In particular, it seems to be more stable to sample selection - in our simulation studies, the regularization parameter chosen by CV has a lower variance than its current alternatives.

Finally, an empirical application to the estimation of the Engel curve in a sample of UK households shows that the cross-validation devised here is quite flexible, and it can be applied when conditional expectation operators are estimated using any available nonparametric technique, such as local polynomial or B-splines. It can therefore accommodate several tastes in the use of nonparametric methods. Consequently, this work goes in the direction of providing a stable and functioning data-driven methodology that can allow an easier implementation of nonparametric instrumental regressions.

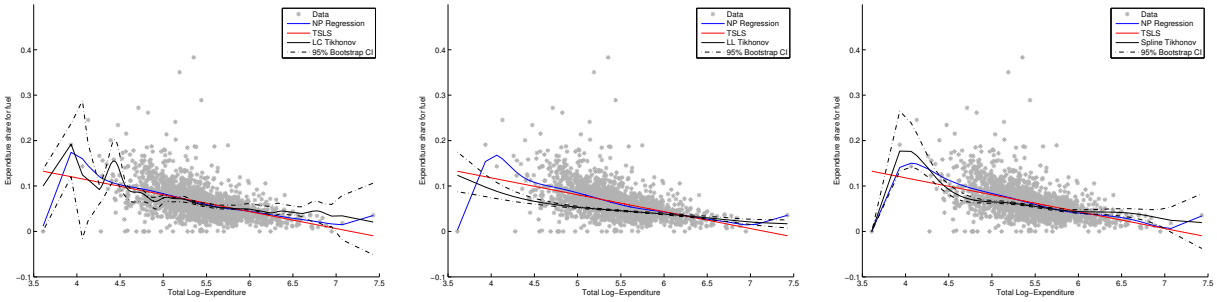
The theoretical results of this work can be extended and improved in several ways. First of all, the rate optimality result should be strengthened to the choice of a particular loss function and

an oracle type inequality should be established. Second, while the theoretical result hinges on the selection of one smoothing parameter, in practical applications, there are at least two smoothing parameters to be selected. This can impact directly the rate of convergence of the estimator, as recently shown by Fève and Florens (2013). Moreover, as already outlined in the introduction, it would be interesting to study the properties of the same criterion in other nonparametric problems in econometrics, where the object of interest is a solution of an *ill-posed* inverse problem.



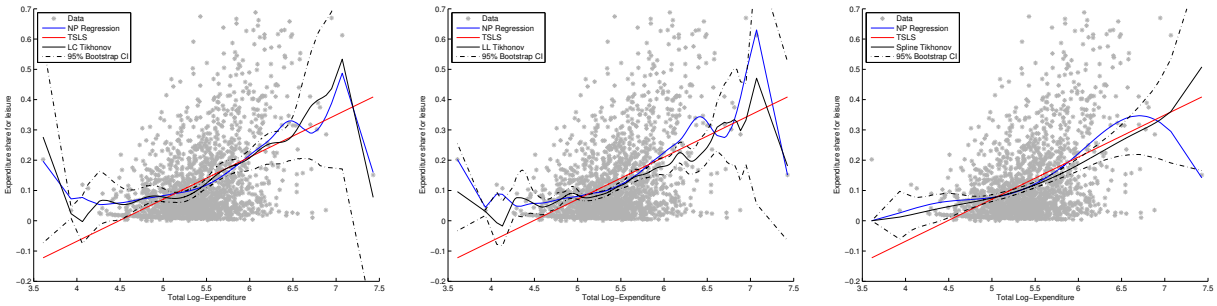
(a) Local Constant (b) Local Linear (c) Cubic B-splines

Figure 1.10: Engel Curve for food



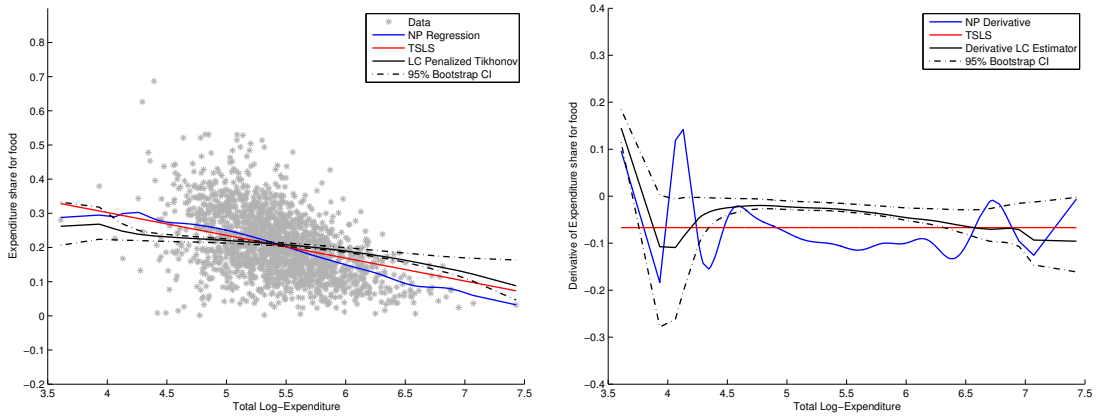
(a) Local Constant (b) Local Linear (c) Cubic B-splines

Figure 1.11: Engel Curve for fuel



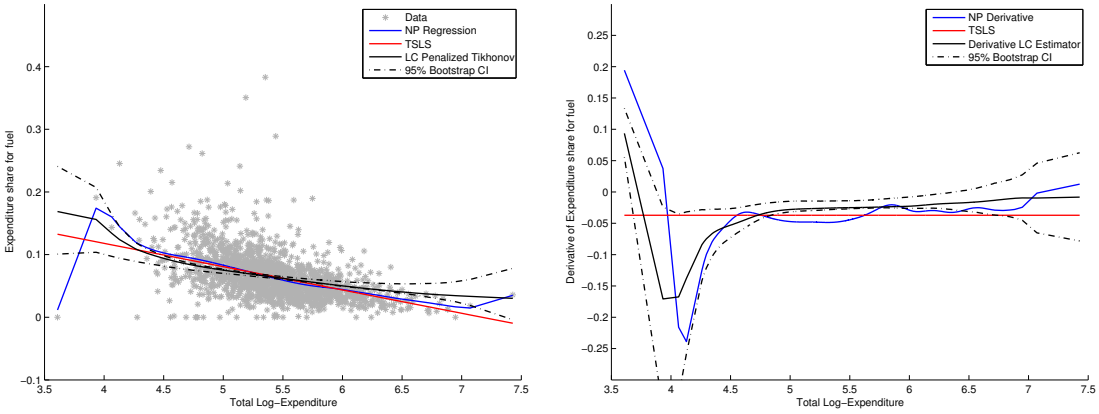
(a) Local Constant (b) Local Linear (c) Cubic B-splines

Figure 1.12: Engel Curve for leisure



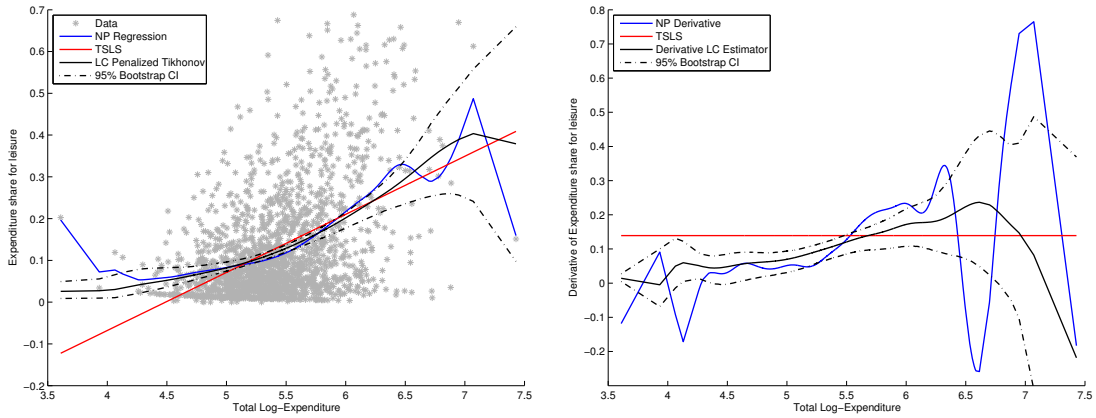
(a) Local Constant Penalized (b) First derivative

Figure 1.13: Engel Curve for food and its derivative



(a) Local Constant Penalized (b) First derivative

Figure 1.14: Engel Curve for fuel and its derivative



(a) Local Constant Penalized (b) First derivative

Figure 1.15: Engel Curve for leisure and its derivative

1.8 Appendix A - Numerical Range of a Bounded Operator

In this section we gather the main definitions and properties of the numerical range of a bounded operator. The results of this section are given without proof. We refer the interested reader to Skoufranis (2012), for a more detailed presentation of the material discussed in this section.

Define by $\mathcal{H}$ a Hilbert space and by $\mathcal{B}(\mathcal{H})$ the class of bounded operators on $\mathcal{H}$. Finally, denote by T an element in $\mathcal{B}(\mathcal{H})$.

Definition 1.8.1. Let $T \in \mathcal{B}(\mathcal{H})$. The numerical range of T, denoted by $N(T)$, is the non-empty set:

$$N(T) := \{\langle T\zeta, \zeta\rangle | \zeta \in \mathcal{H}, \|\zeta\| = 1\}$$

Proposition 1.8.2. *Let $T \in \mathcal{B}(\mathcal{H})$. Then:*

(i) $N(T)$ contains all the eigenvalues of T.

(ii) $N(T)$ is contained in the closed disk of radius $\|T\|$ around the origin.

The following theorems discuss the relation between the numerical range and the spectrum of an operator, denoted here by $\lambda(\cdot)$.

In particular, denote by $\overline{N(T)}$, the closure of the numerical range of T.

Proposition 1.8.3. *Let $T \in \mathcal{B}(\mathcal{H})$. Then $\lambda(T) \subset \overline{N(T)}$.*

Proposition 1.8.4. *Let $T \in \mathcal{B}(\mathcal{H})$ be a normal operator. Then $\overline{N(T)} = conv(\lambda(T))$, where $conv(\lambda(T))$ is the convex hull generated by the spectrum of T.*

The result of the last proposition is the one used for the proof of Theorem (1.3.4).

1.9 Appendix B - Proofs

The following assumption is used in some proofs below.

Assumption 4 (Darolles et al. 2011*a*)**.**

$$\|\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi\|^2 = O_P\left(\frac{1}{N} + h_N^{2\rho}\right)$$

The motivation of this assumption is to avoid the *curse of dimensionality*, by integrating out the instrumental variable. In this way, it is possible to obtain parametric rates of convergence for the variance component. Interested readers are referred to Darolles et al. (2011*a*) for details and a formal proof.

1.9.1 Proof of Corollary (1.3.2)

(i) If $\beta \geq 1$, the second term of the upper bound in (1.3.1) is independent of α. Therefore, the optimal choice of the regularization parameter is obtained by making the variance and the bias term converging at the same speed, which trivially gives the result.

(ii) If $\beta < 1$ and

$$\gamma < 1 - \frac{p+q}{2\rho}\gamma,$$

this implies that:

$$\gamma < \frac{2\rho}{2\rho + p + q},$$

and the second term converges at the speed $N^{\gamma}\alpha_N^{1-\beta}$. Therefore, upon the assumption that $N^{\gamma}\alpha_N^2 \to \infty$, the second term converges to infinity faster, and the bias-variance trade-off gives the rate of convergence for α_N.

(iii) If $\beta < 1$ and

$$\gamma \geq 1 - \frac{p+q}{2\rho}\gamma,$$

this implies that:

$$\gamma \geq \frac{2\rho}{2\rho + p + q},$$

Moreover, to obtain convergence of the MSE to 0, the additional condition:

$$1 - \frac{p+q}{2\rho}\gamma > 0$$

gives the upper bound for γ:

$$\gamma < \frac{2\rho}{p+q}$$

However, upon the restrictions on the rate of convergence of the bandwidth, it is not clear if the second term still converges faster to infinity than the first term. Compute the corresponding bias-variance trade-off for the two terms:

$$\begin{aligned} \frac{1}{N^{\gamma}\alpha_N^2} \approx \alpha_N^{\beta} \quad &\rightarrow \quad \alpha_N \approx N^{-\frac{\gamma}{\beta+2}} \\ \frac{1}{N^{1-\frac{p+q}{2\rho}\gamma}\alpha_N^{1-\beta}} \approx \alpha_N^{\beta} \quad &\rightarrow \quad \alpha_N \approx N^{-1+\frac{p+q}{2\rho}\gamma} \end{aligned}$$

Then, by equalizing the two rates of convergences, one has:

$$\gamma = \frac{2\rho(\beta+2)}{(p+q)(\beta+2)+2\rho}$$

Hence, for γ lower than this threshold, the rate of convergence of the first term is lower than the one of the second term. Otherwise, the rate of the second term is lower than the first term.

1.9.2 Proof of Lemma (1.3.3)

The proof easily follows from the results in Darolles et al. (2011*a*). Consider the estimated conditional expectation of the residuals on the space spanned by the instruments:

$$\hat{T}\hat{\varphi}_{(2)}^{\alpha_N} - \hat{r} = \hat{T}\hat{\varphi}_{(2)}^{\alpha_N} - T\varphi + T\varphi - \hat{r}$$

The last term on the right hand side is the nonparametric estimation error. Therefore, one has:

$$\|T\varphi - \hat{r}\|^2 = \| \left(\hat{T} - T\right) y\|^2 = O_P\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right)$$

Now focus on the first term. Define:

$$M = \left[I + \alpha_N \left(\alpha_N I + T^*T\right)^{-1}\right]$$

Therefore:

$$\begin{aligned}
\hat{T}\hat{\varphi}_{(2)}^{\alpha_N} - T\varphi &= \hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\hat{M}\hat{T}^*\hat{r} - T\varphi \\
&= \hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\hat{M}\hat{T}^*\hat{r} - T\left(\alpha_N I + T^*T\right)^{-1}MT^*T\varphi \\
&+ T\left(\alpha_N I + T^*T\right)^{-1}MT^*T\varphi - T\varphi \\
&= A_1 + A_2
\end{aligned}$$

The second term B is the regularization bias. It can be bounded as follows (Engl et al., 2000):

$$\|A_2\|^2 = O_P\left(\alpha_N^{\min(\beta+1,4)}\right)$$

since a second order iteration for the Tikhonov estimator is considered here. Term A can be finally split into two components:

$$\begin{aligned}
A_1 &= \hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\hat{M}\hat{T}^*\hat{r} - \hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\hat{M}\hat{T}^*\hat{T}\varphi \\
&+ \hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\hat{M}\hat{T}^*\hat{T}\varphi - T\left(\alpha_N I + T^*T\right)^{-1}LT^*T\varphi \\
&= A_{11} + A_{12}
\end{aligned}$$

Since:

$$\|\hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\hat{M}\|^2 = O_P\left(\alpha_N^{-1}\right)$$

from Assumption (4), it follows that:

$$\|A_{11}\|^2 = O_P\left[\frac{1}{\alpha_N}\left(\frac{1}{N} + h^{2\rho}\right)\right]$$

Finally, using some algebra, it is possible to show that:

$$A_{12} = -\alpha_N^2\left[\hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-2} - T\left(\alpha_N I + T^*T\right)^{-2}\right]\varphi$$

which can be further split as follows:

$$\begin{aligned}
A_{12} =&\alpha_N^2 \hat{T}\left[\left(\alpha_N I+\hat{T}^*\hat{T}\right)^{-2}-\left(\alpha_N I+T^*T\right)^{-2}\right]\varphi+\alpha_N^2\left(\hat{T}-T\right)\left(\alpha_N I+T^*T\right)^{-2}\varphi \\
=&\alpha_N^3 \hat{T}\left(\alpha_N I+\hat{T}^*\hat{T}\right)^{-2}\left(\hat{T}^*\hat{T}-T^*T\right)\left(\alpha_N I+T^*T\right)^{-2}\varphi && (A_{12a}) \\
&+\alpha_N^2 \hat{T}\left(\alpha_N I+\hat{T}^*\hat{T}\right)^{-2}\hat{T}^*\hat{T}\left(\hat{T}^*\hat{T}-T^*T\right)\left(\alpha_N I+T^*T\right)^{-2}\varphi && (A_{12b}) \\
&+\alpha_N^2 \hat{T}\left(\alpha_N I+\hat{T}^*\hat{T}\right)^{-2}\left(\hat{T}^*\hat{T}-T^*T\right)T^*T\left(\alpha_N I+T^*T\right)^{-2}\varphi && (A_{12c}) \\
&+\alpha_N^2\left(\hat{T}-T\right)\left(\alpha_N I+T^*T\right)^{-2}\varphi && (A_{12d})
\end{aligned}$$

The proof makes use of the following facts:

$$\begin{aligned}
\| \left(\alpha_N I+\hat{T}^*\hat{T}\right)^{-1} \|^2 &= O_P\left(\frac{1}{\alpha_N^2}\right) \\
\| \left(\alpha_N I+\hat{T}^*\hat{T}\right)^{-1}\hat{T}^* \|^2 &= O_P\left(\frac{1}{\alpha_N}\right) \\
\|\hat{T}\left(\alpha_N I+\hat{T}^*\hat{T}\right)^{-1}\hat{T}^* \|^2 &= O_P\left(1\right) \\
\|\alpha_N\left(\alpha_N I+T^*T\right)^{-1}\varphi\|^2 &= O_P\left(\alpha_N^{min(\beta,2)}\right) \\
\|\alpha_N T\left(\alpha_N I+T^*T\right)^{-1}\varphi\|^2 &= O_P\left(\alpha_N^{min(\beta+1,2)}\right)
\end{aligned}$$

Furthermore, notice that:

$$\hat{T}^*\hat{T}-T^*T=\hat{T}^*\left(\hat{T}-T\right)-\left(\hat{T}^*-T^*\right)T$$

This implies that:

$$\begin{aligned}
\|A_{12a}\|^2 \leq&\|\alpha_N^2 \hat{T}\left(\alpha_N I+\hat{T}^*\hat{T}\right)^{-2}\hat{T}^*\left(\hat{T}-T\right)\alpha_N\left(\alpha_N I+T^*T\right)^{-2}\varphi\|^2 \\
&+\|\alpha_N^2 \hat{T}\left(\alpha_N I+\hat{T}^*\hat{T}\right)^{-2}\left(\hat{T}^*-T^*\right)\alpha_N T\left(\alpha_N I+T^*T\right)^{-2}\varphi\|^2 \\
=&O_P\left[\left(\frac{1}{Nh^{p+q}}+h^{2\rho}\right)\left(\alpha_N^{min(\beta,2)}+\frac{\alpha_N^{min(\beta+1,2)}}{\alpha_N}\right)\right]
\end{aligned}$$

and:

$$\begin{aligned}\|A_{12b}\|^2 \leq & \|\alpha_N \hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-2}\hat{T}^*\hat{T}\hat{T}^*\left(\hat{T}-T\right)\alpha_N\left(\alpha_N I + T^*T\right)^{-2}\varphi\|^2 \\ & + \|\alpha_N \hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-2}\hat{T}^*\hat{T}\left(\hat{T}^*-T^*\right)\alpha_N T\left(\alpha_N I + T^*T\right)^{-2}\varphi\|^2 \\ = & \|\alpha_N \hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\left(\alpha_N I + \hat{T}\hat{T}^*\right)^{-1}\hat{T}\hat{T}^*\left(\hat{T}-T\right)\alpha_N\left(\alpha_N I + T^*T\right)^{-2}\varphi\|^2 \\ & + \|\alpha_N \hat{T}\left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\left(\alpha_N I + \hat{T}\hat{T}^*\right)^{-1}\hat{T}\left(\hat{T}^*-T^*\right)\alpha_N T\left(\alpha_N I + T^*T\right)^{-2}\varphi\|^2 \\ = & O_P\left[\left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\left(\alpha_N^{min(\beta,2)} + \frac{\alpha_N^{min(\beta+1,2)}}{\alpha_N}\right)\right]\end{aligned}$$

In the same way, it is possible to show that:

$$\|A_{12c}\|^2 = O_P\left[\left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\left(\alpha_N^{min(\beta,2)} + \frac{\alpha_N^{min(\beta+1,2)}}{\alpha_N}\right)\right]$$

Finally:

$$\|A_{12d}\|^2 = O_P\left[\left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\alpha_N^{min(\beta,2)}\right]$$

which gives:

$$\|A_{12}\|^2 = O_P\left[\left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\alpha_N^{min(\beta,1)}\right]$$

and the result follows by multiplying each factor for $1/\alpha_N$.

1.9.3 Proof of Theorem (1.3.4)

First notice that minimizing the cross-validation function (1.3.6) is tantamount to minimize the following criterion:

$$CV(\alpha_N) = \| \left(I - Diag\left[\left(\alpha_N I + \hat{T}\hat{T}^*\right)^{-1}\hat{T}\hat{T}^*\right]\right)^{-1}\left(\hat{T}\hat{\varphi}^{\alpha_N} - \hat{r}\right)\|^2$$

Therefore:

$$CV(\alpha_N) \leq \| \left(I - Diag\left[\left(\alpha_N I + \hat{T}\hat{T}^*\right)^{-1}\hat{T}\hat{T}^*\right]\right)^{-1}\|^2 \|\hat{T}\hat{\varphi}^{\alpha_N} - \hat{r}\|^2$$

The norm of the residual sum of squares can be bounded as before, i.e.:

$$\|\hat{T}\hat{\varphi}^{\alpha_N} - \hat{r}\|^2 = O_P\left(\frac{1}{\alpha_N}\left(\frac{1}{N} + h^{2\rho}\right) + \left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\left(1 + \alpha_N^{\min(\beta,0)}\right) + \alpha_N^{\min(\beta+1,2)}\right)$$

which, because of $\beta > 0$, simplifies to:

$$\|\hat{T}\hat{\varphi}^{\alpha_N} - \hat{r}\|^2 = O_P\left(\frac{1}{\alpha_N}\left(\frac{1}{N} + h^{2\rho}\right) + \alpha_N^{\min(\beta+1,2)} + \left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\right)$$

The rest of the proof is to show that:

$$\| \left(Diag\left[I - \left(\alpha_N I + \hat{T}\hat{T}^*\right)^{-1}\hat{T}\hat{T}^*\right]\right)^{-1} \|^2 = O_P\left[\left(\frac{\alpha_N + 1}{\alpha_N}\right)^2\right]$$

First, notice that:

$$I - \left(\alpha_N I + \hat{T}\hat{T}^*\right)^{-1}\hat{T}\hat{T}^* = \alpha_N\left(\alpha_N I + \hat{T}\hat{T}^*\right)^{-1} = \hat{R}_{\alpha_N}$$

Furthermore, for $\alpha_N > 0$, $\hat{R}_{\alpha_N}$ is a normal bounded operator (Carrasco et al., 2007) and its diagonal elements belong to its numerical range. The latter is defined as the convex polygon whose vertices are the eigenvalues of $\hat{R}_{\alpha_N}$ (see, e.g. Herrero, 1991). Denote by d_{ii}, these diagonal entries. Since the eigenvalues of T^*T are bounded in the interval $(0, 1]$, the following inequalities hold:

$$\begin{aligned}\sup_{i\geq 0} d_{ii} &\leq \sup_{i\geq 0}\frac{\alpha_N}{\alpha_N + \lambda_i^2} < 1\\ \inf_{i\geq 0} d_{ii} &\geq \inf_{i\geq 0}\frac{\alpha_N}{\alpha_N + \lambda_i^2} \geq \frac{\alpha_N}{\alpha_N + 1}\end{aligned}$$

Which further implies that:

$$\sup_{i\geq 0}\frac{1}{d_{ii}} \leq \frac{\alpha_N + 1}{\alpha_N}$$

As the eigenvalues of a diagonal operator are equal to its diagonal elements, it follows that:

$$\| \left(Diag\left[\hat{R}_{\alpha_N}\right]\right)^{-1} \|^2 = O_P\left[\left(\frac{\alpha_N + 1}{\alpha_N}\right)^2\right]$$

■

1.9.4 Proof of Theorem (1.4.1)

Throughout this proof and the next one, we make extensive use of the following inequalities (see Engl et al., 2000):

$$\| \left(\alpha I + B^*B\right)^{-1} \|^2 \leq \alpha^{-2}$$
$$\| \left(B^*B\right)^{\mu} \alpha \left(\alpha I + B^*B\right)^{-1} \|^2 \leq \alpha^{2\mu}$$

together with Assumption (3) and inequality (1.4.1), with:

$$0 \leq \nu = \frac{u-s}{a+s} \leq 1$$

which explains why one needs to assume that $u \leq a + 2s$. Furthermore, we suppose that the order of the approximation error for the operator L^{-s} is negligible with respect to the estimation error of the operator T, and we therefore proceed *as if* L^{-s} is known.

We start by with the following decomposition:

$$\begin{aligned}
\|\hat{\varphi}^{(s),\alpha} - \varphi^{(s)}\|^2 &= \|\hat{\varphi}^{(s),\alpha} - \varphi^{(s),\alpha} + \varphi^{(s),\alpha} - \varphi^{(s)}\|^2 \\
&\leq \| \left(\alpha I + \hat{B}^*\hat{B}\right)^{-1} \hat{B}^*\hat{r} - \left(\alpha I + B^*B\right)^{-1} B^*B\varphi^{(s)}\|^2 \\
&+ \| \left(\alpha I + B^*B\right)^{-1} B^*B\varphi^{(s)} - \varphi^{(s)}\|^2 \\
&\leq \left\| \left(\alpha I + \hat{B}^*\hat{B}\right)^{-1} \left[\hat{B}^*\hat{r} - \hat{B}^*\hat{B}\varphi^{(s)}\right] \right\|^2 \qquad (A_1) \\
&+ \left\| \left[\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1} \hat{B}^*\hat{B} - \left(\alpha I + B^*B\right)^{-1} B^*B \right] \varphi^{(s)} \right\|^2 \qquad (A_2) \\
&+ \|\alpha \left(\alpha I + B^*B\right)^{-1} \varphi^{(s)}\|^2 \qquad (A_3)
\end{aligned}$$

The term A_3 is the regularization bias. Consequently:

$$\|A_3\|^2 = \|\alpha \left(\alpha I + B^*B\right)^{-1} \left(B^*B\right)^{\frac{u-s}{2(a+s)}} v\|^2 \leq O_P\left(\alpha^{\frac{u-s}{a+s}} \|\varphi\|_u^2\right)$$

Similarly, the term A_1 can be written as:

$$\begin{aligned}\|A_1\|^2 &= \left\| \left(\alpha I + \hat{B}^*\hat{B}\right)^{-1} L^{-s} \left[\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi\right] \right\|^2 \\ &\leq \left\| \left(\alpha I + \hat{B}^*\hat{B}\right)^{-1} (B^*B)^{\frac{s}{2(a+s)}} \right\|^2 \|\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi\|^2 \\ &\leq O_P\left[\frac{1}{\alpha^{2-\frac{s}{a+s}}}\left(\frac{1}{N} + h_N^{2\rho}\right)\right]\end{aligned}$$

from inequality (1.4.1) and assumption (4). Regarding the term A_2, it is possible to decompose it as follows:

$$\begin{aligned}\|A_2\|^2 &= \left\| \alpha\left[\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1} - (\alpha I + B^*B)^{-1}\right]\varphi^{(s)} \right\|^2 \\ &\leq \left\| \alpha\left[\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\left(\hat{B} - B\right)(\alpha I + B^*B)^{-1}\right]\varphi^{(s)} \right\|^2 \qquad (A_{21}) \\ &+ \left\| \alpha\left[\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\left(\hat{B}^* - B^*\right)B(\alpha I + B^*B)^{-1}\right]\varphi^{(s)} \right\|^2 \qquad (A_{22})\end{aligned}$$

Furthermore:

$$\begin{aligned}\|A_{21}\|^2 &\leq O_P(\alpha)\|\hat{T} - T\|^2 \left\| (B^*B)^{\frac{s}{2(a+s)}} (\alpha I + B^*B)^{-1} (B^*B)^{\frac{u-s}{2(a+s)}} v \right\|^2 \\ &\leq O_P\left[\alpha^{\frac{u-a-s}{a+s}}\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right)\|\varphi\|_u^2\right]\end{aligned}$$

and:

$$\begin{aligned}\|A_{22}\|^2 &\leq O_P(\alpha^{\frac{s}{a+s}})\|\hat{T}^* - T^*\|^2 \left\| (B^*B)^{\frac{1}{2}} (\alpha I + B^*B)^{-1} (B^*B)^{\frac{u-s}{2(a+s)}} v \right\|^2 \\ &\leq O_P\left[\alpha^{\frac{u-a-s}{a+s}}\left(\frac{1}{Nh_N^{p+q}} + h_N^{2\rho}\right)\|\varphi\|_u^2\right]\end{aligned}$$

The result of the theorem follows. ■

1.9.5 Proof of Theorem (1.4.2)

Following the proof of Theorem (1.3.4), minimizing the CV criterion is tantamount to the minimization of:

$$CV(\alpha) = \| \left(I - Diag \left[\left(\alpha I + \hat{B}\hat{B}^* \right)^{-1} \hat{B}\hat{B}^* \right] \right)^{-1} \left(\hat{T}\hat{\varphi}^{\alpha} - \hat{r} \right) \|^2$$

The operator B is a bounded linear operator with finite norm $\|B\|$. Therefore, the diagonal operator is bounded as before, i.e.:

$$\| \left(Diag \left[\alpha \left(\alpha I + \hat{B}\hat{B}^* \right)^{-1} \right] \right)^{-1} \|^2 = O_P \left[\left(\frac{\alpha + \|B\|}{\alpha} \right)^2 \right]$$

Now consider the remaining term. First note that, since $\varphi \in \mathcal{D}(L^u)$, then $\|\varphi\|_u < \infty$.

$$\begin{aligned} \|\hat{T}\hat{\varphi}^{\alpha} - \hat{r}\|^2 &\leq \|\hat{T}\hat{\varphi}^{\alpha} - T\varphi\|^2 + \|T\varphi - \hat{r}\|^2 \\ &\leq \|\hat{T}\hat{\varphi}^{\alpha} - \hat{T}\varphi^{\alpha}\|^2 + \|\hat{T}\varphi^{\alpha} - T\varphi\|^2 + \|T\varphi - \hat{r}\|^2 \\ &= \|A_1\|^2 + \|A_2\|^2 + \|A_3\|^2 \end{aligned}$$

The norm of A_3 corresponds to the nonparametric estimation error, so that:

$$\|A_3\|^2 = O_P \left(\frac{1}{Nh^{p+q}} + h^{2\rho} \right)$$

The squared norm of A_2 can be decomposed as follows:

$$\begin{aligned} \|A_2\|^2 &\leq \|\hat{T}\varphi^{\alpha} - T\varphi^{\alpha}\|^2 + \|T\varphi^{\alpha} - T\varphi\|^2 \\ &= \|\hat{T}L^{-s} \left(\alpha I + B^*B \right)^{-1} B^*T\varphi - TL^{-s} \left(\alpha I + B^*B \right)^{-1} B^*T\varphi\|^2 \\ &+ \|TL^{-s} \left(\alpha I + B^*B \right)^{-1} B^*T\varphi - T\varphi\|^2 \\ &= \|A_{21}\|^2 + \|A_{22}\|^2 \end{aligned}$$

A_{22} corresponds to the regularization bias and should converge to 0 as α approaches 0. One has then:

$$\begin{aligned}\|A_{22}\|^2 &= \|B\left(\alpha I + B^*B\right)^{-1} B^*T\varphi - T\varphi\|^2 = \|\alpha\left(\alpha I + BB^*\right)^{-1} T\varphi\|^2 \\ &= \|\alpha\left(\alpha I + BB^*\right)^{-1} BL^s\varphi\|^2 = \|\alpha\left(\alpha I + BB^*\right)^{-1} B\left(B^*B\right)^{\frac{u-s}{2(a+s)}} v\|^2 \\ &= \|\alpha\left(\alpha I + BB^*\right)^{-1} \left(B^*B\right)^{\frac{a+u}{2(a+s)}} v\|^2 = O_P\left(\alpha^{\frac{a+u}{a+s}}\|\varphi\|_u^2\right)\end{aligned}$$

Now consider the term A_{21}.

$$\begin{aligned}\|A_{21}\|^2 &= \|\left(\hat{T} - T\right) L^{-s}\left(\alpha I + B^*B\right)^{-1} B^*T\varphi\|^2 \leq \|\hat{T} - T\|^2 \|L^{-s}\left(\alpha I + B^*B\right)^{-1} B^*T\varphi\|^2 \\ &= \|\hat{T} - T\|^2 \|\left(\alpha I + B^*B\right)^{-1} B^*T\varphi\|_{-s}^2 \leq \|\hat{T} - T\|^2 \|\left(B^*B\right)^{\frac{s}{2(a+s)}}\left(\alpha I + B^*B\right)^{-1} B^*BL^s\varphi\|^2 \\ &= \|\hat{T} - T\|^2 \|\left(B^*B\right)^{\frac{s}{2(a+s)}}\left(\alpha I + B^*B\right)^{-1} \left(B^*B\right)^{\frac{2a+s+u}{2(a+s)}} v\|^2 = O_P\left[\alpha^{\frac{u}{a+s}}\|\varphi\|_u^2\left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\right]\end{aligned}$$

Finally, consider the term A_1.

$$\begin{aligned}\|A_1\|^2 &= \|\hat{T}\hat{\varphi}^\alpha - \hat{T}\varphi^\alpha\|^2 = \|\hat{T}L^{-s}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\hat{r} - \hat{T}L^{-s}\left(\alpha I + B^*B\right)^{-1} B^*T\varphi\|^2 \\ &\leq \|\hat{B}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\hat{r} - \hat{B}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\hat{T}\varphi\|^2 \\ &+ \|\hat{B}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\hat{T}\varphi - \hat{B}\left(\alpha I + B^*B\right)^{-1} B^*T\varphi\|^2 \\ &= \|A_{11}\|^2 + \|A_{12}\|^2\end{aligned}$$

The term A_{12} can be simplified as follows:

$$\begin{aligned}\|A_{12}\|^2 &= \|\alpha\hat{B}\left[\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1} - \left(\alpha I + B^*B\right)^{-1}\right] L^s\varphi\|^2 \\ &= \|\alpha\hat{B}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\left(\hat{B}^*\hat{B} - B^*B\right)\left(\alpha I + B^*B\right)^{-1} L^s\varphi\|^2 \\ &\leq \|\alpha\hat{B}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\left(\hat{B} - B\right)\left(\alpha I + B^*B\right)^{-1} L^s\varphi\|^2 \qquad (\|A_{12a}\|^2) \\ &+ \|\alpha\hat{B}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\left(\hat{B}^* - B^*\right) B\left(\alpha I + B^*B\right)^{-1} L^s\varphi\|^2 \qquad (\|A_{12b}\|^2) \\ &= O_P\left[\alpha^{\frac{u}{a+s}}\|\varphi\|_u^2\left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\right]\end{aligned}$$

The result arises from the fact that:

$$\begin{aligned}\|A_{12a}\|^2 &\leq \|\alpha\left(\alpha I + \hat{B}\hat{B}^*\right)^{-1}\hat{B}\hat{B}^*\|^2\|\left(\hat{T}-T\right)L^{-s}\left(\alpha I + \hat{B}\hat{B}^*\right)^{-1}L^s\varphi\|^2 \\ &\leq \alpha^2\|\hat{T}-T\|^2\|\left(\alpha I + \hat{B}\hat{B}^*\right)^{-1}(B^*B)^{\frac{u-s}{2(a+s)}}\,v\|^2_{-s} \\ &= O_P\left[\alpha^{\frac{u}{a+s}}\|\varphi\|^2_u\left(\frac{1}{Nh^{p+q}}+h^{2\rho}\right)\right]\end{aligned}$$

and

$$\begin{aligned}\|A_{12b}\|^2 &\leq \|\alpha\hat{B}\left(\alpha I + \hat{B}\hat{B}^*\right)^{-1}\|^2\|L^{-s}\left(\hat{T}^*-T^*\right)B\left(\alpha I + \hat{B}\hat{B}^*\right)^{-1}L^s\varphi\|^2 \\ &\leq \alpha\|\left(\hat{T}^*-T^*\right)\left(\alpha I + \hat{B}\hat{B}^*\right)^{-1}(B^*B)^{\frac{u-s}{2(a+s)}}\,v\|^2_{-s} \\ &= O_P\left[\alpha^{\frac{u}{a+s}}\|\varphi\|^2_u\left(\frac{1}{Nh^{p+q}}+h^{2\rho}\right)\right]\end{aligned}$$

Finally:

$$\begin{aligned}\|A_{11}\|^2 &= \|\hat{B}\left(\alpha I + \hat{B}\hat{B}^*\right)^{-1}L^{-s}\left(\hat{T}^*\hat{r}-\hat{T}^*\hat{T}\varphi\right)\|^2 \\ &\leq \|\hat{B}\left(\alpha I + \hat{B}\hat{B}^*\right)^{-1}\|^2\|\hat{T}^*\hat{r}-\hat{T}^*\hat{T}\varphi\|^2_{-s} \\ &= O_P\left[\alpha^{-\frac{a}{a+s}}\left(\frac{1}{N}+h^{2\rho}\right)\right]\end{aligned}$$

which gives the desired result. ■

Chapter 2

Nonparametric Instrumental Variable Estimation of Binary Response Models

joint with Jean-Pierre Florens

Abstract

We present an instrumental variable approach to the nonparametric estimation of binary outcome regression models with endogenous independent variables. In order to achieve identification, we use the reduced form model associated to the decomposition of the unobservable dependent variable into the space spanned by the instruments, and we suppose disturbances in this reduced form model to have a known distribution. We prove consistency of this estimator and run an extensive simulation study to corroborate its usefulness as a preliminary and exploratory tool. An empirical application demonstrates the performance of the proposed method relative to existing semiparametric estimators.

2.1 Introduction

An important recent literature has considered the nonparametric estimation of the separable instrumental variable model defined by the relation:

$$Y = \varphi(Z) + U \tag{2.1.1}$$

under the assumption, $\mathbb{E}(U|W) = 0$. The variables Y and Z are endogenous (in particular Z and U may be dependent) and W denotes the instruments (see,e.g. Newey and Powell, 2003; Hall and Horowitz, 2005; Carrasco et al., 2007; Darolles et al., 2011*a*; Chen and Pouzo, 2012*a*, and many others). In the majority of these papers, the regression function $\varphi(\cdot)$ is estimated by solving a regularized version of a functional equation.

The objective of this work is to propose a nonparametric estimation of the function $\varphi(\cdot)$ in the case where Y is not directly observed. We assume instead to observe a binary transformation of it, i.e. $\tilde{Y} = \mathbb{1}(Y \geq 0)$.

Previous literature on the topic has examined the semiparametric estimation of binary regression models with continuous endogenous variables (see Blundell and Powell, 2004; Rothe, 2009). In order to correct the endogeneity bias, these authors advocate a control function approach. Identification is achieved by specifying a parametric form for the function φ and estimating nonparametrically

the distribution of the error term (see also Klein and Spady, 1993; Ahn et al., 2004).

In this chapter, we propose instead a nonparametric estimation of φ. We make use of the fact that the variable Y can be also written as:

$$Y = \mathbb{E}(Y|W) + \varepsilon$$

and we suppose the conditional distribution of ε given W to be known. In particular, we consider the case in which the distribution of the errors is normal (Probit model) and logistic (Logit model). Finally, we obtain φ as the solution of the following functional equation:

$$\mathbb{E}(\varphi(Z)|W) = \mathbb{E}(Y|W)$$

When the two sides of this equation are estimated using any nonparametric method, the solution is known to be an *ill-posed* inverse problem, and needs a regularization method. We follow here the approach of Darolles et al. (2011*a*), and explore the properties of a Tikhonov regularized solution in the case where the dependent variable is binary.

Through a simulation study, we show the finite sample properties of our estimator and we acknowledge its usefulness as a preliminary and exploratory tool for binary models with endogenous regressors. Finally, we compare its properties to the semiparametric estimator of Rothe (2009) in an empirical application to interstate migration in the US. We provide evidence that our model can be used as an alternative to existing semiparametric frameworks when there is evidence of nonlinear dependencies in the endogenous variable.

2.2 The Model

Let (Y, Z, W) a random vector in $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$, such that:

$$Y = \varphi(Z) + U \quad \text{with} \quad \mathbb{E}(U|W) = 0 \tag{2.2.1}$$

where $\varphi(\cdot)$ is an unknown function in $\mathbb{L}_z^2$, the space of square integrable functions with respect to

the generating distribution of the data. Model (2.2.1) is equivalent to:

$$\mathbb{E}(\varphi(Z)|W) = r \tag{2.2.2}$$

where $r = \mathbb{E}(Y|W)$, assuming Y square integrable. When Y is directly observable, the standard way to proceed is to estimate r using any nonparametric technique and finally solve the inverse problem to obtain an estimator of φ (see Darolles et al., 2011*a*; Horowitz, 2011, among others).

In this chapter, we consider the estimation of φ in the case where the endogenous variable Y is not observable. Instead, we suppose to have at hand a binary discrete transformation of it $\tilde{Y} = \mathbb{1}(Y \geq 0)$. The additional difficulty in this case is to obtain an estimation of r from $\tilde{Y}$ and W.

Notice that the identification condition of model (2.2.1) remains unchanged in this case. Define $T\varphi = \mathbb{E}(\varphi(Z)|W)$ where $T : \mathbb{L}_z^2 \to \mathbb{L}_w^2$ is the conditional expectation operator. The function φ is still uniquely determined by equation (2.2.2) if T is one to one, or, equivalently, if:

$$T\varphi \overset{a.s.}{=} 0 \quad \Rightarrow \quad \varphi \overset{a.s.}{=} 0 \tag{2.2.3}$$

(see Newey and Powell, 2003; Darolles et al., 2011*a*). We assume this *completeness condition* to hold throughout.

Let us remind that model (2.2.1) can be rewritten as follows (see Chen and Reiss, 2011; Florens and Simoni, 2012)

$$Y = \mathbb{E}(\varphi(Z)|W) + \varepsilon \quad \text{where} \quad \mathbb{E}(\varepsilon|W) = 0$$

which represents the decomposition of Y as the sum of its conditional expectation with respect to W plus a residual term, where:

$$\varepsilon \equiv \varphi(Z) - \mathbb{E}(\varphi(Z)|W) + U$$

Via this decomposition, we have that:

$$\begin{aligned}\mathbb{P}(\tilde{Y}=1|W=w) =& \mathbb{P}\left(Y \geq 0|W=w\right) = \mathbb{P}\left(r(w)+\varepsilon \geq 0|W=w\right) \\ =& 1 - G_{\varepsilon|w}\left(-r(w)\right)\end{aligned}$$

where G is the conditional distribution of the error term, ε, with respect to W.

As usual in binary regression models, we cannot jointly nonparametrically identify the conditional expectation function r and the conditional distribution of the error term $G_{\varepsilon|w}$, unless we are willing to restrict r into a particular class of functions (see Matzkin, 1992). Therefore, we need to make some parametric assumption about either of these terms.

A viable approach would be to replace the unknown conditional expectation function r with some finite parametric specification, e.g.:

$$r = \sum_{k=0}^{J} W^k \beta_k \quad \text{where} \quad \beta_0 = 1$$

One could then estimate the vector of parameters β_k and $G_{\varepsilon|w}$ nonparametrically (see Manski, 1985; Horowitz, 1992; Klein and Spady, 1993; Ichimura, 1993, among others).

An alternative approach is to suppose that the conditional distribution of the error term $G_{\varepsilon|w}$ is known and then obtain an estimator of r by inversion of the known function $G_{\varepsilon|w}$.

The former approach has the advantage of not imposing any parametric restriction on the distribution of the error term, and therefore avoids model misspecification. However, a finite-dimensional parametric approximation of the conditional expectation function can lead to seriously erroneous conclusions if it is incorrect. In our case especially, a wrong inference about r impacts directly the estimation of φ.

In this chapter, therefore, we advocate the latter approach. In fact, if we consider the nonparametric model to be an exploratory tool, we might prefer to misspecify the distribution of the error, but to obtain correct inference about the shape of the function of interest. Another reason to prefer the second model is that, when economic theory can support a specific form of the conditional expectation function, one can impose such a restriction and estimate, either parametrically or nonparametrically, the shape of the distribution G_ε (see Matzkin, 1991, 1992).

In practice, we are going to suppose that the conditional distribution of the disturbances, $G_{\varepsilon|w}$, is either normal or logistic with constant standard deviation. In applications, identification is tantamount to classical Probit and Logit models. Take two solutions φ_1 and φ_2, and the corresponding residual variances σ_1 and σ_2. Write:

$$G_{\sigma_1,w}\left(\mathbb{E}\left[\varphi_1|w\right]\right) = G_{\sigma_2,w}\left(\mathbb{E}\left[\varphi_2|w\right]\right)$$
$$\sigma_1 G_w\left(T\varphi_1\right) = \sigma_2 G_w\left(T\varphi_2\right)$$

If we suppose G to be bijective and using the completeness condition (3.2.5), we have:

$$T\left(\varphi_1 - \frac{\sigma_2}{\sigma_1}\varphi_2\right) = 0 \quad \Rightarrow \quad \varphi_1 - \frac{\sigma_2}{\sigma_1}\varphi_2 = 0$$

Hence, the functions φ_1 and φ_2 are distinguishable only if we assume either that $\sigma_1 = 1$ or, equivalently, that $\|\varphi_1\| = 1$. The main assumption of this chapter is, therefore, about the homoskedasticity of the residuals ε, conditionally on the instruments W. Notice, that we do not require the error term ε to be independent of W.

Our main assumption is tantamount to:

$$Var\left(Y|W=w\right) = Var\left[\left(\varphi(Z)+U\right)|W=w\right] = \sigma^2 \tag{2.2.4}$$

where σ^2 is a constant, independent from the particular realization w of the instruments W.

Two remarks are in order. As in classical Probit and Logit models, our framework breaks down in the presence of heteroskedasticity. The distribution of the error term ε generally depends on W, hence, according to the application we have in mind, it would be more or less reasonable to assume that the conditional distribution of the errors does not vary with the particular realization of the instruments.

Second, it would be possible to characterize a simple linear system of simultaneous equation as a special case of our model. The following example clarifies this statement.

Example 2 (Linear simultaneous equations)**.** Assume for simplicity that $p = q = 1$, so that

$(Z, W) \in \mathbb{R}^2$, and consider model (2.2.1) with:

$$\varphi(Z) = Z\beta$$

and

$$Z = \zeta(W) + V$$

where V is an random noise, such that $\mathbb{E}(V|W) = 0$ and V is correlated with U, so that Z is endogenous. Then, we have that:

$$\varepsilon = U + (Z - \zeta(W))\,\beta = U + V\beta$$

Write the joint conditional variance of the residual components U and V as:

$$Var\begin{pmatrix} U \\ V \end{pmatrix}|W = w\Bigg) = \begin{pmatrix} \tau_U^2(w) & \tau_{UV}(w) \\ \tau_{UV}(w) & \tau_V^2(w) \end{pmatrix}$$

Then:

$$Var\left(\varepsilon|W = w\right) = \tau_U^2(w) + \tau_V^2(w)\beta^2 + 2\beta\tau_{UV}(w)$$

Therefore, our assumption is trivially satified when (U, V) is conditionally homoskedastic. For instance, (see also Heckman, 1978):

$$\begin{pmatrix} U \\ V \end{pmatrix}|W = w\Bigg) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}\right)$$

where τ is a constant in $[-1, 1]$.

Otherwise, one needs to place direct restrictions on the covariance function between U and V in such a way that:

$$\tau_{UV}(w) = \frac{1}{2\beta}\left(\sigma^2 - \tau_U^2(w) - \tau_V^2(w)\beta^2\right)$$

■

Hence, our estimator of r is defined as:

$$\hat{r}(w) = G_{\varepsilon|w}^{-1}\left[\hat{\mathbb{P}}\left(\tilde{Y} = 1|W = w\right)\right] \tag{2.2.5}$$

where $\hat{\mathbb{P}}\left(\tilde{Y} = 1|W = w\right)$ is the nonparametric estimator of the conditional probability function.

Finally, we obtain the function φ as the solution of the linear inverse problem (Carrasco et al., 2007):

$$T\varphi = r \tag{2.2.6}$$

The main issue arising from the non-parametric approach concerns the *ill-posedness* of the inversion of the operator T. The solution of the equation may not exist or is not in general a continuous function of the estimated part of the equation. The estimation is then not consistent in many cases. To cope with the inverse problem, we apply here a regularization method. In particular, we decide to use here the, so-called, *Tikhonov* regularization approach, advocated in Darolles et al. (2011*a*). However, any other regularization method could have been equivalently applied in this case (see, e.g. Horowitz, 2011; Florens and Racine, 2012; Johannes et al., 2013).

The solution of the inverse problem minimizes the following penalized criterion:

$$\varphi^{\alpha} = \arg\min_{\varphi} \|T\varphi - r\|^2 + \alpha\|\varphi\|^2$$

where, α is the regularization parameter which ought to be chosen using an appropriate data-driven method (see, also Fève and Florens, 2010).

2.3 Theoretical Properties

We suppose to observe an iid realization of the random variables $\left(\tilde{Y}, Z, W\right)$, that we denote $\{(\tilde{y}_i, z_i, w_i), i = 1, \ldots, N\}$.[1] We further assume, without loss of generality, that Z and W take values in $[0,1]^p$ and $[0,1]^q$, respectively. For simplicity, define $Q_{\varepsilon} = G_{\varepsilon}^{-1}$. In order to find the regularized solution of (2.2.6), we need to estimate the operator T, its adjoint T^*, and r.

[1] As usual, this assumption could be relaxed by assuming stationarity and mixing, see Hansen (2008)

All the low level assumptions are standard in the nonparametric IV literature, and we refer the interested reader to Darolles et al. (2011*a*) and Horowitz (2011) for a review of these.

We consider univariate generalized kernel functions K_h of order $l \geq 2$, where h is a bandwidth parameter; and the set of functions $\varphi \in \mathbb{C}^s$. We denote by $\rho = \min\{l, s\}$. In order to obtain uniform convergence of the regularization bias, we further suppose that our φ function has regularity $\beta > 0$. This boils down to the so-called *source condition* and it is discussed in details in Carrasco et al. (2007).

Denote by $f_{Z,W}$, f_Z and f_W, the joint and the marginal pdfs of Z and W respectively; and by $K_{W,h}$ and $K_{Z,h}$ the multivariate kernel functions of order l of dimension q and p, respectively. For any couple of functions, φ and ψ, the estimators of T, T^* and r are defined as follows:

$$\begin{aligned}
\left(\hat{T}\varphi\right)(w) &= \int \varphi(z)\frac{\hat{f}_{Z,W}(z,w)}{\hat{f}_W(w)}dz \\
\left(\hat{T}^*\psi\right)(z) &= \int \psi(w)\frac{\hat{f}_{Z,W}(z,w)}{\hat{f}_Z(z)}dw \\
\hat{r} &= Q_\varepsilon\left[\frac{\frac{1}{Nh^q}\sum_{i=1}^{N}\tilde{y}_i K_{W,h}(w-w_i,w)}{\hat{f}_W(w)}\right]
\end{aligned}$$

where $\hat{f}_{Z,W}$, $\hat{f}_Z$, and $\hat{f}_W$ are the usual nonparametric kernel estimators of the joint and marginal pdfs.

Then:

$$\hat{\varphi}^\alpha = \left(\alpha I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\hat{r} \tag{2.3.1}$$

is the estimate our binary nonparametric regression function.

The main difference with Darolles et al. (2011*a*) here is the fact that we cannot explicitly compute the conditional expectation of Y given W, as Y is not observed.

We maintain the following assumption about the cdf G_ε and the corresponding quantile function.

Assumption 5. *The function G_ε is monotone nondecreasing and right continuous. Furthermore, for each $p \in (0,1)$, it admits a generalized inverse, the quantile function, Q_ε, such that*

$Q_\varepsilon\left(G_\varepsilon(\varepsilon_0)\right) \leq \varepsilon_0$. *This inverse is monotone, nondecreasing with continuous and bounded first derivatives.*

Note that this assumption is satisfied by the Normal and the Logistic distribution. It is, however, more general than the case studied in this chapter. Furthermore, the assumption of boundedness of the first derivative of the quantile function is tantamount to the assumption of the conditional pdf, f_ε, being bounded away from zero. In fact, every quantile function, which satisfies assumption (5), can be written as solution of the following ordinary differential equation:

$$\frac{dQ_\varepsilon(p)}{dp} = \frac{1}{f_\varepsilon(Q_\varepsilon(p))}$$

To complete our study of the properties of our estimator, we make here the following high level assumption (a proof is provided in the appendix):

Assumption 6. *There exists $\rho \geq 2$, such that:*

$$\|\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi\|^2 = O_P\left(N^{-1} + h^{2\rho}\right)$$

This assumption is essentially the same as assumption A4 in Darolles et al. (2011*a*, p. 1553). In this case, we are also able to avoid the curse of dimensionality in the instrument by integrating them out. The intuition behind the preservation of this property is that we are simply applying a continuous transformation (the quantile function Q_ε) to our nonparametric estimator of the conditional probability.

With these assumptions, we obtain the same asymptotic properties as in the case where the variable Y is directly observed, i.e.:

$$\|\hat{\varphi}^\alpha - \varphi\|^2 = O_P\left[\frac{1}{\alpha^2}\left(\frac{1}{N} + h^{2\rho}\right) + \left(\frac{1}{Nh^{p+q}} + h^{2\rho}\right)\alpha^{(\beta-1)\wedge 0} + \alpha^{\beta\wedge 2}\right]$$

2.4 Estimation

Our estimator of the regression function φ is obtained as follows:

(i) We estimate nonparametrically the conditional expectation operator, T, and the conditional probability function $\mathbb{P}(\tilde{Y} = 1|w)$.

(ii) We invert the know conditional distribution function, in order to get $\hat{r}$, as described in (2.2.5).

(iii) We estimate the adjoint operator T^*, and find the Tikhonov regularized solution φ^α.

Step (i)

Define $p(w) = \mathbb{P}(\tilde{Y} = 1|w)$, the regression function in interest of our binary nonparametric regression model.

Signorini and Jones (2004) extensively discuss, among other methods, the use of local constant versus local linear logit regression in the class of binary models. They conclude that local linear logit regression has to be preferred over a local constant specification, although the difference is not so clear cut. Moreover, in this case, potential disadvantages of the local linear logit is that it does not ensure that the probability to be bounded between 0 and 1; and it does not have a closed form expression (as the weighted objective function is nonlinear in the parameter of interest) and requires a numerical optimization procedure at each estimation point.

Therefore, we decide to preserve the simplicity of the estimation and apply a standard Nadayara-Watson estimator[2], i.e.:

$$\hat{p}(w) = \frac{\sum_{i=1}^{N} \tilde{y}_i K_{h_w}(w_i - w)}{\sum_{i=1}^{N} K_{h_w}(w_i - w)} = \hat{T}\tilde{y}$$

with bandwidth parameters h_w.

Step (ii)

The main assumption of this chapter is that the conditional distribution of the error term ε is known. Therefore, to retrieve the estimator of conditional expectation function, $\hat{r}$, we simply

[2]It would be also possible in some cases to use variable kernel method as bias reduction technique for the local constant estimator, as advocated in Hazelton (2007).

use the quantile function associated to the distribution G_ε, and the estimator of the conditional probability obtained in step (i) (see equation 2.2.5).

Step (iii)

We finally obtain the nonparametric instrumental regression function by solving (2.2.6), using a Tikhonov regularization method (see equation 2.3.1).

The adjoint operator T^* defines the conditional expectation of all square integrable functions of W given Z. Therefore, a natural nonparametric estimator is:

$$\hat{T}^*\hat{r} = \frac{\sum_{i=1}^{N} \hat{r}_i K_{h_z}(z_i - z)}{\sum_{i=1}^{N} K_{h_z}(z_i - z)}$$

with bandwidth parameter, h_z.

Finally, in order to derive the value of the regularization parameter, we adopt the cross validation criterion proposed in chapter 1.

Using the optimal selection criterion, we obtain the first step Tikhonov estimator of the regression function as described in (2.3.1).

As described in Fève and Florens (2010), it is also possible to update the smoothing parameters for the conditional expectation functions $\mathbb{E}(\varphi(z)|w)$ and $\mathbb{E}(\mathbb{E}(\varphi(z)|w)|z)$, using our first step estimation of the function φ. We discuss the advantages versus the disadvantages of a two step estimation in this context in the next session.

2.5 Finite sample behavior

In this section we provide a Monte-Carlo simulation to explore the finite sample properties of our estimator. The numerical example is calibrated on the empirical application presented in the next section. We consider a real endogenous variable Z and two instruments W_1 and W_2.

The data generating process is as follows:

$$\begin{aligned} Y =& \mathbb{E}\left(\varphi(Z)|W\right) + \varepsilon \\ Z =& 0.15W_1 + 0.16W_2 + \eta \end{aligned}$$

where:

$$\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix} \right)$$
$$\eta \sim \mathcal{N}\left(0, (0.17)^2\right)$$

The residual term ε is generated according to a Normal, a Logistic and a mixture of normal distributions, with mixing coefficients 0.8 and 0.2, i.e. $\varepsilon|w \sim 0.8\mathcal{N}\left(-1, 0.05\right) + 0.2\mathcal{N}\left(4, 0.15\right)$. The latter simulation scheme, adapted from Rothe (2009), has been employed to assess the performance of our estimation under asymmetric distribution of the error term. The standard deviation of the disturbance ε has been set equal to 0.05 and it is taken as known; w_i, η_i and ε_i are mutually independent, for every i.

We employ two specifications for the function φ: it is chosen equal to $-z^2$, and to $-0.075e^{-|z|}$ (Darolles et al., 2011*a*; Florens and Simoni, 2012). These functional forms are employed as we can easily compute the corresponding conditional expectation functions. Define:

$$\Gamma(w_1, w_2) = 0.15w_1 + 0.16w_2$$

Then:

$$\mathbb{E}\left(Z^2|W = w\right) = \sigma_\eta^2 + \Gamma^2(w_1, w_2)$$

and:

$$\begin{aligned} \mathbb{E}\left(0.075e^{-|Z|}|W = w\right) = 0.075e^{0.5\sigma_\eta^2} \Bigg[& e^{-\Gamma(w_1,w_2)} \left(1 - \Phi\left(\sigma_\eta - \frac{\Gamma(w_1, w_2)}{\sigma_\eta}\right)\right) \\ & + e^{\Gamma(w_1,w_2)} \Phi\left(-\sigma_\eta - \frac{\Gamma(w_1, w_2)}{\sigma_\eta}\right) \Bigg] \end{aligned}$$

where Φ denotes the cdf of a standard normal distribution.

We work with a sample size of $N = 1000$, and we estimate the model both under a Probit ($G_\varepsilon \sim \mathcal{N}$) and a Logit ($G_\varepsilon \sim$ Logistic) specification. We run the simulation using each time 250 simulated samples of the residuals ε.

We use standard Gaussian kernels. The regularization parameters is computed as explained in section (2.4). The bandwidth parameters are obtained using leave-one-out cross validation[3].

Figures (2.1) and (2.3) report the estimation results when using a Probit specification of the model. Figures (2.2) and (2.4) report instead the results using a Logit specification. For each figure, we plot the true function (dashed light-grey line), against the mean of the first step estimator (grey line), and the median of the second step estimator (black line). We also plot their respective 90% simulated confidence intervals (dotted-dashed lines).

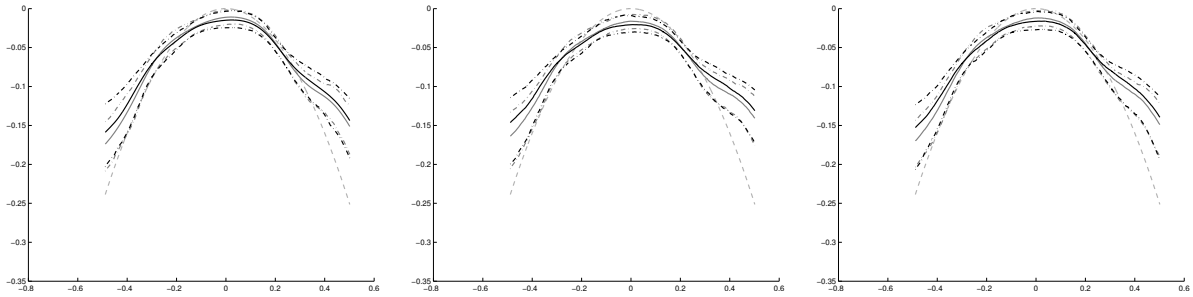
As expected, there is not a significant advantage in choosing between a Probit and a Logit specification of the model, as the two display similar results. In both cases, the first step estimator, $\hat{\varphi}_1$, performs better in terms of bias, while it has in general a greater variance than the second step estimator. This might be due to the fact that we generally undersmooth when computing the estimators of $\mathbb{E}(\hat{\varphi}_1(z)|w)$ and $\mathbb{E}(\mathbb{E}(\hat{\varphi}_1|w)|z)$, with respect to the estimation of $p(w)$, and of $\mathbb{E}(\mathbb{E}(\hat{r}|w)|z)$. This is compensated computationally by a larger value of the regularization parameter, which decreases the variance, but at a cost of a much larger regularization bias.[4] Therefore, we suggest using the first step estimator in this context.

Furthermore, the regularity of the function of interest does change the quality of our results. As a matter of fact, our estimator performs much better in the case where we take a very regular function (z^2) compared to the case where the function is highly irregular ($e^{-|z|}$). This is particularly evident when the distribution of the error term is not symmetric and we estimate using a Logistic specification.

2.6 An empirical application: interstate migration in the US

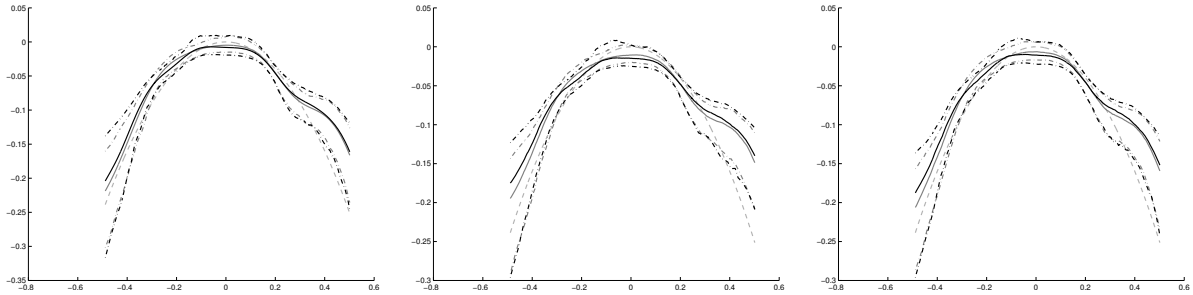
We now apply the proposed approach for the estimation of a binary choice model of interstate migration in the United States. The sample is drawn from the 2003 wave of the *Panel Study of*

[3] Codes, in MatLab and R, are available upon request.
[4] MSE comparison not reported here indicates that the second step estimator has to be preferred.



(a) $\varepsilon|w \sim \mathcal{N}$ (b) $\varepsilon|w \sim$ Logistic (c) $\varepsilon|w \sim$ Mixture

Figure 2.1: Estimation of the regression function $\varphi(z) = -z^2$ using a Probit specification. The true function (dashed light grey line) is plotted against the median of the first step (dark grey line) and the second step (black line) Tikhonov estimators, and their simulated confidence intervals.



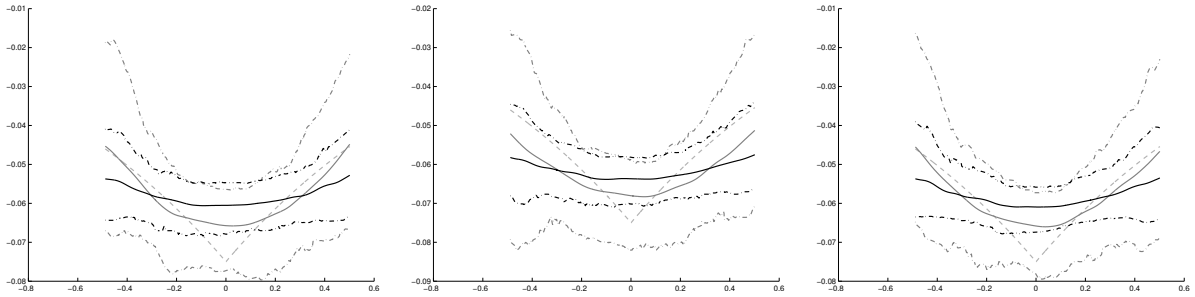
(a) $\varepsilon|w \sim \mathcal{N}$ (b) $\varepsilon|w \sim$ Logistic (c) $\varepsilon|w \sim$ Mixture

Figure 2.2: Estimation of the regression function $\varphi(z) = -z^2$ using a Logit specification. The true function (dashed light grey line) is plotted against the median of the first step (dark grey line) and the second step (black line) Tikhonov estimators, and their simulated confidence intervals.

Income Dynamics (PSID), a large household panel survey conducted in the US.

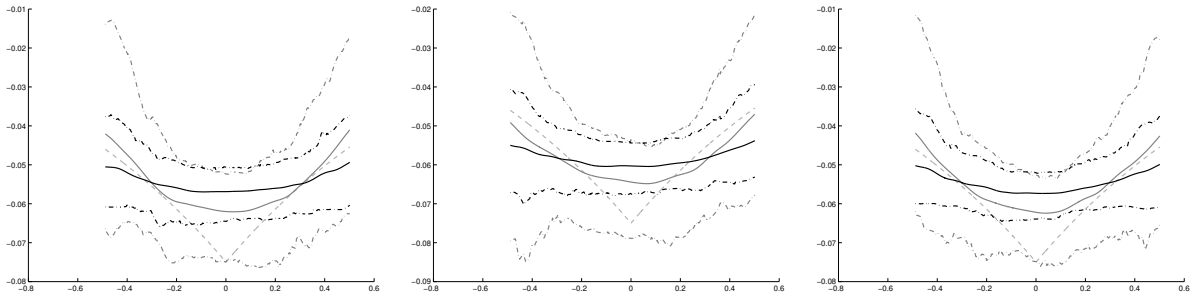
The choice to move to another US state may be related to higher expected income in the new state of residence. However, income is expected to increase, if and only if the individual decides to move. This makes income a potentially endogenous dependent variable.

Following Dong (2010) and Escanciano et al. (2011), we construct a sample of non-student male household heads, aged 22 to 69, with positive labor income during the year 2002-2003. To avoid results driven by outliers, we trim those individuals whose labor income is below the 0.01 and above the 99.9 percentile. We then obtain information about migration by comparing the state of residence declared in 2003, with the state of residence in the following waves of the panel (2005,2007 and 2009). In this way, we obtain a sample of 3642 observations. The binary endogenous dependent



(a) $\varepsilon|w \sim \mathcal{N}$ (b) $\varepsilon|w \sim$ Logistic (c) $\varepsilon|w \sim$ Mixture

Figure 2.3: Estimation of the regression function $\varphi(z) = -0.075e^{-|z|}$ using a Probit specification. The true function (dashed light grey line) is plotted against the fist step (dark grey line) and the second step (black line) Tikhonov estimators, and their simulated confidence intervals (dotted-dashed lines).



(a) $\varepsilon|w \sim \mathcal{N}$ (b) $\varepsilon|w \sim$ Logistic (c) $\varepsilon|w \sim$ Mixture

Figure 2.4: Estimation of the regression function $\varphi(z) = -0.075e^{-|z|}$ using a Logit specification. The true function (dashed light grey line) is plotted against the fist step (dark grey line) and the second step (black line) Tikhonov estimators, and their simulated confidence intervals (dotted-dashed lines).

variable $\tilde{Y}$ is defined as follows:

$$\tilde{Y} = \begin{cases} 1 & \text{if the household head has moved in the years 2004-2009} \\ 0 & \text{otherwise} \end{cases}$$

Due to attrition, we only observe $Y = 1$ for roughly 10% of the sample. The endogenous covariate Z is the log of the reported labor income. We also use a set of control variables X, such as a college dummy, the log of age and the log of family size. In order to instrument the endogenous variable Z, we have chosen the log of utility expenditure (such as gas, electricity, water, etc.) and the log of transport costs[5]. These instrumental variables are clearly unlikely to be correlated with

[5]Some descriptive statistics for these variables are given in Table (3.6).

the choice of migration. However, they might be a very good proxy of income as higher expenses in utilities are generally related to a bigger house; and higher transport costs might indicate higher expenditure on leisure[6].

	Mean	St.Dev	Min	Max
Migration Decision	0.09	0.29	0.00	1.00
Log Income	10.45	0.81	5.30	12.21
Log Utilities Expenditure	5.32	0.73	1.61	8.76
Log Transport Costs	4.88	0.72	0.69	8.41
Log Age	3.69	0.28	3.09	4.23
College	0.59	0.49	0.00	1.00
Log Family Size	1.02	0.51	0.00	2.30

Table 2.1: Summary statistics from the Panel Study Income Dynamics.

Since we introduce a number of exogenous variables, we decide to use the following semiparametric model:

$$\tilde{Y} = \mathbb{1}\left(\mathbb{E}\left(\varphi(Z)|W, X\right) + X\beta + \varepsilon \geq 0\right)$$

It appears that our partially linear specification is supported against the null of a fully parametric model, as the Hsiao et al. (2007) test for the linear probability model rejects the latter in favor of the former.[7] Our main assumption becomes here about the distribution of the error term given X and W. Thus:

$$\varepsilon|W, X \sim \mathcal{N}\left(0, 1\right)$$

In order to estimate φ and β, we use an approach similar to backfitting.

(i) We estimate the conditional probability of $\tilde{Y}$ given X and W. Finally, we obtain $\hat{r}$ by inversion of the known conditional cdf of ε.

(ii) For a given value of β, we solve the inverse problem:

$$\hat{T}\varphi = \hat{r} - X\beta$$

where $\hat{T}$ is now the estimator of the conditional expectation operator onto the space of

[6]The instruments have been tested using a parametric specification. They pass the weak-identification test using the Kleibergen-Paap rank LM statistic (Kleibergen and Paap, 2006).

[7]We also test our partially linear specification against a set of nonparametric alternatives, using the cross validation procedure proposed by Härdle et al. (2000). It appears that our partially linear model does not beat any other possible nonparametric alternative. However, we maintain such a specification to simplify the description of the estimator.

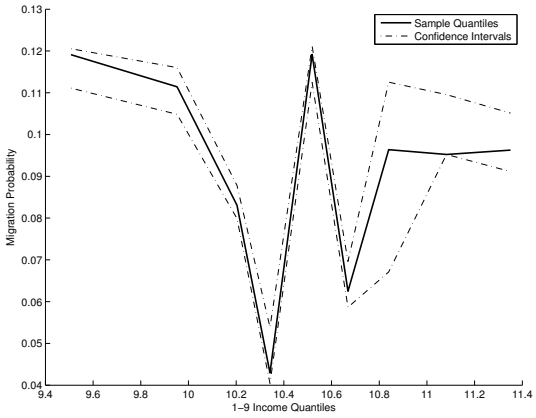


Figure 2.5: Average probability of migration by income quantile.

(X, W).

(iii) For $\hat{\mathbb{E}}\left(\hat{\varphi}^{\alpha_N}(z)|x,w\right)$ given, we estimate β using a simple parametric probit, where we control for the conditional expectation of $\hat{\varphi}^{\alpha_N}$. Optimality and $\sqrt{N}$-consistency of the estimated β follows from Florens et al. (2012).

The backfitting algorithm iterates the last two steps up to convergence of the following minimization criterion:

$$SSR(\alpha_N, \hat{\beta}) = \frac{1}{N\alpha_N} \left\| \hat{\mathbb{P}}\left(\tilde{y}|w,x\right) - \Phi\left[\hat{\mathbb{E}}\left(\hat{\varphi}^{\alpha_N}(z)|w,x\right) + x\hat{\beta}\right] \right\|^2$$

where Φ denotes the standard normal distribution. An initial value for β should be selected and should be not too far from the true value. In many cases 0 may be a suitable initial value.

Following the results in Burda (1993), we expect the coefficient associated to age and family size to be negative. Accordingly, the coefficient associated to the college dummy is expected to be positive. The effect of income is, however, not clear. For low revenue types, the probability of migration is higher, as they might want to move in order to improve their status. Using a linear approximation of φ and several parametric and semiparametric specifications, Dong (2010) indeed finds that migration probability is decreasing when labor income is increasing. The same result is confirmed in Escanciano et al. (2011). However, by plotting the average probability of interstate migration by income quantile (figure 2.5), it appears that probability is decreasing, but not in a linear fashion. This leaves rooms for nonparametric specification of the income effect in this context. We therefore employ our nonparametric procedure to the estimation of φ. For completeness, we compare our

result with the semiparametric specification of Rothe (2009), i.e. we estimate the model:

$$\tilde{Y} = \mathbb{1}\left(Z\gamma + X_1 + X_2\beta_2 + U \geq 0\right) \tag{2.6.1}$$

$$Z = \zeta(W) + V \tag{2.6.2}$$

where the matrix X is partitioned into X_1, a vector of college dummies, and X_2, a matrix of logarithmic age and family size. For identification reasons, we set the coefficient associated with the college dummy to be equal to 1. We remind that the additional identification condition with endogeneity is:

$$\mathbb{E}\left(U|W,V\right) = \mathbb{E}\left(U|V\right)$$

Since we do not observe V, we obtain a consistent estimator of it, $\hat{V}$, using the auxiliary regression model in (2.6.2). The link function ζ is estimated nonparametrically using leave-one-out bandwidths. Finally, we maximize the following log-likelihood function conditionally on the index, $Z\gamma + X_2\beta_2$, and the estimated residual $\hat{V}$[8]:

$$\log \mathcal{L}\left(\gamma, \beta_2, h\right) = \sum_{i=1}^{N}\left[\tilde{y}_i \hat{\mathbb{P}}(U|z_i\gamma + x_{2i}\beta_2, \hat{v}_i) + (1-\tilde{y}_i)\left(1 - \hat{\mathbb{P}}(U|z_i\gamma + x_{2i}\beta_2, \hat{v}_i)\right)\right]$$

where $\left\{\hat{\mathbb{P}}(U|z_i\gamma + x_{2i}\beta_2, \hat{v}_i), i = 1, ..., N\right\}$ is the nonparametric estimator of the conditional cdf of U, with bandwidth h. Notice that the log-likelihood function is jointly maximized in the coefficients, γ and β_2, and the vector of bandwidths h.

	SP-SI	SP-IV
	Migration	**Decision**
Log Income	-0.785	
	(0.488)	
Log Age	-2.168	-0.874
	(0.645)	(0.106)
College	1	0.402
	(-)	(0.065)
Log Family Size	-0.455	-0.191
	(0.248)	(0.058)

Table 2.2: Summary of regression results from SP-SI (column 1) and SP-IV (column 2) models. Standard Errors in brackets.

[8] See Rothe (2009) for a detailed explanation of the estimation procedure.

Table (2.2) reports the results of the estimation using the semiparametric single index model (SPSI, column 1), versus the linear part of our semiparametric instrumental variable estimation (SPIV, column 2). The standard errors are obtained using bootstrap in the former case, while in our semiparametric specification we simply retrieve them from the parametric probit model. The result for the coefficients are not very different in the two model specifications and have the expected sign. It has to be noticed that the coefficient associated to family size is not significant in the SPSI model. Turning our attention to the coefficient associated to the endogenous variable in the SPSI, we can see that its value is negative as expected and, therefore, consistent with existing evidence. However, this coefficient is barely significant. Based on previous observations on the nonlinear decay of the average probability, this does not come as a surprise since a linear specification might not be sufficient to capture the relation between income and migration decision.

Figure (2.6) draws the nonparametric instrumental variable estimator of the impact of income on migration probabilities. Bootstrap confidence intervals are obtained using the method discussed in chapter 3. We can observe that the function is indeed not monotonic. The income effect is marginally positive for low income values, and it then nonmonotonically decreases towards higher income. This nonlinear trend may be due to the fact that low income individuals may find convenient to move to a new state, especially if this displacement is associated with better living conditions and higher expected income. However, they may not have adequate means or opportunities to move elsewhere, especially if we consider that low income is often associated with low education and low skill jobs. This would explain while the curve is initially increasing. However, as income increases, everything else being equal, people have less incentives to relocate. This is consistent with existing evidence in the literature, as discussed above.

2.7 Conclusions

We propose in this chapter a very simple nonparametric instrumental variable approach to binary outcome models in presence of endogenous regressors, we prove its consistency and draw its finite sample properties via a simulation study. Our empirical application shows that our estimator is easy to apply and very flexible and can be used as as alternative framework to existing semiparametric models for endogenous regressors.

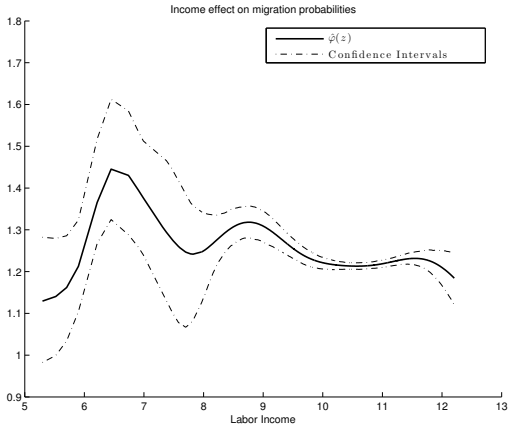


Figure 2.6: Functional estimator of the impact of income on migration decisions.

2.8 Appendix

2.8.1 Proof of Assumption 6

We denote by $\hat{r}^*$ the unfeasible estimator of the conditional expectation of Y given W.

Remember that $\hat{r} = Q_\varepsilon(\hat{p}(w))$. We start by considering a Taylor expansion of the quantile function $Q_\varepsilon(\hat{p}(w))$, around $G_\varepsilon(\hat{r}^*)$.

$$\begin{aligned} Q_\varepsilon(\hat{p}(w)) =& Q_\varepsilon(G_\varepsilon(\hat{r}^*)) + Q_\varepsilon^{'}(G_\varepsilon(\hat{r}^*))(\hat{p}(w) - G_\varepsilon(\hat{r}^*)) + o(|\hat{p}(w) - G_\varepsilon(\hat{r}^*)|^2) \\ \leq& \hat{r}^* + Q_\varepsilon^{'}(G_\varepsilon(\hat{r}^*))(\hat{p}(w) - G_\varepsilon(\hat{r}^*)) + o(|\hat{p}(w) - G_\varepsilon(\hat{r}^*)|^2) \end{aligned}$$

by Assumption 5. Then:

$$\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi = \hat{T}^*\hat{r}^* + \hat{T}^*Q_\varepsilon^{'}(G_\varepsilon(\hat{r}^*))(\hat{p}(w) - G_\varepsilon(\hat{r}^*)) \tag{2.8.1}$$

where, for simplicity, we omit higher order terms.

We consider the Hilbert-Schmidt norm of the term on the lhs of 2.8.1:

$$\begin{aligned} &\|\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi\|^2 \\ &\leq 2\|\hat{T}^*\hat{r}^* - \hat{T}^*\hat{T}\varphi\|^2 + 2\|\hat{T}^*Q_\varepsilon^{'}(G_\varepsilon(\hat{r}^*))(\hat{p}(w) - G_\varepsilon(\hat{r}^*))\|^2 \\ &= 2\|A_1\|^2 + 2\|A_2\|^2 \end{aligned}$$

Using assumption A4 in Darolles et al. (2011*a*, p. 1553), we can show that:

$$\|A_1\|^2 = O_P\left(N^{-1} + h^{2\rho}\right)$$

Now we turn to A_2. By the properties of the quantile function, the boundedness of the conditional density of the disturbances, and the definition of $\hat{p}(w)$, and $\hat{T}^*$, we obtain:

$$\begin{aligned}
A_2 =& \int \left[\frac{1}{f_\varepsilon\left(Q_\varepsilon(G_\varepsilon(\hat{r}^*))\right)}\left(\frac{\frac{1}{Nh^q}\sum_{i=1}^N \tilde{y}_i K_h(w-w_i,w)}{\hat{f}(w)} - G_\varepsilon(\hat{r}^*)\right)\right]\frac{\hat{f}(w,z)}{\hat{f}(z)}dw \\
=& \int \left[\frac{1}{f_\varepsilon\left(Q_\varepsilon(G_\varepsilon(\hat{r}^*))\right)}\left(\frac{1}{Nh^q}\sum_{i=1}^N (\tilde{y}_i - G_\varepsilon(\hat{r}^*))K_h(w-w_i,w)\right)\right]\frac{\hat{f}(w,z)}{\hat{f}(z)\hat{f}(w)}dw \\
\leq& \int \left[\frac{1}{\inf_\varepsilon[f_\varepsilon(\varepsilon)]}\left(\frac{1}{Nh^q}\sum_{i=1}^N (\tilde{y}_i - G_\varepsilon(\hat{r}^*))K_h(w-w_i,w)\right)\right]\frac{\hat{f}(w,z)}{\hat{f}(z)\hat{f}(w)}dw \\
\leq& O_p(1)\int \left[\left(\frac{1}{Nh^q}\sum_{i=1}^N (\tilde{y}_i - G_\varepsilon(\hat{r}^*))K_h(w-w_i,w)\right)\right]\frac{\hat{f}(w,z)}{\hat{f}(z)\hat{f}(w)}dw \\
=& \int B_N(w)\frac{\hat{f}(w,z)}{\hat{f}(z)\hat{f}(w)}dw = \tilde{A}_2
\end{aligned}$$

By the uniform convergence properties of kernel density estimators (Hansen, 2008; Darolles et al., 2011*b*), it is possible to show that:

$$\tilde{A}_2 = \int B_N(w)\frac{f(w,z)}{f(w)f(z)}dw + O_p\left(\int B_N(w)\frac{f(w,z)}{f(w)f(z)}dw\right)$$

Notice that, $\tilde{y}_i - G_\varepsilon(\hat{r}^*)$ is iid uniform between $[-1,1]$, so that uniformly in z:

$$\tilde{A}_2 = O_P\left(N^{-1} + h^{2\rho}\right)$$

Following the proof of Darolles et al. (2011*b*).

Chapter 3

Implementation, Simulations and Bootstrap in Nonparametric Instrumental Variable Estimation

joint with Frédérique Fève and

Jean-Pierre Florens

Abstract

We present a rather thorough investigation of the use of regularization methods for the estimation of nonparametric regression models with instrumental variables. We consider various version of Tikhonov, Landweber-Fridman and Galerkin regularization. We review data-driven techniques for the sequential choice of the smoothing and the regularization parameters. Through intensive Monte-Carlo simulations, we discuss the finite sample properties of each regularization method and the validity of wild bootstrap confidence bands in this context. Finally, we investigate the use of these methodologies in the estimation of the Engel curve for food for a sample of rural households in Pakistan.

3.1 Introduction

Instrumental variables are popular in econometrics to achieve identification and perform inference in the presence of endogenous explanatory variables. Empirical applications of this framework are vast, e.g. structural estimation of the Engel curve (Blundell et al., 2007), of demand functions (Hoderlein and Holzmann, 2011) or of returns to education in a homogeneous population (Blundell et al., 2005).

However, in many empirical application, 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.

On the contrary, 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).

However, while nonparametric estimation with instrumental variables (also known as nonparametric instrumental regression) has recently received enormous attention in the theoretical literature

(see, e.g. Darolles et al., 2011*a*; Horowitz, 2011, and references therein), it remains 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 which defines it is not continuous. Hence, the estimation of this type of models requires, beside the usual selection of the smoothing parameter for the nonparametric regression, 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.

The tuning of the latter parameter constitutes an additional layer of complication and it has to be tackled with the appropriate method. Data-driven techniques for the choice of regularization parameter in the framework of nonparametric instrumental regressions are presented in the first chapter of this manuscript and in Fève and Florens (2010); Florens and Racine (2012), and Horowitz (2012).[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.

The contribution of this work is therefore to review several regularization techniques that can be applied when the explanatory variable is endogenous and the regression function is estimated nonparametrically using instrumental variables. We consider the simple framework of an additive separable model, with a single endogenous covariate, a single instrument and without additional exogenous regressors. We analyze the performances of several version of Tikhonov (Darolles et al., 2011*a*), Landweber-Fridman (Johannes et al., 2013; Florens and Racine, 2012) and Galerkin (Cardot and Johannes, 2010; Horowitz, 2011) regularizations in the case where both the smoothing and the regularization parameters are chosen using data-driven methods.

Moreover, we assess the performances of *wild bootstrap* to obtain pointwise confidence intervals

[1] The few notables exceptions we are aware of are Blundell et al. (2007); Hoderlein and Holzmann (2011) and Sokullu (2010)

[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).

in this framework. Confidence bands may be extremely important to draw conclusions about the variability of the estimation and to assess unusual features of the estimated regression curve. Moreover, in this context, they can serve to test for the exogeneity of the independent variable (Blundell and Horowitz, 2007). However, nonparametric instrumental regressions lack of a general procedure to obtain them. Chen and Pouzo (2012*b*); Horowitz and Lee (2012) and Santos (2012) study bootstrap in nonparametric instrumental regressions and prove its validity but only in the very specific framework of Galerkin regularization. The wild bootstrap presented in this work is instead of more general applicability and, in particular, it can be used independently of the regularization scheme under consideration.

The chapter is structured as follows. In section (3.2), we present the main framework. We review carefully each regularization scheme, and we discuss its practical implementation in section (3.3). In sections (3.4) and (3.5), we describe the structure of the Monte-Carlo experiment, and expose the bootstrap procedure and its validity. In section (3.6), we present an application to the estimation of the Engel curve for food using a cross section database of Pakistan households. Finally, section (3.7) concludes.

3.2 The main framework

We focus our analysis on a simple framework characterized by a triplet of random variables $(Y, Z, W) \in \mathbb{R}^3$, verifying the following model:

$$Y = \varphi(Z) + U \tag{3.2.1a}$$

$$\mathbb{E}(U|W) = 0 \tag{3.2.1b}$$

This model is a regression type model, where the usual mean independence condition $\mathbb{E}(U|Z) = 0$ is replaced by condition (3.2.1b). This specification has been extensively studied in econometrics in order to account for the possible *endogeneity* of Z (i.e. the lack of independence between the covariate Z and the error U), under the name of instrumental variable regression. In particular, recent literature has investigated the nonparametric estimation of the function $\varphi(\cdot)$ in (3.2.1a) (see,e.g. Newey and Powell, 2003; Hall and Horowitz, 2005; Carrasco et al., 2007; Darolles et al.,

2011*a*; Chen and Pouzo, 2012*a*, among others).

The main specificity of the model considered here is that $\varphi(\cdot)$ has to be found as the solution of an integral equation of the first kind, i.e.

$$\mathbb{E}(\varphi(Z)|W) = \mathbb{E}(Y|W) \tag{3.2.2}$$

which leads to a linear inverse problem. However, this problem is generally *ill-posed* (see Engl et al., 2000). To briefly illustrate the matter, denote by $r = \mathbb{E}(Y|W)$, and $T\varphi = \mathbb{E}(\varphi(Z)|W)$, so that (3.2.2) now writes:

$$T\varphi = r \tag{3.2.3}$$

We assume that the triplet (Y, Z, W) is characterized by its joint cumulative distribution function F, dominated by the Lebesgue measure. Denote by f its probability density function. We consider the space of square integrable function relative to the true F and we denote, for instance, by $\mathbb{L}_z^2$, the space of square integrable functions of Z only. We further assume that $Y \in \mathbb{L}_z^2$ and $r \in \mathbb{L}_w^2$. The operator T defines the following linear mapping:

$$\begin{aligned} T: \quad & \mathbb{L}_Z^2 \to \mathbb{L}_W^2 \\ (T\varphi)(w) &= \int \varphi(z) f(z|w) dz \end{aligned}$$

In order to solve (3.2.3), we also require its adjoint T^*, which is defined as follows:

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

and

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

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{L}_Z^2$ or in $\mathbb{L}_W^2$.

The operators T and T^* are taken to be compact (see, e.g. Carrasco et al., 2007; Darolles et al., 2011*a*), 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) $T\varphi_i = \lambda_i \psi_i$

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

For every othonormal sequence $\psi_i \in \mathbb{L}_W^2$ and $\phi_i \in \mathbb{L}_Z^2$. Using the singular value decomposition of T, we can rewrite equation (3.2.3) as:

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

where $\varphi_j = \langle \varphi, \phi_j \rangle$ and $r_j = \langle r, \phi_j \rangle$ are the Fourier coefficients of φ and r, respectively. We point out that compacteness it is not a simplifying assumption in this context, but describes a realistic framework in which the eigenvalues of the operator are declining to zero. Assuming that the eigenvalues are bounded below is relevant for other econometric models, but it is not realistic in the case of continuous nonparametric instrumental variable estimation.

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

$$T\varphi \overset{a.s.}{=} 0 \quad \Rightarrow \quad \varphi \overset{a.s.}{=} 0 \tag{3.2.5}$$

(see Newey and Powell, 2003; Darolles et al., 2011*a*; Andrews, 2011; D'Haultfoeuille, 2011). This *completeness condition* is assumed to hold throughout, and it guarantees that the eigenvalues of the operator T are strictly positive, although converging to 0 at some rate.

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:

$$\varphi = \sum_{j=1}^{\infty} \frac{r_j}{\lambda_j} \psi_j \tag{3.2.6}$$

The ill-posedness in (3.2.3) 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 $j \to \infty$, which entails the *ill-posedness* of the problem. This leads to a non consistent estimation of the function φ.

(ii) The right hand side of the equation needs to be estimated. This approximation introduces a

further estimation error component which renders the ill-posedness of the problem even more severe.

Therefore, the problem in (3.2.3) 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 denominator in (3.2.6) does not blow up. One could add to every eigenvalue λ_j a small constant term. This constant term *controls* the rate of decay of the λ_j's to 0 (Tikhonov regularization). Another approach would be to replace the infinite sum in (3.2.6) by a finite approximation of it, and estimate the Fourier coefficients by projection on an arbitrary function basis of the instruments and the endogenous variable (Galerkin 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 the *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 to the direct operator inversion.

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

3.3 Implementation of the regularized solution

Once we have chosen our preferred nonparametric estimator (local constant kernels, local polynomials, splines), the implementation of regularization methods requires, beside the choice of the smoothing parameter for the nonparametric regression, the selection of a regularization constant in order to cope with the *ill-posedness* of the inverse problem.

Despite a correspondence between the smoothing and the regularization parameters clearly exists, their simultaneous choice is, to the best of our knowledge, not feasible. The most judicious approach is to select them sequentially. As a matter of fact, it seems that the regularization parameter adjusts to the choice of the smoothing parameter in a reasonable set of values.[3]

For practical applications, it is essential to obtain data-driven techniques for the selection of both types of parameters. There is already a vast literature about the selection of the smoothing parameter for nonparametric regressions (for a review, see Härdle, 1990; Li and Racine, 2007). Hence,

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

here we mainly focus our attention on the methods for the optimal selection of the regularization parameter, and we suppose that the smoothing parameter has been chosen using our preferred data-driven approach.

Given the smoothing parameter, an inadequate choice of the regularization parameter has a substantial impact on the final estimation: 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.

In the following, we suppose to observe an iid realization of the random variables (Y, Z, W), which we denote $\{(y_i, z_i, w_i), i = 1, \ldots, N\}$.

The linear operator T and the rhs of (3.2.3), r, can be estimated using our favorite nonparametric regression technique (e.g., local polynomials, regression splines). Finally, we need to choose a regularization rule, which identifies our solution as function of our nonparametric estimates of r and T. The remainder of this section reviews the regularization methods we undertake in this chapter, and discusses, for each of them, a criterion for the data-driven choice of the regularization parameter.

3.3.1 Tikhonov Regularization

The Tikhonov regularization method (TK henceforth) is based on the minimization of the following criterion function (Darolles et al., 2011*a*):

$$\|T\varphi - r\|^2 + \alpha\|\varphi\|^2 \tag{3.3.1}$$

which leads to find the function φ as the solution of the following system of equations:

$$\alpha\varphi + T^*T\varphi = T^*r \tag{3.3.2}$$

Notice that, in this equation, only the right hand side can be estimated from the data, while the left hand side depends on the unknown function φ. The conditional expectation of Y given W is estimated as, $\hat{r} = \hat{T}y$, where $\hat{T}$ corresponds to the matrix of kernel weights (see Fève and Florens,

2010) or to the orthogonal projection of the y's on the space spanned by the spline basis of W. Similarly, the adjoint operator T^* is estimated as the conditional expectation function of $\mathbb{E}(\hat{r}|Z)$. For each of these estimator, a smoothing parameter is chosen using least square cross validation. Finally, a first step estimator of φ is obtained by replacing these estimators in (3.3.2), i.e.,

$$\hat{\varphi}^{\alpha} = \left(\alpha I + \hat{T}^*\hat{T}\right)^{-1} \hat{T}^* r \tag{3.3.3}$$

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

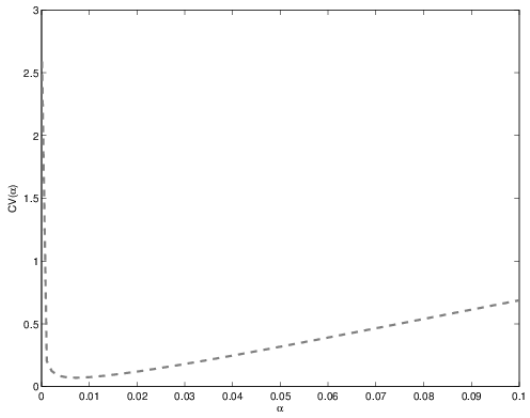


Figure 3.1: Criterion function for the optimal choice of α in Tikhonov regularization

In order to choose the regularization parameter α, we adopt the cross validation approach developed in Chapter 1. A typical shape of this criterion function can be found in figure (3.3.1).

Once an initial estimate of φ is obtained, it would be possible to select new smoothing parameters for the estimation of the left hand side of (3.3.2). That is, to replace T and T^*, on the lhs with the matrices of weights obtained from the estimation of $\mathbb{E}(\hat{\varphi}^{\alpha}|W)$ and of $\mathbb{E}(\hat{\mathbb{E}}(\hat{\varphi}^{\alpha}|W)|Z)$, respectively; and to finally iterate the choice of the regularization parameter for these new smoothing parameters.

However, there is not a theoretical (or practical) evidence that the iterative approach improves the estimation. As a matter of fact, the quality of this scheme strongly depends on the first step estimator. If the latter poorly approximates the function of interest, we cannot, in general, be sure to converge to a better outcome. Thus, in this chapter, we only consider the performance of the first step TK estimator.

3.3.2 Landweber-Fridman Regularization

The Landweber-Fridman (LF henceforth) regularization consists of an iterative approach, which is meant to avoid the inversion of a large matrix (Johannes et al., 2013). If we multiply both sides of equation (3.2.3) by T^*, the solution φ can be written as:

$$cT^*T\varphi = cT^*r$$

where c is a scalar constant, such that $\|T^*T\| < 1/c$. The iterative approach is about finding a fixed point of the system of equations. Therefore, by adding and subtracting φ on the left hand side, we obtain the recursive solution:

$$\varphi_{j+1} = \varphi_j + cT^* \left(r - T\varphi_j\right), \quad \forall j = 0, 1, \ldots \tag{3.3.4}$$

or equivalently:

$$\varphi^M = c \sum_{j=0}^{M-1} \left(I - cT^*T\right)^j T^*r \tag{3.3.5}$$

where M is the total number of iterations needed to reach the solution. M plays here the role of regularization parameter. As M diverges to infinity the regularized solution in (3.3.5) converges to the true φ. Asymptotically, it can be shown that $M \simeq 1/\alpha$, where α is the regularization parameter in the Tikhonov approach (see, e.g. Florens and Racine, 2012).

In order to implement the LF regularization, we use the iterative scheme from equation (3.3.4). We proceed as follows:

(i) We compute smoothing parameters h_0, for the estimation of r, and of $\mathbb{E}(r|Z)$. As for TK regularization, this allows us to obtain $\hat{T}_{h_0}$ and $\hat{T}^*_{h_0}$, first step estimators of the operators T and T^*, where subscripts are used to stress the dependence on a specific value of the smoothing parameter.

(ii) We set the initial condition $\hat{\varphi}_0 = c\hat{T}^*_{h_0}\hat{r}_{h_0}$. This is consistent with equation (3.3.5) for $j = 0$.

(iii) Using $\hat{\varphi}_0$, we update smoothing parameters for the estimation of $\mathbb{E}(\hat{\varphi}_0|W)$, and of $\mathbb{E}(\mathbb{E}(Y - \hat{\varphi}_0|W)|Z)$. Define these new smoothing parameters as h_1. We therefore obtain updated

estimators of the operators, $\hat{T}_{h_1}$ and $\hat{T}^*_{h_1}$.[4]

(iv) By equation (3.3.4), we compute $\hat{\varphi}_1$ as:

$$\hat{\varphi}_1 = \hat{\varphi}_0 + c\hat{T}^*_{h_1}\left(\hat{r}_{h_0} - \hat{T}_{h_1}\hat{\varphi}^0\right)$$

(v) For $j = 2, 3, \ldots$, we repeat steps (iii) and (iv), until the following criterion is minimized (see also Florens and Racine, 2012):

$$SSR(j) = j\left\|\hat{T}\hat{\varphi}_j - \hat{r}\right\|^2 \quad , \quad j = 1, 2, \ldots$$

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 j in order to admit a minimum. A typical shape of this function is reported in figure (3.2). 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, we iterate at least until $j = c^{-1}N^{1/4}$.[5] The shape of the function can then be checked *ex-post* for local minima.

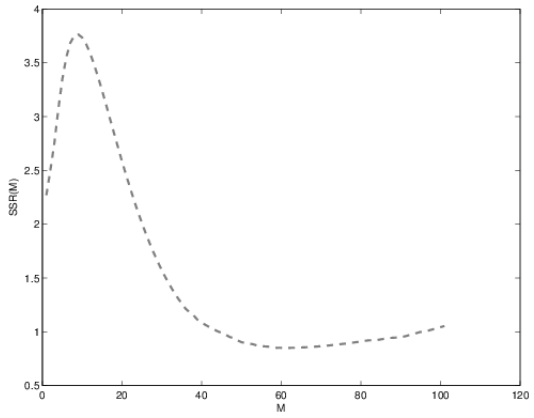


Figure 3.2: Stopping function for Landweber-Fridman regularization

[4]Updated smoothing seems natural, in this context, to account for the relation between regularization and smoothing parameters. It also appears that the this strategy is MSE minimizing. We would like to thank Jeffrey S. Racine for insightful discussions on this topic.

[5]This stopping rule is justified by the fact that the Tikhonov regularization parameter $\alpha \simeq N^{-\frac{1}{4}}$ asymptotically (Darolles et al., 2011*a*) Since $M \simeq 1/\alpha$, it follows $M \simeq N^{1/4}$. We then multiply by the inverse of the constant as convergence towards the solution is slower as c decreases.

3.3.3 Galerkin Regularization

The Galerkin type of regularization (GK henceforth) consists on truncating the infinite sum in (3.2.6), by a finite approximation on an *arbitrary* basis (see, e.g. Cardot and Johannes, 2010; Horowitz, 2011).

Fix an orthonormal basis $\{\phi_j, j = 1, \ldots, J\}$ (e.g., B-Splines, Wavelets, Hermite polynomials,etc.), which does not necessarily correspond to the natural basis of operators T and T^*. Take an integer $J_n < \infty$, the solution given by Galerkin regularization can be written as:

$$\varphi^{J_n} = \sum_{j=1}^{J_n} \beta_j \phi_j \tag{3.3.6}$$

where $\beta_j = \langle \varphi, \phi_j \rangle$ are the Fourier coefficients, associated to the decomposition of φ on the space spanned by the basis functions, and the superscript J_n denotes again the dependence of the solution on the truncation parameter.

The implementation of this method is very simple: we need to estimate the Fourier coefficients β_j, for $j = 1, \ldots, J_n$ in (3.3.6), upon the choice of an orthonormal family of basis functions and of the truncation parameter J_n.

To the best of our knowledge, a theoretically justified rule for choosing the former is not available. We therefore decide to use cubic B-spline basis (Blundell et al., 2007; Horowitz, 2011). For every value of J_n, we obtain an estimator of the Fourier coefficients as follows:

(i) Define the two matrices of basis functions:

$$\mathcal{W}_n = [\phi_1(w), \ldots, \phi_{J_n}(w)] \qquad \mathcal{Z}_n = [\phi_1(z), \ldots, \phi_{J_n}(z)]$$

and the vector of Fourier coefficients, $\beta = \{\beta_1, \ldots, \beta_{J_n}\}$

(ii) Then:

$$\varphi^{J_n} = \sum_{j=1}^{J_n} \beta_j \phi_j = \mathcal{Z}_n \beta$$

(iii) We proceed as in a standard two stages least square problem and we obtain our estimator of

β as:

$$\hat{\beta} = \underset{\beta \in \mathcal{B}_{J_n}}{\arg\min} \left(Y - \mathcal{Z}_n \beta\right)' \left(\mathcal{W}_n \mathcal{W}_n'\right) \left(Y - \mathcal{Z}_n \beta\right)$$

where $\mathcal{B}_{J_n}$ is the parameter space that depends on the choice of J_n. This finally gives:

$$\hat{\beta} = \left(\mathcal{Z}_n' \mathcal{W}_n \mathcal{W}_n' \mathcal{Z}_n\right)^{-1} \left(\mathcal{Z}_n' \mathcal{W}_n \mathcal{W}_n' Y\right)$$

For the choice of the regularization parameter J_n, we follow the data driven method proposed by Horowitz (2012). Define $\mathcal{H}_{J_n,s}$ the Sobolev space of functions with s square integrable derivatives, whose decomposition is truncated at J_n. Define further:

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

Blundell et al. (2007) call ρ_{J_n} the sieve measure of ill-posedness. As $n \to \infty$, to obtain consistency of the estimator, we require $\rho_{J_n} \left(J_n^3/n\right)^{\frac{1}{2}} \to 0$ and $\rho_{J_n} \left(J_n^4/n\right)^{\frac{1}{2}} \to \infty$. We therefore need to find a value of J_n which satisfies these requirements. Such a value can be defined as:

$$J_{n_0} = \underset{J=1,2,\ldots}{\arg\min} \left\{\rho_J^2 J^{3.5}/n \quad : \quad \rho_J^2 J^{3.5}/n - 1 \geq 0\right\}$$

i.e., J_{n_0} is the smallest integer such that $\rho_J^2 J^{3.5}/n \geq 1$. The method for determining a feasible estimate of J_{n_0} 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 $\hat{T}_J^* \hat{T}_J$, where $\hat{T}_J^*$ and $\hat{T}_J$ are the estimators of the conditional expectation operators truncated at J.

(ii) Finally, define:

$$\hat{J}_{n_0} = \underset{J=1,2,\ldots}{\arg\min} \left\{\hat{\rho}_J^2 J^{3.5}/n \quad : \quad \hat{\rho}_J^2 J^3.5/n - 1 \geq 0\right\}$$

A typical shape of this criterion is drawn is figure (3.3).

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 2SLS estimator. Mariano (1972), in an influential paper, shows that the 2SLS estimator only possesses moments of order $p - q + 1$,

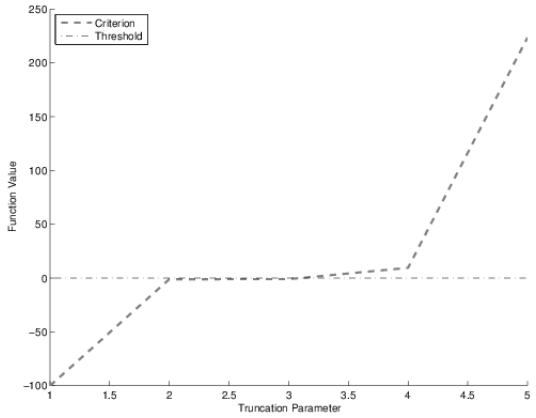


Figure 3.3: Choice of $\hat{J}_n$ for Galerkin regularization.

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 have only finite mean but infinite variance. In order to obtain a finite variance in our sample, we therefore include an additional term in the matrix $\mathcal{W}_n$, so that its dimension is $J_n + 1$.[6]

3.3.4 Penalization by derivatives

The last approach presented in this work does not point out towards the realization of the regularization scheme, but rather to the methodological fact that we can use the restriction in (3.2.3) to obtain φ as the integral of its derivatives of any order. Therefore, we can regularize the derivative of the function of interest, instead of the function itself, in order to obtain an estimator that is smoother and less oscillating than the ones previously discussed.

We solely focus on the case when the penalization is on the first derivative of the function. This framework may be particularly relevant in economic applications as researchers are often interested in marginal effects. For instance, one could be interested in the estimation of demand elasticities, rather than the demand function itself.

In this section we thus work with functions having square integrable first derivative, i.e. $\varphi' \in \mathbb{L}^2_z$.

[6] Simulations ran with the same dimension for both matrices show indeed that the variance of the GK estimator becomes arbitrarily large when we do not correct for this effect.

Define the first order differential operator L. We can rewrite equation (3.2.3) as follows:

$$\begin{aligned} TL^{-1}L\varphi &= r \\ TL^{-1}\varphi^{'} &= r \\ B\varphi^{'} &= r \end{aligned}$$

where $B = TL^{-1}$. We can then obtain $\varphi^{'}$ as the solution of this equation, and, by definition, $\varphi = L^{-1}\varphi^{'}$, where L^{-1} corresponds to the integral operator.

The main obstacle in the implementation of this estimator is to find the adjoint of the operator B, defined as:

$$B^* = \left(TL^{-1}\right)^* = \left(L^{-1}\right)^* T^*$$

This definition requires to find the adjoint of the first order integral operator L^{-1}. Following Florens and Racine (2012), we have, for all $\psi \in \mathbb{L}_Z^2$, that:

$$\left(L^{-1}\right)^* \psi(z) = -\left(\int_z^{\infty} \psi(u)du - \int \psi(u)du\right)$$

Now define a generic function λ, such that, $\lambda^{'} \in \mathbb{L}_W^2$; f_Z and S_Z, the pdf and the survivor function of Z, respectively; f_W, the pdf of W; and, finally,

$$S(u,w) = -\frac{\partial}{\partial w}\mathbb{P}\left(Z \geq u, W \geq w\right)$$

Then the adjoint operator, B^*, is such that:

$$\left(B^*\lambda\right)(u) = \frac{1}{f_Z(u)}\int \lambda(w)\left(S(u,w) - S_Z(u)f_W(w)\right)dw$$

The pdf and the survivor function can be estimated using nonparametric kernels. Suppose $K_h(\cdot)$ to be a continuous, positive, and bounded kernel, for a given bandwidth h, and define $\bar{\mathcal{K}}_h(a) = 1 - \int_{-\infty}^{a} K_h(b)db$. We then have:

$$\left(\hat{B}^*\lambda\right)(u) = \frac{1}{\hat{f}_Z(u)}\left\{\frac{1}{N}\sum_{i=1}^{N}\left[\bar{\mathcal{K}}_h\left(u - z_i\right)\lambda(w_i)\right] - \hat{S}_Z(u)\left(\frac{1}{N}\sum_{i=1}^{N}\lambda(w_i)\right)\right\}$$

For the selection of the bandwidth parameter h, we apply least squares cross validation. For the estimation of K and r, we can again apply any nonparametric technique. The corresponding smoothing parameters are chosen by cross validation.

The integral operator L^{-1} is approximated using a trapezoidal rule. I.e.

$$\left(\hat{L}^{-1}\varphi'\right)_i = \sum_{l=1}^{i} \varphi'_l \left(z_l - z_{l-1}\right) \quad , \quad i = 1, \ldots, N$$

where z_0 is normalized to be the smallest value taken by the random variable Z in the sample. Finally, $\hat{B} = \hat{T}\hat{L}^{-1}$.

Notice that, the operator L^{-1} is a proper inverse of L only on the space of centered functions, i.e. when $\mathbb{E}(\varphi) = 0$. Therefore, the estimator is identified up to a constant term. However, by the structural equation in (3.2.1*a*), we have that $\mathbb{E}(\varphi) = \mathbb{E}(y)$. Then, our final estimator is recentred, in order to have the same sample expectation as the dependent variable.

The implementation is based on both TK and LF regularization.

(i) **TK**. The derivative of the solution satisfies the following system of normal equations:

$$\hat{B}^*\hat{B}\varphi' = \hat{B}^* r \tag{3.3.7}$$

Notice that, in this case, the estimation is extremely simplified with respect to the case studied in Florens and Racine (2012). As a matter of fact, the normalization of the estimated adjoint operator $\hat{B}^*$ by the pdf of Z is not necessary, since both sides of (3.3.7) are multiplied by it. Moreover, we do not need to recenter the solution of this problem, as *a fortiori*, the mean of the function φ is the same as the mean of y, up to the regularization bias. With TK penalization of the first derivative, the solution is written as:

$$\varphi^\alpha = L^{-1}\varphi'^\alpha = L^{-1}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^* r$$

For the selection of α, we apply the same cross validation criterion presented above (see also Fève and Florens, 2013, for an application).

(ii) **LF**. The LF iterative solution writes:

$$\varphi_{j+1}^{'} = \varphi_{j}^{'} + cT^{*}\left(r - T\varphi_{j}^{'}\right), \quad \forall j = 0, 1, \ldots \tag{3.3.8}$$

where:

$$\varphi_{j} = L^{-1}\varphi_{j}^{'} - \mathbb{E}\left(L^{-1}\varphi_{j}^{'}\right)$$

with the initial condition:

$$\varphi_{0}^{'} = c\frac{1}{\hat{f}_{Z}}\left[\hat{S}r - \hat{S}_{Z}\hat{\mathbb{E}}_{N}\left(r\right)\right]$$

Finally:

$$\varphi_{j+1} = L^{-1}\varphi_{j+1}^{'} - \mathbb{E}\left(L^{-1}\varphi_{j+1}^{'}\right) + \mathbb{E}\left(y\right)$$

The smoothing parameters for the estimation of the pdf and the survivor functions are not updated from iteration to iteration (see also Florens and Racine, 2012). The choice of the smoothing parameters for the estimation of the operator T and the stopping criterion are, instead, identical to the baseline case.

3.4 Monte-Carlo Simulations

In this section, we analyse the performances of the various estimators previously discussed using data-driven methods. In particular, we consider the application of these regularizations under distinct nonparametric estimations. We inspect the behavior of local constant, local linear and B-splines estimation associated with TK and LF; local constant estimation with penalized first derivative; and finally a B-spline estimation for GK.

Couple of caveats are in order. The goal of this simulation study *is not* to compare the performance of the various estimation techniques, but rather to show the effectiveness of the data-driven approaches presented in this chapter and test the validity of the bootstrap, discussed in the next section. Our objective is not to specifically drive the empirical researcher towards one of these methods. By contrast, we may want to encourage to use various estimators simultaneously. Moreover, a simulation study which aims at comparing the various regularization techniques would be flawed by definition. This is because different regularities of the joint distribution of the endoge-

nous variables and the instruments, and smoothness of the true regression function are driving the degree of *ill-posedness* of the inverse problem. On the one hand, the estimators presented here may be more or less sensitive to these regularities; on the other hand, many choices related to the implementation are still not backed by valid theoretical arguments, and might be suboptimal for a particular design of the data.

The numerical example used in this chapter is based on the framework adopted by Darolles et al. (2011*a*), Florens and Simoni (2012) and Florens and Racine (2012). The main data generating process follows equation (3.2.1*a*):

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

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

$$\begin{aligned} W \sim & \mathcal{N}\left(0, 10^2\right) \\ V \sim & \mathcal{N}\left(0, (0.5)^2\right) \\ U \sim & \mathcal{N}\left(0, (0.05)^2\right) \\ Z = & \frac{1}{1 + \exp\left(-(0.1W + 40U + V)\right)} \\ Y = & Z^2 + U \end{aligned}$$

The main difference with the numerical examples reported in other papers is that the endogenous variable, Z, is a nonseparable function of the instrument, W, and the disturbances, U and V. The companion code for this chapter has been programmed in Matlab and it is available upon request from the authors.

We work with a modest sample size of 500 observations and we draw 1000 replications of the error terms V and U. Since the regressor Z is changing for each of these replications, we evaluate each estimator of φ on a grid of 500 equispaced points in $(0, 1)$.

When using B-splines, we fix the order of the basis to 4 (cubic splines), and we compute the optimal number of knots using either least squares cross validation (TK and LF) or the method developed

in Horowitz (2012) (GK). An important remark about the B-spline estimation is about the choice of knots. The boundary knots are placed at the minimum and the maximum of the observed data. We then place the interior knots uniformly between the two boundaries. The impact of free-knots (Stone, 2005) or quantile knots is not explored here and left to further research.[7]

For local constant and local linear estimation, the bandwidth parameters are all obtained by least squares cross validation (Li and Racine, 2007).

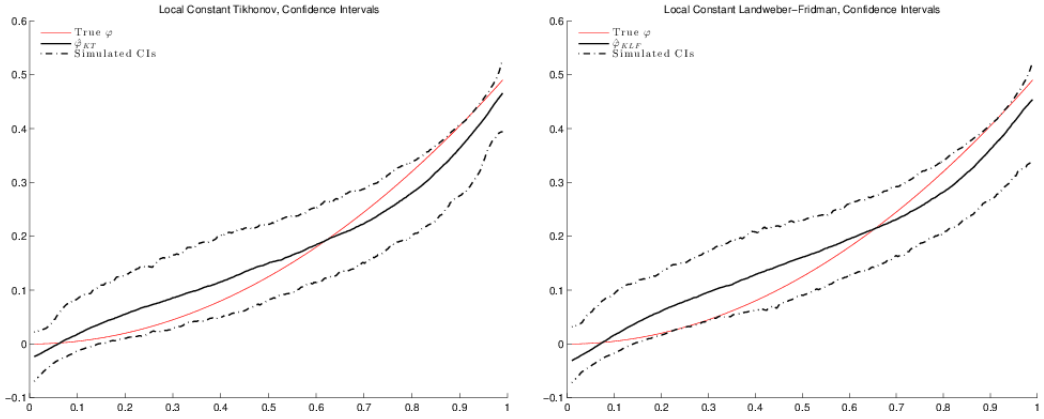
Notice that the use of least squares cross validation in this context is only of practical relevance, and it can be replaced by other methods. Possible alternatives include rule of thumb smoothing, maximum likelihood cross validation, or a modified AIC criterion (Hurvich et al., 1998). Notice, that all these methods are known to balance the trade-off between variance and bias for nonparametric regressions. In practice, following the discussion presented in chapter 1, this also seems appropriate in the case of nonparametric instrumental regressions.

Figures (3.4), (3.5), (3.6) and (3.7) report the results of our simulations for the local constant, local linear, B-splines and penalized first derivative local constant estimators. On the left panel of each figure, we draw the TK regularized solution; the LF solution is instead on the right panel. Figure (3.8) presents the same results for GK with B-splines. The red line in each figure is the true function φ. The thick black line is the median value of the regression function, obtained from simulations, at each evaluation point and the dashed lines give the 95% confidence intervals.

The comparison of the various estimators in terms of Mean Integrated Square Error (MISE), median Mean Square Error (MSE), variance and bias is given in Table (3.1). All estimators have roughly comparable performances. A comparison of the MISE shows that the Penalized Local Constant TK and the Spline TK estimators are those giving the best results for our simulation scheme. They generally have a lower bias and a lower variance compared to all other estimators. The GK regularization also gives good fitting of the true regression function. Its bias is very low, while its variance is substantially larger than the one of other estimators.

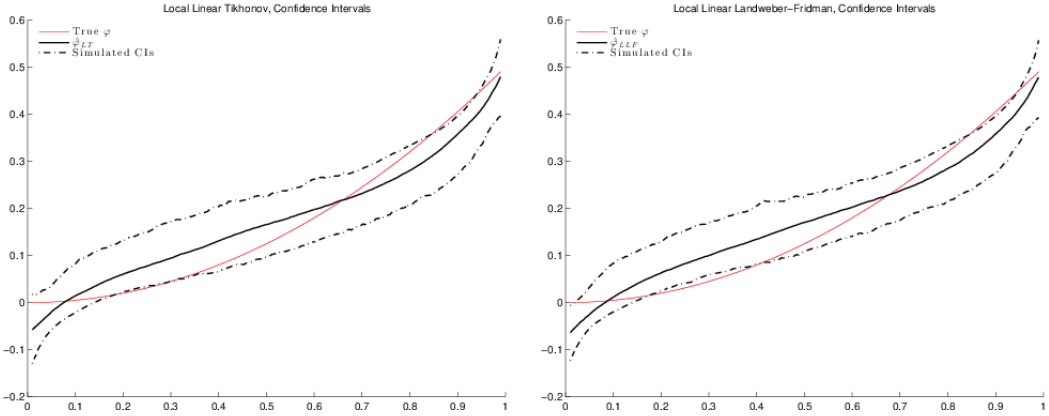
The Local Constant and Local Linear kernel estimators (both with TK and LF) present a larger bias. It is difficult to say whether higher bias comes from the selection of the smoothing or the

[7] Another important aspect to consider is that the position of the knots can be chosen adaptively to ensure the best fitting of the regressions curve (see Ma and Racine, 2013). This type of adaptive selection can be used with the *crsiv* function in R (Racine and Nie, 2012).



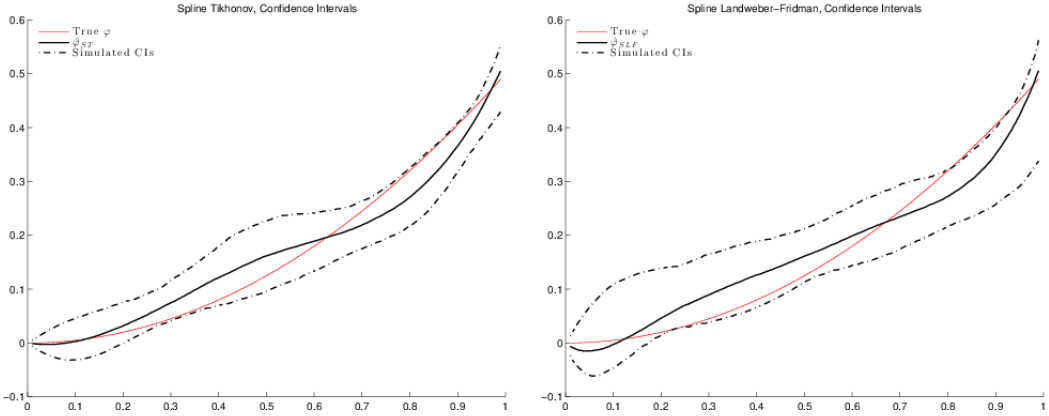
(a) Tikhonov (b) Landweber-Fridman

Figure 3.4: Simulations results using Local Constant Kernels



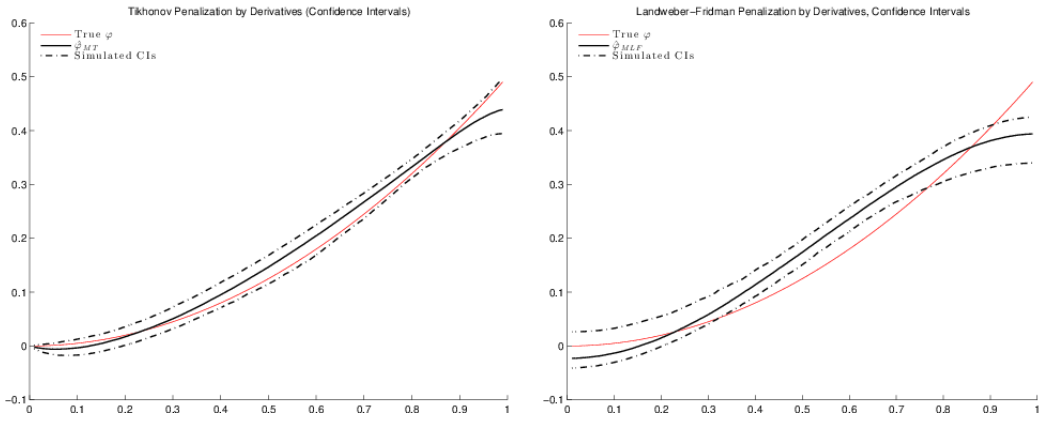
(a) Tikhonov (b) Landweber-Fridman

Figure 3.5: Simulations results using Local Linear Kernels



(a) Tikhonov (b) Landweber-Fridman

Figure 3.6: Simulations results using B-Splines



(a) Tikhonov (b) Landweber-Fridman

Figure 3.7: Simulations results using Local Constant Kernel with penalized first derivative

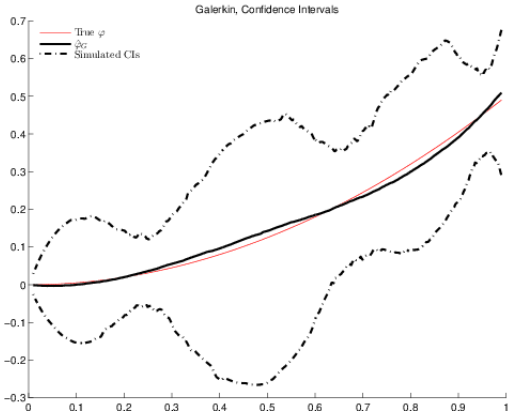


Figure 3.8: Simulations results using Galerkin with B-splines

regularization parameter. Variances are comparable across estimators both for LF regularization and TK regularization. Notice that the local constant and local linear estimator have higher median variance under TK rather than under LF. The opposite holds true for the spline and the penalized local constant. This latter result is consistent with the bias-variance trade off.

In order to explore further the differences between bias and variance for the different estimators, we report, in Table (3.2), summary statistics for the regularization parameter, by estimation type. Concerning both LF and TK regularization, it is clear from this table that the choice of the regularization parameter goes into the expected direction. In TK scheme, when α is selected to be small, the estimation bias is reduced, as it is the case for the Spline and for the Penalized Local Constant. This is consistent with the fact that splines tend to smooth more the regression function and therefore lead to select a smaller value of the regularization parameter. By contrast, in the Penalized Local Constant, the regularization is carried onto the first derivative of the function,

	MISE	MSE	Bias	Variance
Local Constant TK	0.00214	0.00219	0.01079	0.00119
Local Linear TK	0.00253	0.00260	0.01703	0.00104
Spline TK	0.00148	0.00129	0.00329	0.00057
Penalized TK	0.00039	0.00029	0.00872	0.00012
GK	0.01830	0.01344	0.00085	0.01336
Local Constant LF	0.00256	0.00278	0.01678	0.00117
Local Linear LF	0.00253	0.00256	0.01937	0.00084
Spline LF	0.00218	0.00196	0.01080	0.00087
Penalized LF	0.00163	0.00112	0.01942	0.00018

Table 3.1: MISE and Median MSE, Bias and Variance for each estimator.

which gives a smoother solution for the inverse problem (and a smaller value of the regularization parameter). The Local Constant and Local Linear estimators lead to a more rough estimation of the conditional expectation functions and, therefore, the data-driven criterion selects a larger value of the regularization parameter.

The same effect holds for the LF regularization. When the number of iterations, M, increases, the bias decreases and the variance rises. In this case too, nonparametric methods that lead to a smoother estimation of the regression function (as B-splines and local linear kernels) converge towards a larger number of iteration, i.e. lower regularization. While local constant kernels reach convergence, on average, for a lower value of M. The penalized local constant estimator is the one having the higher mean (and median) number of iterations, as its solution is smoother. This is reflected in practice by a lower bias and a larger variance of this estimator, as reported in Table (3.1).

A final remark is about the choice of the number of knots for the B-spline basis in the GK scheme. As it can be seen from Table (3.2), the optimal criterion is very conservative as it selects a small number of knots. For our simulation scheme, the data-driven criterion almost always selects 3 interior knots and, in some particular cases, 4 interior knots.

Finally, we also report in Table (3.3) 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 the truncation parameter plays in this case the role of regularization and smoothing constant. Therefore, it is not necessary to implement any type of CV criterion for the tuning of the smoothing parameter, which can be computationally very costly. Moreover, the

	Mean	Median	St.Dev	Min	Max
Local Constant TK ($\times 10^5$)	1288.3	1082.2	724.5	122.5	6043.2
Local Linear TK ($\times 10^5$)	716.0	391.8	959.4	3.9	11251.9
Spline TK ($\times 10^5$)	872.7	866.3	439.9	0.0	2794.7
Penalized TK ($\times 10^5$)	77.6	64.0	54.2	6.3	600.3
GK	3.0	3.0	0.2	3.0	4.0
Local Constant LF	45.9	46.0	14.0	9.0	118.0
Local Linear LF	68.4	44.0	82.1	10.0	1000.0
Spline LF	57.5	56.0	17.6	12.0	131.0
Penalized LF	256.8	248.5	85.0	87.0	641.0

Table 3.2: Summary statistics for the regularization parameter.

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 this work is 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.[8]

	Mean	Median	St.Dev	Min	Max
Local Constant TK	16.64	15.20	5.00	10.27	43.35
Local Linear TK	35.96	31.16	18.11	14.20	227.92
Spline TK	16.49	16.57	2.08	4.03	24.72
Penalized TK	13.30	12.54	2.81	8.13	29.35
GK	0.06	0.06	0.02	0.01	0.27
Local Constant LF	502.44	481.24	169.98	105.78	1304.09
Local Linear LF	2265.71	1139.09	2591.45	194.02	23390.28
Spline LF	720.31	656.09	337.61	110.41	2614.35
Penalized LF	1887.46	1819.83	615.88	285.43	4631.69

Table 3.3: CPU time for each estimator (in seconds).

[8] An additional comment about LF is that, although updating the regularization parameter at each iteration may be a MSE minimizing strategy, the gain in terms of MSE may not be sufficient to justify such a high computational time. This point is not explored in this work and it is left to further research.

3.5 Wild Bootstrap in Nonparametric IV

3.5.1 Resampling from sample residuals in Nonparametric Regression Models

In standard nonparametric regressions without endogeneity, the general theory of bootstrap is presented in Härdle and Bowman (1988); Cao-Abad (1991); Härdle and Marron (1991) and Hall (1992). To overview briefly the most common approaches , suppose for the moment that the variable Z can be considered as exogenous and that we want to estimate the following model:

$$Y = m(Z) + U \qquad \mathbb{E}(U|Z) = 0$$

In this case, bootstrap boils down to replace any occurrence of the unknown distribution of the error term by the empirical distribution function. However, this empirical distribution function cannot be observed in practice and it is obtained using an initial estimate $\hat{m}$ of the regression function. The sample residuals are then computed as:

$$\hat{u} = y - \hat{m}(z)$$

and then recentered, so that $\mathbb{E}(\hat{u}) = 0$. Bootstrap residuals, u^*, are finally obtained by sampling with replacement from the recentered $\hat{u}$. A bootstrap sample is then generated as follows:

$$y^* = \hat{m}(z) + u^*$$

For simplicity, we refer to this technique in the following as *naïve bootstrap*.

Resampling directly from the empirical distribution requires exchangeability of the residuals and thus homoskedasticity. The latter condition can be relaxed under the so-called *wild bootstrap* (see Härdle and Marron, 1991; Härdle and Mammen, 1993).

Under this framework, the i^{th} bootstrap error u_i^* is derived directly from the corresponding estimated residual $\hat{u}_i$. The new random variable u_i^* has a two point distribution $\hat{G}_i = \gamma\delta_a + (1-\gamma)\delta_b$, defined through the parameters γ, a and b, and where δ_a and δ_b denote point measures at a and b, respectively. The values of these parameters are computed so that the new random variable

matches the first three moments of the original residuals, i.e. $\mathbb{E}(u_i^*) = 0$, $\mathbb{E}(u_i^{*2}) = \hat{u}_i^2$, and $\mathbb{E}(u_i^{*3}) = \hat{u}_i^3$. Some algebra reveals that the parameters γ, a and b satisfying this property at each location are $\gamma = (5+\sqrt{5})/10$, $a = \hat{u}_i(1-\sqrt{5})/2$, and $b = \hat{u}_i(1+\sqrt{5})/2$.

3.5.2 Residuals in Nonparametric IV model

In the presence of endogeneity and when the regression function is estimated nonparametrically, bootstrap confidence intervals have been proposed by Chen and Pouzo (2012*a*), Horowitz and Lee (2012), and Santos (2012). While the first two papers solely deal with the case in which the function of interest is estimated using sieves, Santos (2012) presents a method which is of a more general interest and it is closely related to the one presented in this chapter. In fact, the approach we present is very simple to implement, and can be used irrespectively of the method applied to obtain the nonparametric estimator of φ. The theoretical properties of this bootstrap approach are not studied in this chapter and left to further research.

In nonparametric instrumental regressions, bootstrapping directly the residuals from the main structural equation, while it may work in practice, is theoretically flawed. This is because, direct sampling implies modifying the dependence structure between the endogenous covariate Z and the error term U.

An alternative approach, that has been undertaken by Sokullu (2010), is to bootstrap directly from the joint distribution of (Z, W). If we specify the following triangular model:

$$Y = \varphi(Z) + U \tag{3.5.1}$$

$$Z = g(W, V) \tag{3.5.2}$$

it would be possible, after estimation of the functions φ and g, to consistently estimate the errors U and V and then draw observations from their joint empirical distribution. However, this approach breaks down the basic rationale for using instrumental variables, which is exactly not to specify a functional relation between Z and W. Moreover, structural estimation of the function g in model (3.5.1) requires assumption on the error term V, which may not be satisfied in practice. Alternatively, we could take an additively separable form for the function g but this approach

seems more suited when the endogenous model is estimated using control functions.

An alternative procedure would be to sample from the residual of the statistical inverse problem. That is, define the errors in the following way:

$$\eta = r - T\varphi \tag{3.5.3}$$

By drawing from the error term η, we could generate bootstrap samples r^* and then estimate φ^* as the solution of the inverse problem:

$$r^* = T\varphi$$

However, the error in equation (3.5.3) is a functional residual. To consistently bootstrap from it, we can write its Fourier decomposition as follows:

$$\eta = \sum_{j=0}^{\infty} \frac{\langle \eta, \phi_j \rangle}{\lambda_j} \phi_j$$

We can then resample an iid sequence of Fourier coefficients and generate a bootstrap sample of the error term η from a truncated version of this infinite sum. While this approach is consistent with the inverse problem theory as framed in the context of nonparametric IV, both its application and its asymptotic properties are deemed intricate. The former requires to compute truncated sums of iid series to obtain a single realization of the bootstrap errors. This increases computational costs tremendously. The latter needs to make assumptions about the truncation parameter, that needs to diverge to infinity as the sample size increases. Moreover, the value of this truncation parameter should be objectively determined in finite samples.

To circumvent these issues, the approach proposed here is, instead, to resample residuals from the conditional moment equation obtained by projecting the dependent variable Y on the space spanned by the instruments W (see also Chen and Reiss, 2011; Florens and Simoni, 2012), i.e.:

$$\varepsilon = Y - \mathbb{E}(\varphi(Z)|W) \tag{3.5.4}$$

This model can be used to construct the sampling distribution of Y given the function φ. In the

spirit of Florens and Simoni (2012), we can redefine our operators as follows:

$$T_N : \mathbb{L}_Z^2 \to \mathbb{R}^N \tag{3.5.5}$$

$$T_N^* : \mathbb{R}^N \to \mathbb{L}_Z^2 \tag{3.5.6}$$

and the inverse problem would be the one defined by the sample counterpart of equation (3.5.4), i.e.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \mathbb{E}(\varphi(Z)|W = w_1) \\ \vdots \\ \mathbb{E}(\varphi(Z)|W = w_N) \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix} \tag{3.5.7}$$

$$y_{(N)} = T_N\varphi + \varepsilon_{(N)}, \qquad \text{with} \qquad y_{(N)}, \varepsilon_{(N)} \in \mathbb{R}^N \tag{3.5.8}$$

The residual component, $\varepsilon_{(N)}$, is thus not defined as a functional residual in this equation and it lays on a finite dimensional space. Hence, standard bootstrap techniques that draws directly from the empirical distribution of the residuals ε can be applied. Notice that this approach is much simpler than the direct bootstrap from equation (3.5.3). A potential criticism is that, resampling from (3.5.4), leads to bootstrap only the dependent variable Y and not the endogenous component Z. However, by the definition of the error term ε in (3.5.4), we have that:

$$Y^* = \mathbb{E}(\varphi(Z)|W) + \varepsilon^* = (\varphi(Z) + U)^*$$

Then, by holding constant the conditional expectation of φ given W, we are modifying the value of $\varphi(Z) + U$. Therefore, *we are changing the realization of the function φ and the error term U, simultaneously, for a given realization of the instrument W*. This appears to be equivalent to bootstrap directly from the joint distribution of the errors (U, V), as in (3.5.1), at least in some particular cases.

Example 3 (Linear simultaneous equations)**.** Consider the following triangular model:

$$Y = Z\beta + U$$

$$Z = \zeta(W) + V$$

where V is an random noise, such that $\mathbb{E}(V|W) = 0$ and V is correlated with U, so that Z is endogenous. Then, we have that:

$$\varepsilon = U + (Z - \zeta(W))\,\beta = U + V\beta$$

Therefore, bootstrap directly from the error ε is equivalent to bootstrap from the joint distribution of (U, V). ■

Furthermore, the mean independence condition, $\mathbb{E}(U|W) = 0$, guarantees that the projected residuals are not related to the regressors. However, the estimated residual from (3.5.4) is, by definition of conditional expectation, a function of the instruments W. In general, it is not possible to suppose this function to be constant and, therefore, wild bootstrap is advocated here, in order to cope with this source of heteroskedasticity.[9]

A further advantage of using wild bootstrap in this general nonparametric context is that it correctly takes into account the bias arising from the nonparametric estimation, without requiring additional estimation of the latter or a suboptimal (undersmoothed) curve estimation (Härdle and Marron, 1991). The nonparametric bootstrap advocated in, e.g., Chen and Pouzo (2012*b*) may be suboptimal in the general case, as the bootstrap bias arising from the nonparametric estimation will be equal to 0. Morover, bootstrapping from the residuals allows to sample under the null of a statistical test, if, for instance, the researcher would like to test for a particular shape of the function φ.

Call $\hat{T}$ the estimated conditional expectation operator, acting onto the space spanned by W. The estimated residuals are defined as follows:

$$\hat{\varepsilon}_i(w) = y_i - \hat{T}\hat{\varphi}(z_i) \quad \forall i = 1, \ldots, N$$

Define further the bootstrap residual $\varepsilon_i^*(w)$ which is drawn with probability γ from the two point distribution $\hat{G}_i$, with realizations $a(w) = \hat{\varepsilon}_i(w)(1 - \sqrt{5})/2$, and $b(w) = \hat{\varepsilon}_i(w)(1 + \sqrt{5})/2$. This

[9] We are aware that, despite its flexibility, wild bootstrap may cause greater variability and, ultimately, undercoverage. We do not explore this point further in the chapter. Interested readers are referred to Kauermann and Carroll (2001) and Kauermann et al. (2009).

residual is ultimately used to construct bootstrap observations as follows:

$$y^* = \hat{T}\hat{\varphi}(z) + \varepsilon^*(w)$$

A bootstrap estimator, $\varphi^*(z)$, is then obtained by solving the inverse problem:

$$\hat{T}\varphi = r^*$$

with $r^* = \hat{T}y^*$. In order to retrieve the bootstrap estimator, smoothing parameters for the nonparametric estimation of the conditional expectation operators are held constant. While in the theory of wild bootstrap (Härdle and Marron, 1991; Ferraty et al., 2010), it is generally required to simulate bootstrap observations y^*, from an oversmoothed version of the regression function, simulation studies often suggest the usage of the same bandwidth. Oversmoothing seems particularly difficult to implement, as it is not clear, to the best of our knowledge, how to practically choose the new smoothing constant.

The regularization parameter is also held fixed. However, in order to match the asymptotic distribution, we need to deal with the specific features of each regularization procedure. In particular, it is important to notice that it would be impossible to match the asymptotic distribution of the nonparametric IV estimator when the regularization bias converges to 0. In finite samples, it is known that regularization methods lead to a bias. Therefore, the bootstrap has to match the distribution of each estimator around the *biased* version of the true value of the function.

(i) **TK**: For a fixed value of the regularization parameter α, an asymptotic bias arises in the distribution of the estimator (Carrasco et al., 2013). Confidence intervals have to be recentred according to this bias. We know that (see Darolles et al., 2011*a*):

$$\varphi^\alpha - \varphi = -\alpha \left(\alpha I + T^*T\right)^{-1} T^*T\varphi$$

Hence, we have that:

$$\hat{\varphi}^\alpha - \varphi^\alpha = \hat{\varphi}^\alpha - \varphi + \alpha \left(\alpha I + T^*T\right)^{-1} T^*T\varphi \tag{3.5.9}$$

which is the object whose distribution we would like to match.

If we replace φ, T, T^*, and α with their sample counterparts, and $\hat{\varphi}^{\alpha}$ with the bootstrap estimator $\hat{\varphi}^{*\alpha}$, we can approximate the object in (3.5.9) by:

$$\hat{\varphi}^{*\alpha} - \hat{\varphi}^{\alpha} + \alpha_N \left(\alpha_N I + \hat{T}^*\hat{T}\right)^{-1} \hat{T}^*\hat{T}\hat{\varphi}^{\alpha} \tag{3.5.10}$$

(ii) **LF**: The LF estimation is tantamount to TK regularization as long as the number of iterations is asymptotically proportional to the inverse of the α parameter, i.e. $M \approx 1/\alpha$. Therefore, the LF estimator is unbiased as M goes to infinity, i.e.:

$$\|\varphi^M - \varphi\| = \left\| c \sum_{k=0}^{M-1} \left(I - cT^*T\right)^k T^*T\varphi - \varphi \right\| \xrightarrow[M\to\infty]{} 0$$

For a fixed finite number of iterations M, there exists again a regularization bias. The object, whose asymptotic distribution is studied is, as before:

$$\hat{\varphi}^M - \varphi^M = \hat{\varphi}^M - \varphi + c \sum_{k=0}^{M-1} \left(I - cT^*T\right)^k T^*T\varphi \tag{3.5.11}$$

This object can be approximated as above by replacing φ, T, T^*, and M with their sample counterparts, and $\hat{\varphi}^M$ with the bootstrap estimator $\hat{\varphi}^{*M}$.

(iii) **GK**: In this case, the regularization is achieved by the truncation of the basis, so that, for any basis of order J, we have:

$$\|\varphi^J - \varphi\| = \left\| \sum_{k=J+1}^{\infty} \lambda_j \kappa_j \varphi_j \right\|$$

However, it is not possible to control explicitly for this bias. In fact,

$$\|\varphi^J - \varphi\| = \left\| \mathcal{Z} \left(\mathcal{Z}'\mathcal{W}\mathcal{W}'\mathcal{Z}\right)^{-1} \mathcal{Z}'\mathcal{W}\mathcal{W}'\mathcal{Z}\beta - \varphi \right\|$$

is identically equal to zero for any fixed value of J, and would require the computation of the entire series for $J \to \infty$, which is clearly unfeasible. In this case, we therefore simply apply wild bootstrap to the residuals without correcting for the estimated regularization bias (see

Horowitz and Lee, 2012, for a different approach to bootstrap).

In order to show the validity of our bootstrap procedure, we compare the distribution of the estimator of φ obtained using the Monte-Carlo simulations in the previous section with the distribution obtained over each bootstrap replication, given the values of the smoothing and the regularization parameters.

Since properties of the bootstrap and coverage probabilities are given pointwise, we evaluate the properties of the bootstrap for 7 values of the endogenous variable Z. In particular, we select a vector Q of values of Z, which contains percentiles 1, 5, 25, 50, 75, 95, and 99. To facilitate comparison, all distributions are standardized. With a slight abuse of notations, we thus denote by φ the value of the function, for a particular realization of the endogenous variable Z.

The comparison between the simulated density, $f(\varphi)$, of $\hat{\varphi} - \varphi$, and the bootstrap densities, $f^*(\varphi)$ of $\hat{\varphi}^* - \hat{\varphi}$, at each point Q, is carried in the following way.

(i) For each simulated sample, we compute the value of the smoothing and the regularization parameter and we construct 1000 bootstrap estimators, $\hat{\varphi}^*$, obtaining a matrix of size $N \times 1000$.

(ii) We keep the information about the matrix of bootstrap estimate $\hat{\varphi}^*$ for the elements in Q.

(iii) We repeat steps (i) and (ii) for 1000 simulated samples.

(iv) For each element of the vector Q, we obtain a matrix of bootstrap values of size 1000×1000, where the smoothing and the regularization parameters are constant across columns.

(v) We obtain bootstrap densities, $f^*(\varphi)$, from the row elements of this matrix and we compare them with the simulated density, $f(\varphi)$.

In order to obtain an objective measure of distance between these objects, we compute absolute deviations between an appropriate nonparametric estimator of the former and the latter density.[10] We use standard Gaussian kernels where the optimal bandwidth for $\hat{f}(\varphi)$ is computed using maximum likelihood cross validation and it is held constant for $\hat{f}^*(\varphi)$.

[10] See also Ferraty et al. (2010), for a similar approach to the validity of bootstrap.

In particular, we use the total variational distance as reference measure (Liese and Vajda, 2006). This measure is defined as follows:

$$TV_{\varphi} = \frac{1}{2} \int |f^*(\varphi) - f(\varphi)| d\varphi$$

Figures (3.9), (3.10), (3.11), (3.12), (3.13), (3.14), (3.15), (3.16) and (3.17) present the comparison between the density of the estimator $\hat{\varphi}$ at each point of the vector Q (where the median has been excluded for ease of presentation). The thin gray lines represent the densities obtained by bootstrap; while the dashed thick black line is the distribution obtained from the simulations. It appears clearly that the simulated errors can be fairly well approximated by the bootstrapped errors.

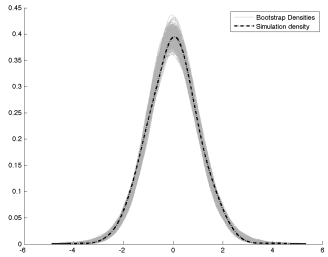
	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
Local Constant TK	0.0261	0.0253	0.0451	0.0279	0.0270	0.0381	0.0418
Local Linear TK	0.0710	0.0799	0.0783	0.1979	0.1073	0.0355	0.0664
Spline TK	0.0205	0.0183	0.0730	0.0520	0.0677	0.0541	0.0431
Penalized TK	0.0236	0.0228	0.0298	0.0475	0.0668	0.0211	0.0297
GK	0.0659	0.0737	0.0313	0.0748	0.0505	0.0649	0.0747
Local Constant LF	0.0273	0.0237	0.0328	0.0410	0.0227	0.0284	0.0373
Local Linear LF	0.0307	0.0420	0.0414	0.0604	0.0848	0.0620	0.0392
Spline LF	0.0603	0.0811	0.0689	0.1098	0.0417	0.0508	0.0953
Penalized LF	0.0563	0.0565	0.0734	0.0521	0.0546	0.0475	0.0470

Table 3.4: Median Variational Distance at each point of the vector Q.

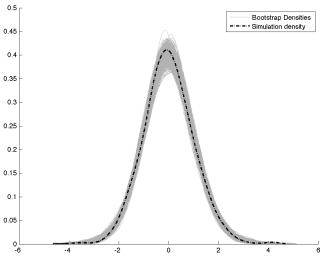
Finally, Table (3.4) reports the median value of the variational distance for each element in Q.[11] The median variational distance is below 0.1 for the majority of the estimators and it therefore confirms that the bootstrap density approximates the true density fairly well. However, its performance deteriorates in the case of GK regularization. Also, in the case of Local Linear TH, the variational distance seems to increase around the median. However, its values remain under 0.3, which can be considered as being reasonable in this setting (see also Ferraty et al., 2010).

To conclude, we present pointwise coverage probabilities for the bootstrap for each value in Q and the usual nominal values for confidence bands: 90%, 95%, and 99%. Table (3.5) reports the median value of coverage probabilities for each one of the estimators considered in this work. It is clear that the confidence bands obtained by bootstrap cover the true function very well and

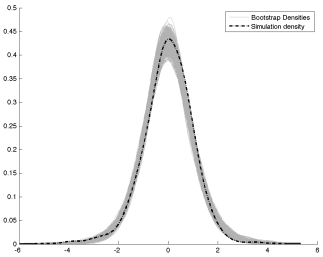
[11] Figures (3.23), (3.24), (3.25), (3.26) and (3.27) in the Appendix report also a box plot comparison of Total Variational Distance.



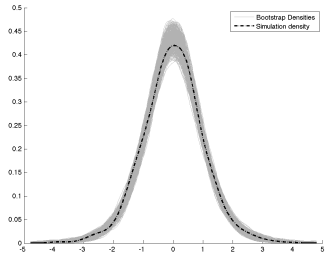
(a) 1% percentile



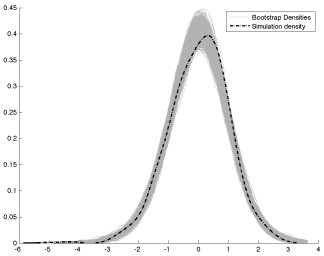
(b) 5% percentile



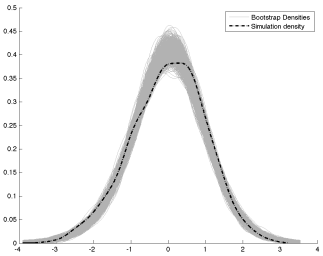
(c) 25% percentile



(d) 75% percentile

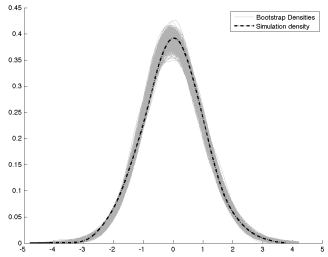


(e) 95% percentile

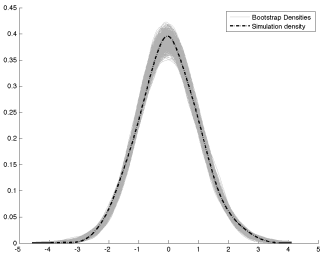


(f) 99% percentile

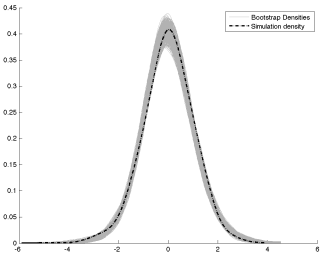
Figure 3.9: Simulation vs Bootstrap Densities for Local Constant Tikhonov.



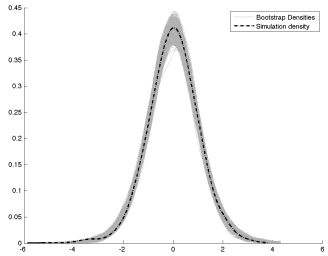
(a) 1% percentile



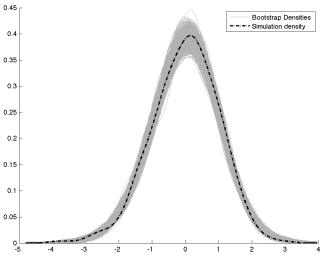
(b) 5% percentile



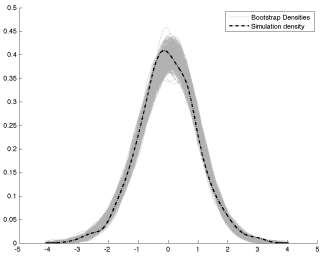
(c) 25% percentile



(d) 75% percentile

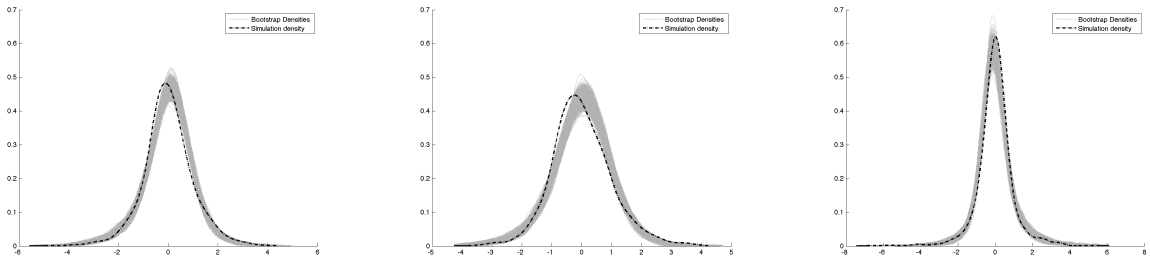


(e) 95% percentile

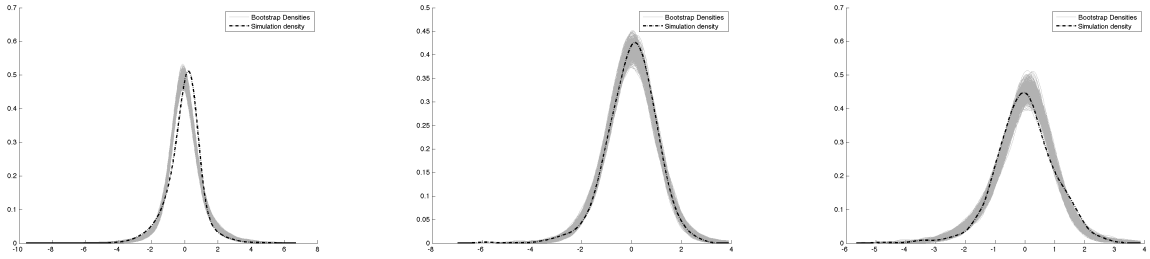


(f) 99% percentile

Figure 3.10: Simulation vs Bootstrap Densities for Local Constant Landweber-Fridman.

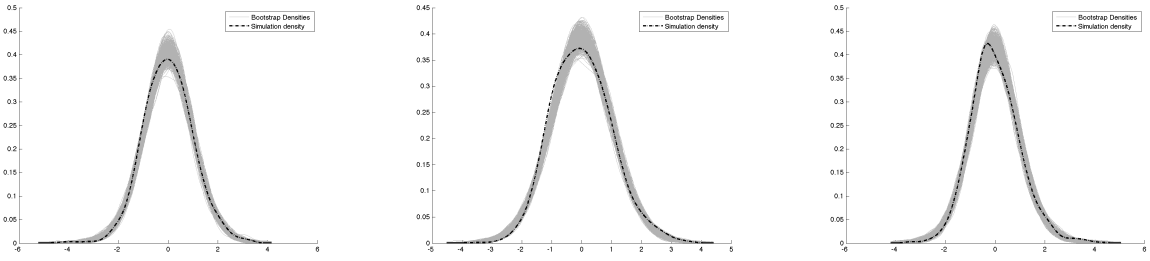


(a) 1% percentile (b) 5% percentile (c) 25% percentile

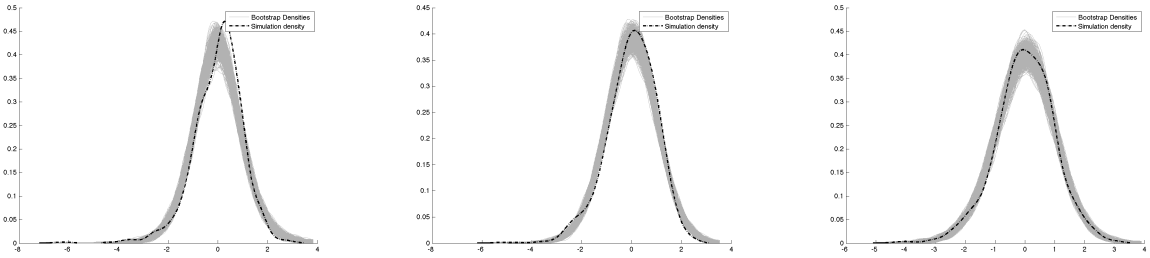


(d) 75% percentile (e) 95% percentile (f) 99% percentile

Figure 3.11: Simulation vs Bootstrap Densities for Local Linear Tikhonov.

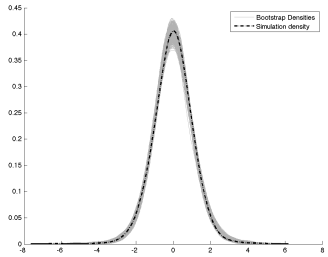


(a) 1% percentile (b) 5% percentile (c) 25% percentile

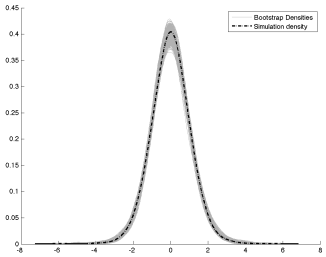


(d) 75% percentile (e) 95% percentile (f) 99% percentile

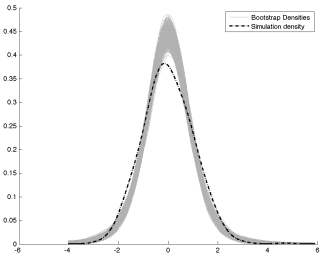
Figure 3.12: Simulation vs Bootstrap Densities for Local Linear Landweber-Fridman.



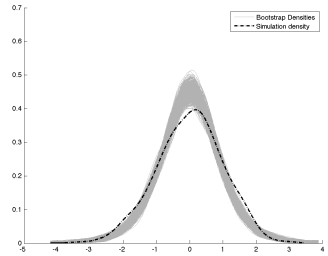
(a) 1% percentile



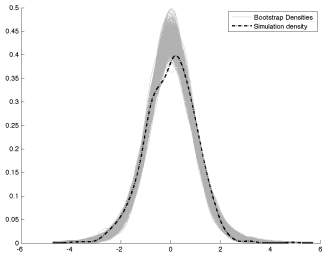
(b) 5% percentile



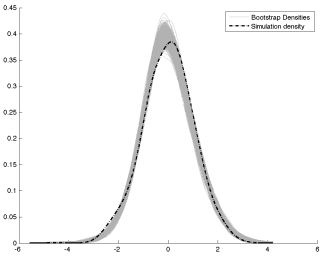
(c) 25% percentile



(d) 75% percentile

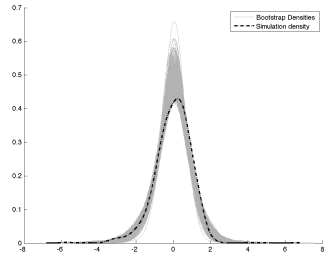


(e) 95% percentile

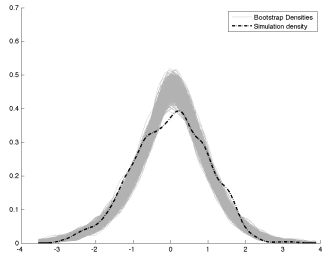


(f) 99% percentile

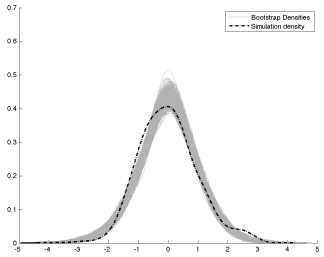
Figure 3.13: Simulation vs Bootstrap Densities for Spline Tikhonov.



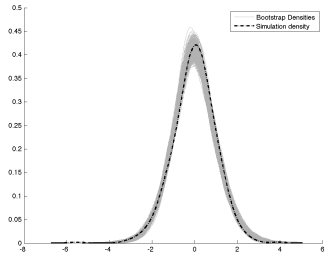
(a) 1% percentile



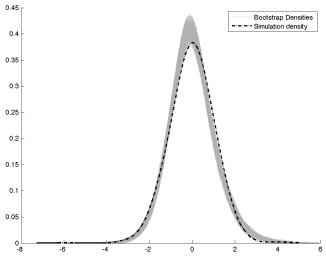
(b) 5% percentile



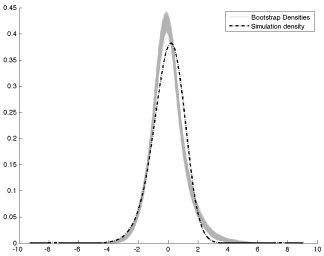
(c) 25% percentile



(d) 75% percentile

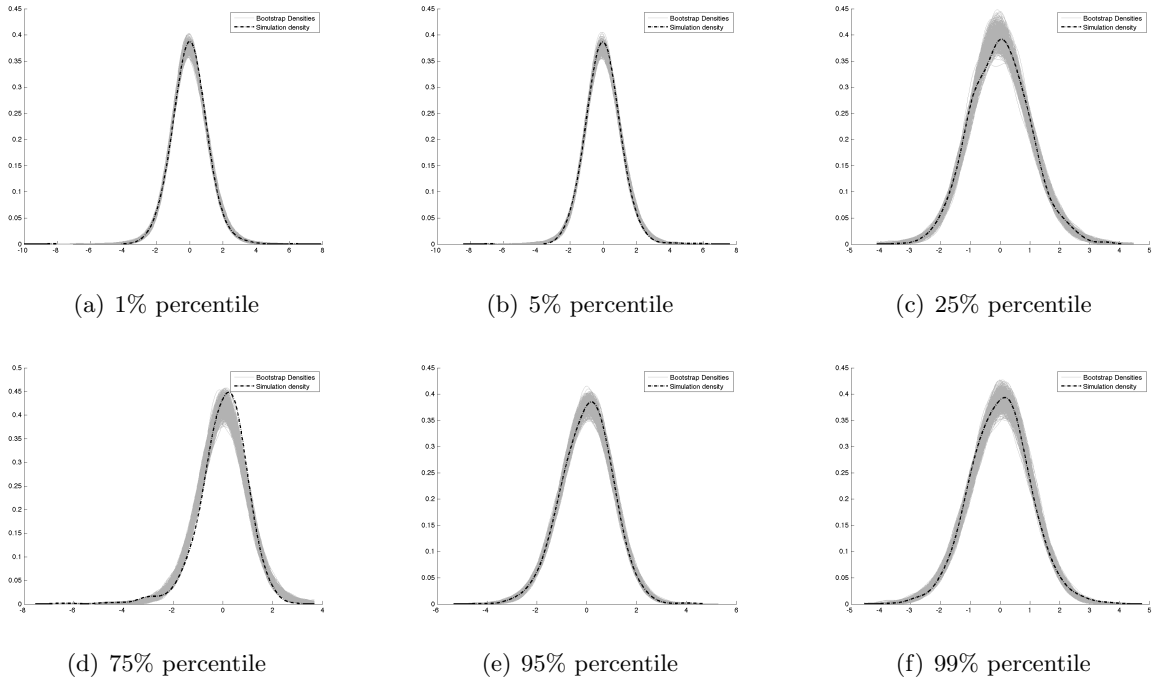


(e) 95% percentile



(f) 99% percentile

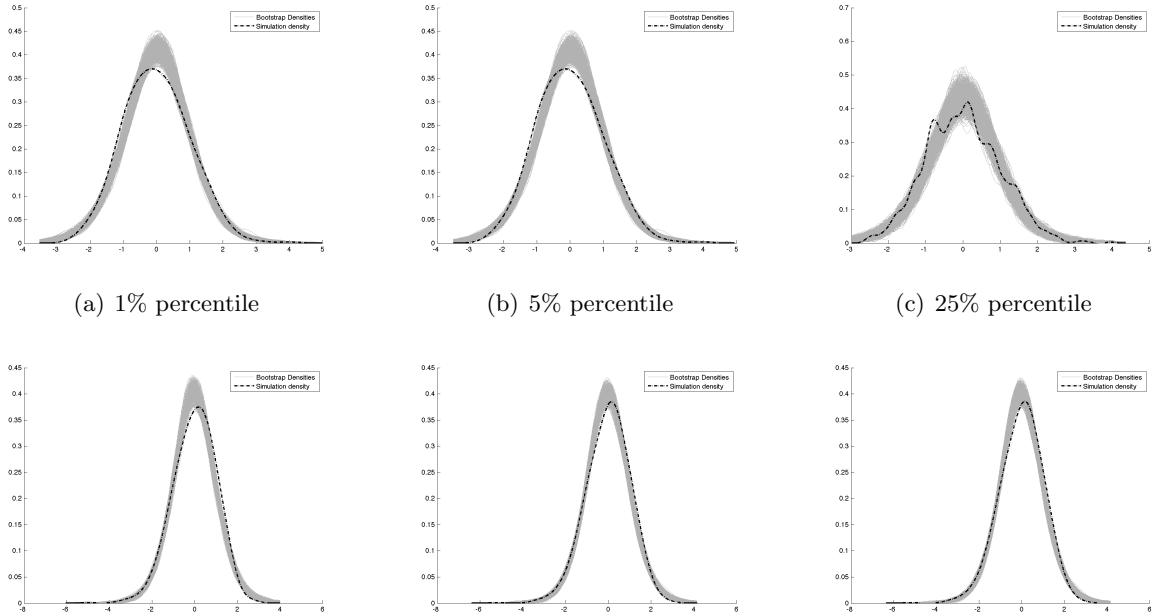
Figure 3.14: Simulation vs Bootstrap Densities for Spline Landweber-Fridman.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

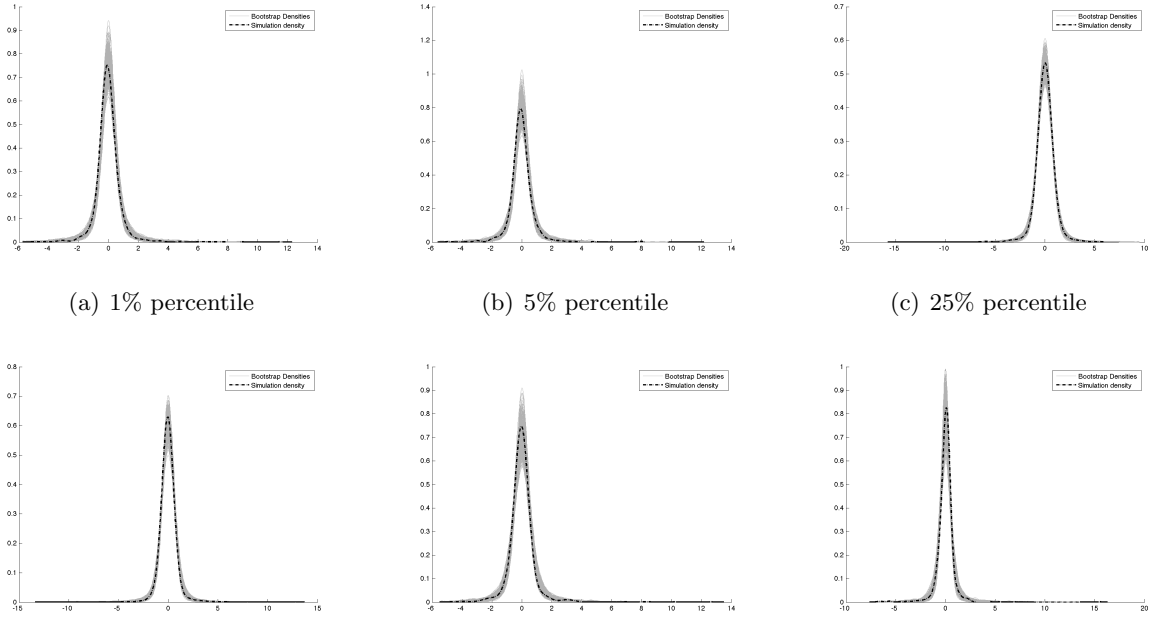
Figure 3.15: Simulation vs Bootstrap Densities for Local Constant Tikhonov with Penalized first derivative.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

Figure 3.16: Simulation vs Bootstrap Densities for Local Constant Landweber-Fridman with Penalized first derivative.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

Figure 3.17: Simulation vs Bootstrap Densities for Splines Galerkin.

that the bootstrap probabilities are very close to the nominal ones. This demonstrates further the applicability and the good properties of wild bootstrap to obtain pointwise confidence bands in the case of nonparametric models estimated with instrumental variables.

3.6 An empirical application: estimation of the Engel curve for food in rural Pakistan

In this last section, we present an empirical application to the estimation of the Engel curve for food. 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 therefore sufficient to regress the share of the household's budget spent for a given commodity (or group of commodities) over the total 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 re-

		Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
	Local Constant TK	0.8940	0.9030	0.9090	0.8970	0.9050	0.9080	0.8980
	Local Linear TK	0.8920	0.8980	0.9020	0.8810	0.8930	0.9100	0.8970
	Spline TK	0.9000	0.9040	0.8730	0.9110	0.8730	0.8950	0.8920
	Penalized TK	0.9070	0.9060	0.8935	0.9000	0.9200	0.9070	0.9000
90%	GK	0.9110	0.9060	0.9040	0.9110	0.9170	0.9140	0.9070
	Local Constant LF	0.8950	0.8970	0.9030	0.9030	0.9050	0.9090	0.8970
	Local Linear LF	0.9040	0.9090	0.9020	0.9070	0.9110	0.9070	0.8900
	Spline LF	0.9220	0.9030	0.9130	0.9550	0.9100	0.9030	0.9220
	Penalized LF	0.9010	0.9030	0.8970	0.8970	0.9140	0.9100	0.9080
	Local Constant TK	0.9560	0.9530	0.9560	0.9470	0.9500	0.9490	0.9500
	Local Linear TK	0.9450	0.9450	0.9440	0.9440	0.9470	0.9550	0.9500
	Spline TK	0.9560	0.9560	0.9610	0.9520	0.9550	0.9540	0.9430
	Penalized TK	0.9550	0.9560	0.9480	0.9530	0.9560	0.9490	0.9490
95%	GK	0.9530	0.9530	0.9480	0.9550	0.9540	0.9510	0.9510
	Local Constant LF	0.9430	0.9480	0.9530	0.9430	0.9510	0.9550	0.9450
	Local Linear LF	0.9500	0.9510	0.9500	0.9435	0.9560	0.9480	0.9430
	Spline LF	0.9620	0.9630	0.9460	0.9730	0.9530	0.9540	0.9560
	Penalized LF	0.9590	0.9590	0.9590	0.9560	0.9610	0.9560	0.9560
	Local Constant TK	0.9940	0.9900	0.9840	0.9910	0.9910	0.9930	0.9960
	Local Linear TK	0.9870	0.9870	0.9870	0.9910	0.9910	0.9890	0.9900
	Spline TK	0.9890	0.9890	0.9970	0.9870	0.9970	0.9930	0.9950
	Penalized TK	0.9910	0.9930	0.9920	0.9910	0.9820	0.9880	0.9840
99%	GK	0.9890	0.9890	0.9880	0.9920	0.9880	0.9910	0.9880
	Local Constant LF	0.9930	0.9940	0.9890	0.9930	0.9870	0.9880	0.9860
	Local Linear LF	0.9880	0.9900	0.9870	0.9860	0.9860	0.9910	0.9880
	Spline LF	0.9890	0.9970	0.9840	0.9880	0.9880	0.9890	0.9860
	Penalized LF	0.9940	0.9940	0.9960	0.9970	0.9920	0.9920	0.9920

Table 3.5: Pointwise coverage probabilies of wild bootstrap.

gressor. Blundell et al. (2007) suggest using other sources of income as a suitable instrument for total expenditure.

In the following, to simplify notation, we denote by the random variable Y, the share of expenditure in a given consumption good; by Z, the total log expenditure of the household; and, by 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 household. This specification allows for a non-parametric modelling of the endogenous variable Z, 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) uses a partially linear model, in which Z 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 Z. They claim that, when using a control function approach with W as 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 maintain a high level of simplicity and we model the relationship as follows:

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

where φ represents the shape of the Engel curve. Since our simplified model ignores specific household and geographical characteristics, we reduce heterogeneity by considering only 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. We therefore end up using a sample of 5691 observations.

In our database, food, as a broad aggregate of 82 commodities, accounts on average for about 51% of the total household expenditure in Punjab (see table 3.6).

In the original work of Bhalotra and Attfield (1998), it is shown that the Engel curve for food it is decreasing, as predicted by Engel's law, and has a quadratic shape. This latter result is of great

	Mean	St.Dev	Min	Max
Log PC Expenditure	5.61	0.49	4.22	8.07
Log PC Income	5.63	0.52	3.98	8.00
Budget share food	0.51	0.10	0.07	0.83

Table 3.6: Summary statistics

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 *test* the robustness of previous results and provide some additional evidence using our simplified nonparametric instrumental variable approach. To compare our fully nonparametric specification with a quadratic model which also takes into account the endogeneity issue, we consider the following model, which is estimated using a control function approach:

$$Y = \beta_1 Z + \beta_2 Z^2 + \gamma V + U \tag{3.6.1}$$

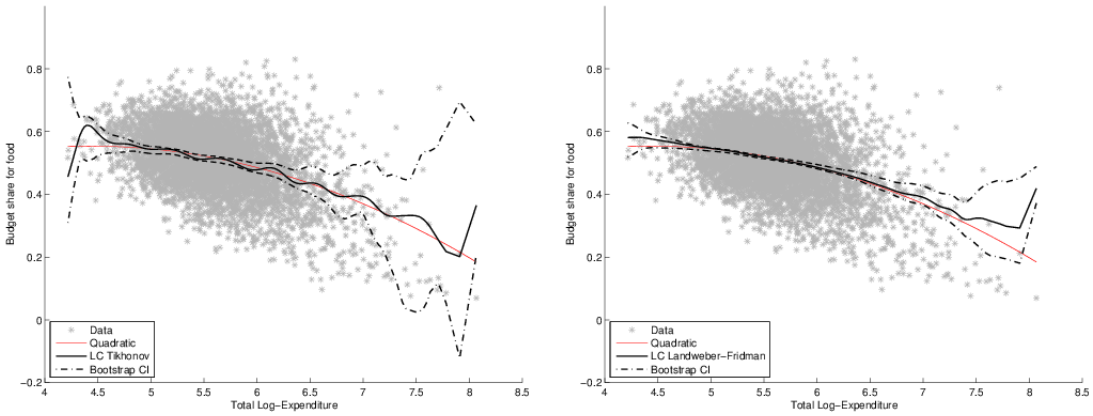
$$Z = \zeta(W) + V \tag{3.6.2}$$

$$\mathbb{E}\left(U|W,V\right) = \mathbb{E}\left(U|V\right) \tag{3.6.3}$$

The link function ζ is estimated using local constant kernels and cross validation bandwidth. The coefficients $(\beta_1, \beta_2, \gamma)$ are instead estimated using simple OLS. The results are summarized in table (3.7). We can see that all coefficients are significant. The one associated with the quadratic component is very small but significantly negative.

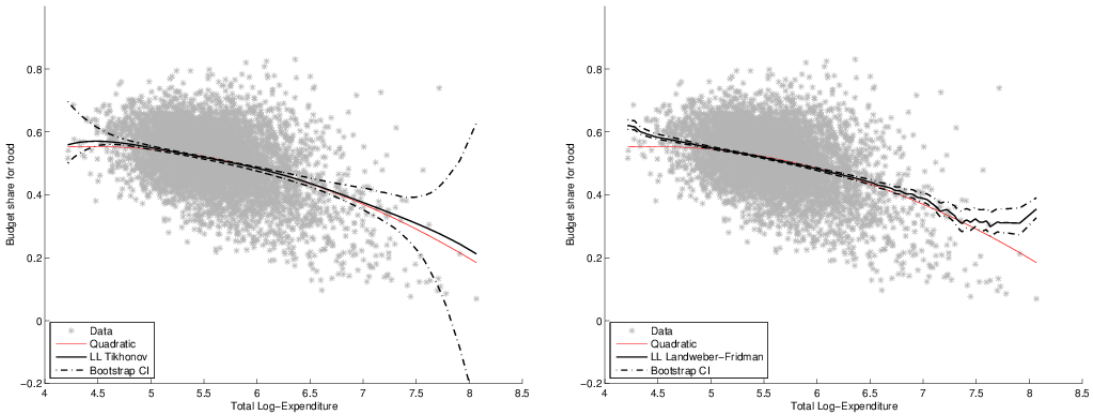
The results of the estimation of the Engel curve for Pakistan data are reported in Figures (3.18), (3.19), (3.20), (3.21) and (3.22). For each kind of nonparametric estimator (local constant, local linear, B-splines and penalized local constant), we present the outcome both using TK and LF regularizations. The final figure (3.22) draws the GK estimator that uses B-spline bases. For each figure, we also consider the 95% bootstrap confidence intervals and we draw the quadratic fitting obtained using the control function approach in (3.6.1).

The results are widely consistent across the various frameworks. Note that the local constant estimation coupled with TK regularization does not give visually nice results. This can be due to the fact that optimal regularization parameter is under-regularizing, which causes the *bumps* in



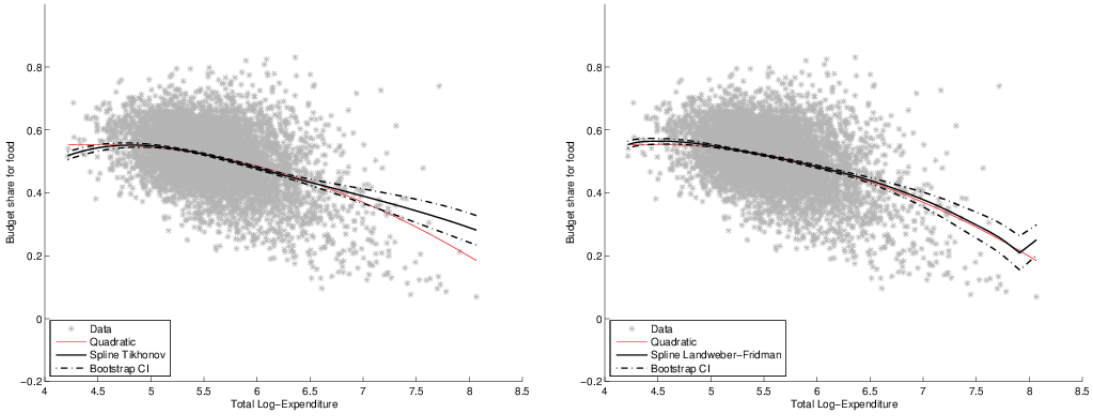
(a) Tikhonov (b) Landweber-Fridman

Figure 3.18: Estimation of the Engel Curve for food (local constant)



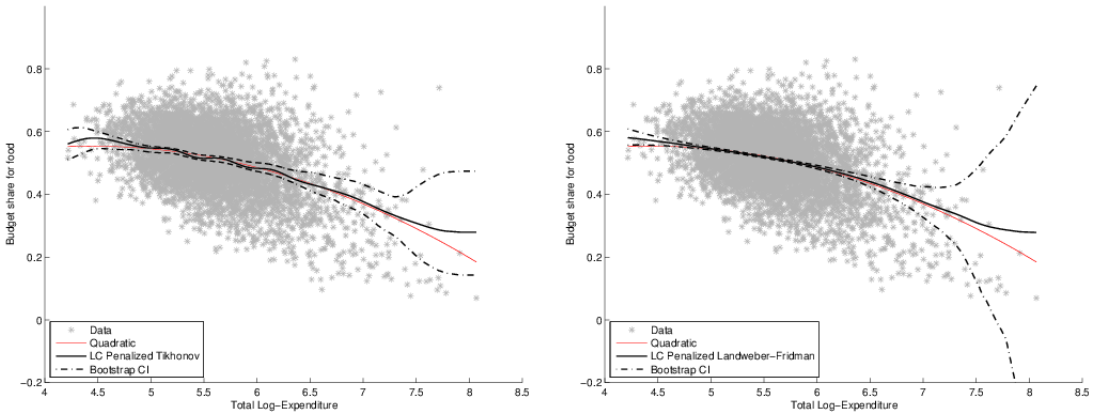
(a) Tikhonov (b) Landweber-Fridman

Figure 3.19: Estimation of the Engel Curve for food (local linear)



(a) Tikhonov (b) Landweber-Fridman

Figure 3.20: Estimation of the Engel Curve for food (splines)



(a) Tikhonov (b) Landweber-Fridman

Figure 3.21: Estimation of the Engel Curve for food (Penalized local constant)

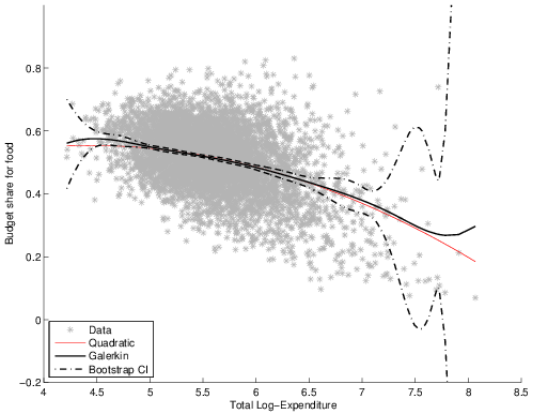


Figure 3.22: Galerkin estimation of the Engel Curve for food

the estimated regression function. It is also instructive to observe that these bumps disappear in figure (3.21), right panel, when we are penalizing the first derivative instead. This gives a much smoother solution for the regression function. Another important computational aspect to stress is that, as mentioned above, LF regularization holds the advantage of not requiring the inversion of the large data matrix and therefore can be a more appealing solution than TK in this case. However, computational time might increase because of the numerical update of the smoothing parameters at each iteration. This makes the two estimators, at least with our sample size, roughly comparable in terms of computational time.

However, the most interesting information is that the nonparametric estimators are not unanimously suggesting a quadratic relation between the total budget and the food share in rural Pakistan. The quadratic specification cannot be rejected at the 95% level by the majority of our

	Log PC Expenditure	Log PC Expenditure Sq	$\hat{V}$
Coefficient	0.245	-0.028	-0.087
Std Error	0.0027	0.0005	0.0063

Table 3.7: Results from model (3.6.1). Dependent variable: share of budget for food.

models. This result is largely partial and does not control for the heterogeneity in our sample. Nonetheless, we stress here that, even a simple nonparametric estimation which controls for the possible endogeneity of the total budget, could be used as an indirect test to support a given parametric model.

3.7 Conclusions

This chapter presents a deep investigation 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 provide a useful review of the data driven approaches that have been proposed so far for the selection of the regularization parameter. A simple and valid approach for obtaining pointwise bootstrap confidence intervals is also discussed and its properties derived by means of simulations. Finally, an application to the estimation of the Engel curve for food, in a sample of household in rural Pakistan shows its practical usefulness.

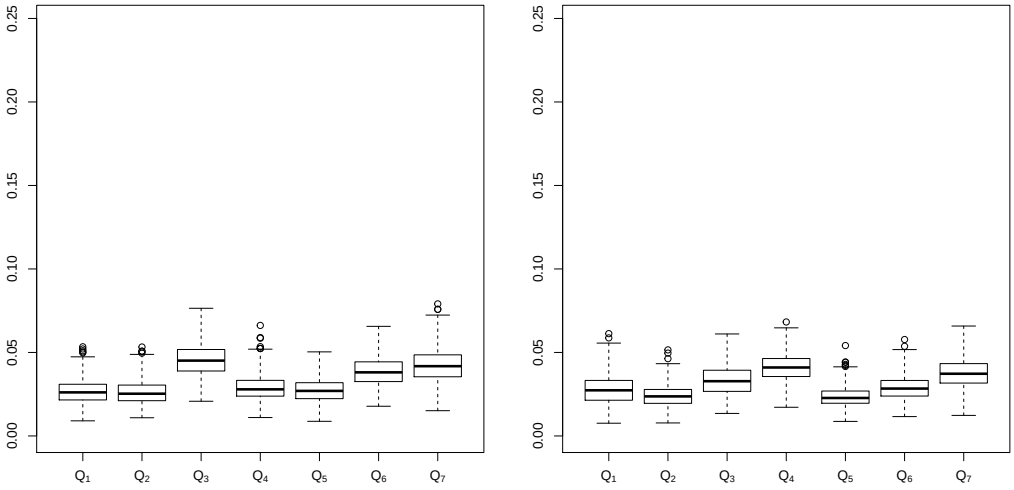
Our intention is to give a unified and simple presentation of the several regularization procedures that can be considered when applied researchers would like to keep the flexibility of nonparametric estimation in presence of endogenous regressors. Our aim is to narrow the gap between the theoretical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

Without delving further into the specific matter of the estimation of the Engel curve, we point out the relevance of the use of nonparametric instrumental regressions, and, more in general, of nonparametric methods, in applied studies. Despite the fact that parametric model are faster to compute and easier to present to the general audience, they may lay on assumptions about the function of interest that can reveal to be unrealistic and may ultimately add more structural

information than the data themselves. This can ultimately lead to substantially different results and hence conclusions in terms of policy considerations and inference about the behavior of economic agents. Moreover, computational issues for nonparametric estimators do not seem to be relevant anymore, and a variety of semiparametric structures can be used in order to ease computational burden, control for heterogeneity in the sample, and obtain parametric rate of convergence (Blundell et al., 2007; Florens et al., 2012).

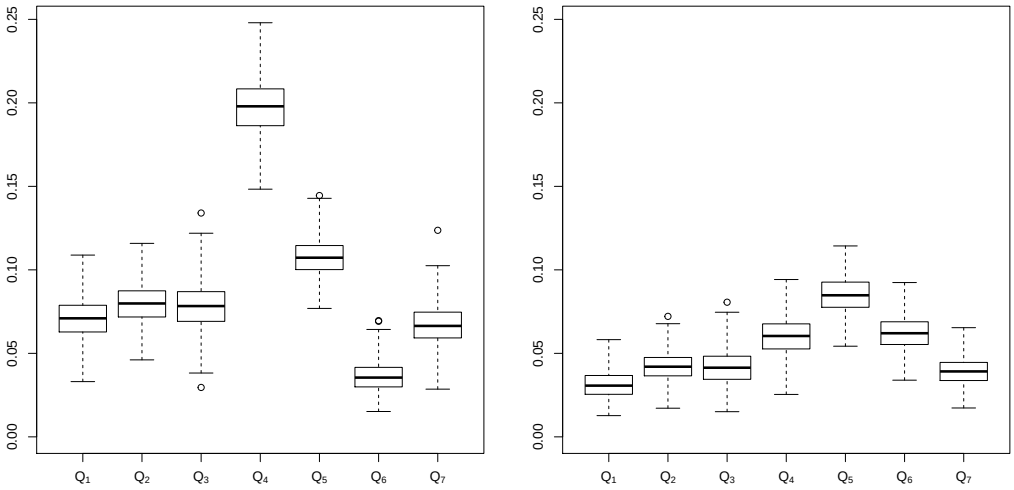
Our analysis is partial, as we do not explore the properties of the various estimators under several simulation schemes and several degrees of *ill-posedness* of the inverse problem. However, we see this work as a useful first step to make nonparametric instrumental regression readily available to applied economists and econometricians.

3.8 Appendix



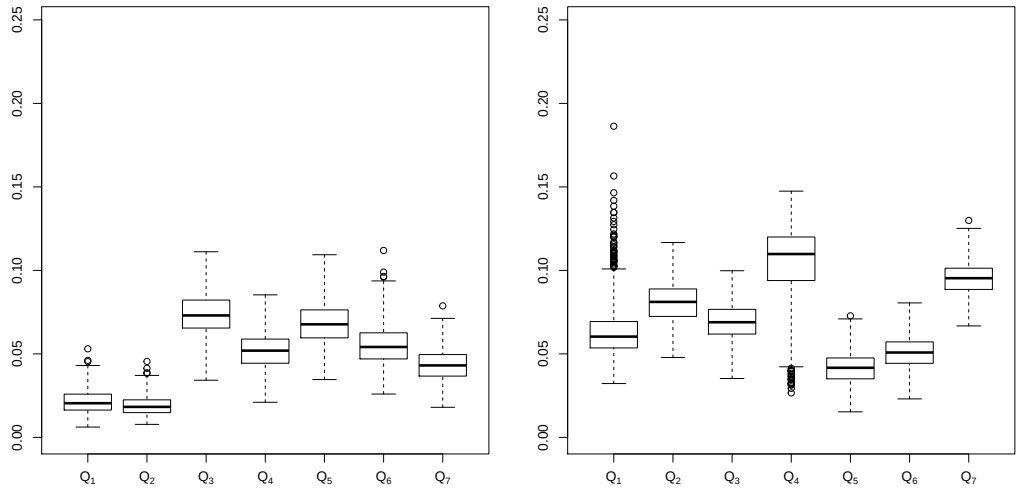
(a) Tikhonov (b) Landweber-Fridman

Figure 3.23: Box plot Total Variational Distance, Local Constant Kernels.



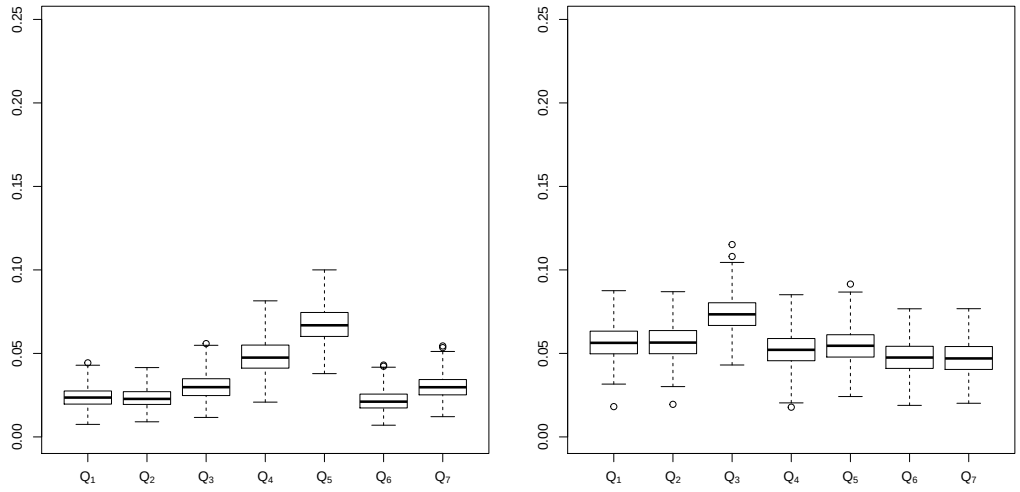
(a) Tikhonov (b) Landweber-Fridman

Figure 3.24: Box plot Total Variational Distance, Local Linear Kernels.



(a) Tikhonov (b) Landweber-Fridman

Figure 3.25: Box plot Total Variational Distance, B-Splines.



(a) Tikhonov (b) Landweber-Fridman

Figure 3.26: Box plot Total Variational Distance, Penalized Local Constant Kernels.

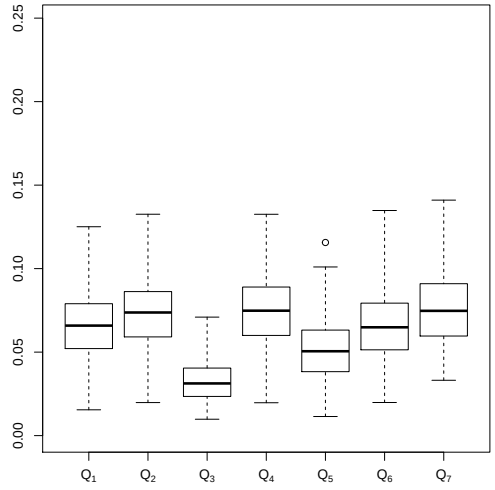


Figure 3.27: Box plot Total Variational Distance, Galerkin.

Final Conclusions

Research is a very lengthy book in which the introduction is very slow and the core is exciting, full of answers, but also of unsolved matters. As we proceed to the next chapter, we may find some new answers and solutions but we are left with new and exciting issues we want to face.

This thesis contributes to the literature on nonparametric estimation in additive separable regression models with endogeneity.

We provide a set of new tools for the data-driven choice of the regularization parameter and for obtaining pointwise confidence intervals using wild bootstrap. Moreover, we extend the current framework to embed the case in which only a binary transformation of the dependent variable is observed.

Of the many issues tackled in this work, we have probably only scratched the surface and future research can proceed in several directions.

Although the literature on nonparametric instrumental regressions is very much established at the moment, many aspects could be further developed. The properties and the validity of the wild bootstrap explored in Chapter 3 need to be analytically derived. Moreover, some further steps are required to make the model more handy for applied researcher. As a matter of fact, regression models in applied microeconometrics often include many control variables as heterogeneity in the sample is extremely important. Beside the partially linear specification studied in Florens et al. (2012), there is not a straightforward way to include exogenous regressors in the picture. Considering a nonseparable function of both endogeneous and exogenous regressors can become very cumbersome in presence of many exogenous variables, although the curse of dimensionality can be mitigated by using infinite order polynomial regressions as studied in Hall and Racine (2013).

An additive separable nonparametric structure, estimated using backfitting techniques could be a nice and viable solution to this problem; although an interesting line of research would be to study the estimation of nonparametric instrumental models with exogenous regressors using infinite order

polynomials.

Finally, the selection of the regularization parameter should be extended to the case of more practical relevance in which we choose two different bandwidths for the estimation of the conditional expectation operator and its adjoint. The theory has to be revised to allow for this more general case. Furthermore, new techniques on linear optimization could leave room for the simultaneous selection of the bandwidth and the regularization parameter.

Bibliography

Ahn, H., Ichimura, H. and Powell, J. (2004), Simple Estimators for Monotone Index Models, Manuscript, Department of Economics, UC Berkeley. 65

Andrews, D. W. K. (2011), 'Examples of L^2-Complete and Boundedly-Complete Distributions', *Cowles Foundation Discussion Paper* **1801**. 15, 90

Banks, J., Blundell, R. and Lewbel, A. (1997), 'Quadratic Engel Curves and Consumer Demand', *Review of Economics and Statistics* **79**(4), pp. 527–539. 9

Bhalotra, S. and Attfield, C. (1998), 'Intrahousehold Resource Allocation in Rural Pakistan: a Semiparametric Analysis', *Journal of Applied Econometrics* **13**(5), 463–480. 121, 123

Blanchard, G. and Mathé, P. (2012), 'Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration', *Inverse Problems* **28**(11), 1–24. 27

Blundell, R., Chen, X. and Kristensen, D. (2007), 'Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves', *Econometrica* **75**(6), 1613–1669. 8, 9, 41, 44, 86, 87, 96, 97, 121, 123, 124, 128

Blundell, R., Dearden, L. and Sianesi, B. (2005), 'Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **168**(3), 473–512. 86

Blundell, R. and Horowitz, J. (2007), 'A Non-Parametric Test of Exogeneity', *Review of Economic Studies* **74**(4), 1035–1058. 43, 88

Blundell, R. W. and Powell, J. L. (2004), 'Endogeneity in Semiparametric Binary Response Models', *Review of Economic Studies* **71**, 655–679. 64

Breunig, C. and Johannes, J. (2011), 'Adaptive Estimation of Functionals in Nonparametric Instrumental Regressions', *Mimeo* . 12

Burda, M. C. (1993), 'The determinants of East-West German migration: Some first results', *European Economic Review* **37**(2-3), 452 – 461. 80

Canay, I. A., Santos, A. and Shaikh, A. M. (2013), 'On the Testability of Identification in Some Nonparametric Models with Endogeneity', *Econometrica* (Forthcoming). 15

Cao-Abad, R. (1991), 'Rate of Convergence for the Wild Bootstrap in Nonparametric Regression', *The Annals of Statistics* **19**(4), pp. 2226–2231. 108

Cardot, H. and Johannes, J. (2010), 'Thresholding projection estimators in functional linear models', *Journal of Multivariate Analysis* **101**(2), 395 – 408. 87, 96

Carrasco, M. and Florens, J.-P. (2011), 'A Spectral Method for Deconvolving a Density', *Econometric Theory* **27**, 546–581. 12

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. 8, 14, 30, 56, 64, 70, 71, 88, 89

Carrasco, M., Florens, J.-P. and Renault, E. (2013), Asymptotic Normal Inference in Linear Inverse Problems, *in* J. S. Racine, A. Ullah and L. Su, eds, 'Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics'. 29, 30, 31, 113

Chen, X., Chernozhukov, V., Lee, S. and Newey, W. K. (2013), 'Local Identification of Nonparametric and Semiparametric Models', *Econometrica* (Forthcoming). 15

Chen, X. and Pouzo, D. (2012*a*), 'Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals', *Econometrica* **80**(1), 277–321. 8, 64, 89, 109

Chen, X. and Pouzo, D. (2012*b*), 'Sieve Quasi Likelihood Ratio Inference on Semi/nonparametric Conditional Moment Models', *Cowles Foundation Working Paper Series* . 88, 112

Chen, X. and Reiss, M. (2011), 'On Rate Optimality for Ill-Posed Iinverse Problems in Econometrics', *Econometric Theory* **27**(3), 497–521. 16, 66, 110

Conway, J. (2000), *A Course in Operator Theory*, Graduate Studies in Mathematics, American Mathematical Society. 14

Darolles, S., Fan, Y., Florens, J. P. and Renault, E. (2011*a*), 'Nonparametric Instrumental Regression', *Econometrica* **79**(5), 1541–1565. 2, 8, 10, 15, 16, 18, 19, 21, 27, 32, 34, 51, 52, 64, 65, 66, 70, 71, 72, 75, 84, 87, 88, 89, 90, 92, 95, 102, 113

Darolles, S., Fan, Y., Florens, J. P. and Renault, E. (2011*b*), 'Supplement to Nonparametric Instrumental Regression', *Econometrica Online Appendix* . 84

D'Haultfoeuille, X. (2011), 'On the Completeness Condition in Nonparametric Instrumental Problems', *Econometric Theory* **27**, 460–471. 15, 90

Dong, Y. (2010), 'Endogenous Regressor Binary Choice Models without Instruments, with an Application to Migration', *Economics Letters* **107**(1), 33 – 35. 77, 80

Engl, H. W., Hanke, M. and Neubauer, A. (2000), *Regularization of Inverse Problems*, Vol. 375 of *Mathematics and Its Applications*, Kluwer Academic Publishers, Dordrecht. 17, 23, 27, 28, 29, 32, 53, 57, 87, 89

Escanciano, J. C., Jacho-Chavez, D. and Lewbel, A. (2011), Identification and Estimation of Semiparametric Two Step Models, Technical report, Boston College. 77, 80

Ferraty, F., Van Keilegom, I. and Vieu, P. (2010), 'On the Validity of the Bootstrap in Non-Parametric Functional Regression', *Scandinavian Journal of Statistics* **37**(2), 286–306. 113, 115, 116

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). 12, 17, 18, 23, 26, 28, 32, 34, 37, 70, 74, 87, 91, 92

Fève, F. and Florens, J.-P. (2013), 'Non Parametric Analysis of Panel Data Models with Endogenous Variables', *Journal of Econometrics* **Forthcoming**. 37, 47, 100

Florens, J. P. and Heckman, J. J. (2003), 'Causality and Econometrics', *Mimeo* . 1

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. 28, 29, 30, 31

Florens, J.-P., Johannes, J. and Van Bellegem, S. (2012), 'Instrumental Regressions in Partially Linear Models', *The Econometrics Journal* **15**(2), 304–324. 80, 128, 133

Florens, J.-P. and Racine, J. (2012), 'Nonparametric Instrumental Derivatives', *Mimeo* . 29, 36, 45, 70, 87, 94, 95, 99, 100, 101, 102

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. 66, 75, 102, 110, 111

Gagliardini, P. and Scaillet, O. (2012), 'Tikhonov regularization for nonparametric instrumental variable estimators', *Journal of Econometrics* **167**(1), 61–75. 12

Golub, G. H., Heath, M. and Wahba, G. (1979), 'Generalized Cross-Validation as a Method for Choosing a good Ridge Parameter', *Technometrics* **21**(11), 215—223. 11, 33

Groetsch, C. (1993), *Inverse Problems in the Mathematical Sciences*, Theory and Practice of Applied Geophysics Series, Vieweg. 11

Hall, P. (1992), 'On Bootstrap Confidence Intervals in Nonparametric Regression', *The Annals of Statistics* **20**(2), pp. 695–711. 108

Hall, P. and Horowitz, J. L. (2005), 'Nonparametric Methods for Inference in the Presence of Instrumental Variables', *Annals of Statistics* **33**(6), 2904–2929. 8, 32, 34, 64, 88

Hall, P. and Racine, J. S. (2013), 'Infiinite Order Cross-Validated Local Polynomials Regressions', *WP - McMaster University, Department of Economics* **5**. 133

Hansen, B. E. (2008), 'Uniform Convergence Rates for Kernel Estimation with Dependent Data', *Econometric Theory* **24**(03), 726–748. 17, 70, 84

Hansen, P. C. (1992), 'Numerical Tools for Analysis and Solution of Fredholm Integral Equations of the First Kind', *Inverse Problems* **8**(6), 849–872. 11

Härdle, W. (1990), *Applied Nonparametric Regression*, Econometric Society Monographs, Cambridge University Press. 91

Härdle, W. and Bowman, A. W. (1988), 'Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands', *Journal of the American Statistical Association* **83**(401), pp. 102–110. 108

Härdle, W., Liang, H. and Gao, J. (2000), *Partially Linear Models*, Contributions to Statistics Series, Heidelberg: Physica-Verlag. 79

Härdle, W. and Mammen, E. (1993), 'Comparing Nonparametric Versus Parametric Regression Fits', *The Annals of Statistics* **21**(4), pp. 1926–1947. 108

Härdle, W. and Marron, J. S. (1985), 'Optimal Bandwidth Selection in Nonparametric Regression Function Estimation', *The Annals of Statistics* **13**(4), 1465–1481. 22

Härdle, W. and Marron, J. S. (1991), 'Bootstrap Simultaneous Error Bars for Nonparametric Regression', *The Annals of Statistics* **19**(2), pp. 778–796. 108, 112, 113

Hausman, J. A., Newey, W. K., Ichimura, H. and Powell, J. L. (1991), 'Identification and Estimation of Polynomial Errors-in-Variables Models ', *Journal of Econometrics* **50**(3), 273 – 295. 9

Hazelton, M. L. (2007), 'Bias Reduction in Kernel Binary Regression', *Computational Statistics and Data Analysis* **51**(9), 4393 – 4402. 73

Heckman, J. J. (1978), 'Dummy Endogenous Variables in a Simultaneous Equation System', *Econometrica* **46**(4), pp. 931–959. 69

Herrero, D. A. (1991), 'Diagonal Entries of a Hilbert Space Operator', *Rocky Mountain Journal of Mathematics* **21**(2), 857–865. 56

Hoderlein, S. and Holzmann, H. (2011), 'Demand Analysis as an Ill-posed Inverse Problem with Semiparametric Specification', *Econometric Theory* **27**, 609–638. 42, 44, 86, 87

Horowitz, J. L. (1992), 'A Smoothed Maximum Score Estimator for the Binary Response Model', *Econometrica* **60**(3), 505–31. 67

Horowitz, J. L. (2011), 'Applied Nonparametric Instrumental Variables Estimation', *Econometrica* **79**(2), 347–394. 2, 8, 41, 66, 70, 71, 87, 96

Horowitz, J. L. (2012), 'Adaptive Nonparametric Instrumental Variables Estimation: Empirical Choice of the Regularization Parameter', *Mimeo - NorthWestern University* . 12, 41, 87, 97, 103

Horowitz, J. L. and Lee, S. (2012), 'Uniform Confidence Bands for Functions Estimated Nonparametrically with Instrumental Variables', *Journal of Econometrics* **168**(2), 175 – 188. 88, 109, 115

Hsiao, C., Li, Q. and Racine, J. S. (2007), 'A Consistent Model Specification Test with Mixed Discrete and Continuous Data', *Journal of Econometrics* **140**(2), 802 – 826. 79

Huang, C., Hsing, T. and Cressie, N. (2011), 'Spectral Density Estimation Through a Regularized Inverse Problem', *Statistica Sinica* **21**(3), 1115–1144. 12

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. 22, 103

Ichimura, H. (1993), 'Semiparametric Least squares (SLS) and Weighted SLS Estimation of Single-Index Models', *Journal of Econometrics* **58**(1–2), 71 – 120. 67

Johannes, J., Bellegem, S. V. and Vanhems, A. (2011), 'Convergence Rates for Ill-posed Inverse Problems with an Unknown Operator', *Econometric Theory* **27**(3), 1–24. 31

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. 70, 87, 94

Kauermann, G. and Carroll, R. J. (2001), 'A Note on the Efficiency of Sandwich Covariance Matrix Estimation', *Journal of the American Statistical Association* **96**(456), pp. 1387–1396. 112

Kauermann, G., Claeskens, G. and Opsomer, J. D. (2009), 'Bootstrapping for Penalized Spline Regression', *Journal of Computational and Graphical Statistics* **18**(1), 126–146. 112

Kleibergen, F. and Paap, R. (2006), 'Generalized Reduced Rank Tests using the Singular Value Decomposition', *Journal of Econometrics* **133**(1), 97 – 126. 79

Klein, L. (1990), The Concept of Exogeneity in Econometrics, *in* R. Carter, J. Dutta and A. Ullah, eds, 'Contributions to Econometric Theory and Application', Springer New York, pp. 1–22. 1

Klein, R. W. and Spady, R. H. (1993), 'An Efficient Semiparametric Estimator for Binary Response Models', *Econometrica* **61**(2), 387–421. 65, 67

Krein, S. and Petunin, Y. (1966), 'Scales of Banach Spaces', *Russian Math. Survey* **21**(2), 89–168. 28

Kress, R. (1999), *Linear Integral Equations*, Applied mathematical sciences, Springer-Verlag. 9, 14, 90

Lewbel, A. (1991), 'The Rank of Demand Systems: Theory and Nonparametric Estimation', *Econometrica* **59**(3), pp. 711–730. 9

Li, Q. and Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press. 22, 27, 42, 91, 103

Liese, F. and Vajda, I. (2006), 'On Divergences and Informations in Statistics and Information Theory', *Information Theory, IEEE Transactions on* **52**(10), 4394 –4412. 116

Lukas, M. A. (1993), 'Asymptotic Optimality of Generalized Cross-Validation for Choosing the Regularization Parameter', *Numerische Mathematik* **66**(1), 41–66. 11, 33, 35

Lukas, M. A. (2006), 'Robust Generalized Cross-Validation for choosing the Regularization Parameter', *Inverse Problems* **22**(5), 1883–1902. 11, 35

Ma, S. and Racine, J. (2013), 'Additive Regression Splines With Irrelevant Categorical and Continuous Regressors', *Statistica Sinica* **23**, 515–541. 103

Manski, C. F. (1985), 'Semiparametric Analysis of Discrete Response : Asymptotic Properties of the Maximum Score Estimator', *Journal of Econometrics* **27**(3), 313–333. 67

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. 97

Marteau, C. and Loubes, J.-M. (2012), 'Adaptive Estimation for an Inverse Regression Model with Unknown Operator', *Statistics & Risk Modeling* **29**(3), 215–242. 11

Mathé, P. and Tautenhahn, U. (2011), 'Regularization under General Noise Assumptions', *Inverse Problem* **27**(3), 35–41. 27

Matzkin, R. L. (1991), 'Semiparametric Estimation of Monotone and Concave Utility Functions for Polychotomous Choice Models', *Econometrica* **59**(5), 1315–27. 67

Matzkin, R. L. (1992), 'Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models', *Econometrica* **60**(2), 239–70. 67

Morozov, V. (1967), 'Choice of a Parameter for the Solution of Functional Equations by the Regularization Method', *Sov. Math. Doklady* **8**, 1000–1003. 27

Neal, R. M. (2003), 'Slice Sampling', *Annals of Statistics* **31**(3), 705–767. 33

Newey, W. K. and Powell, J. L. (2003), 'Instrumental Variable Estimation of Nonparametric Models', *Econometrica* **71**(5), 1565–1578. 8, 64, 66, 88, 90

Panel Study of Income Dynamics (2003). 76

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge University Press. 1

Racine, J. S. and Nie, Z. (2012), *crs: Categorical Regression Splines*. R package version 0.15-18. **URL:** *http://CRAN.R-project.org/package=crs* 103

Rothe, C. (2009), 'Semiparametric estimation of binary response models with endogenous regressors', *Journal of Econometrics* **153**(1), 51 – 64. 64, 65, 75, 81

Santos, A. (2012), 'Inference in nonparametric instrumental variables with partial identification', *Econometrica* **80**(1), 213–275. 88, 109

Signorini, D. F. and Jones, M. C. (2004), 'Kernel Estimators for Univariate Binary Regression', *Journal of the American Statistical Association* **99**(465), 119–126. 73

Skoufranis, P. (2012), Numerical Ranges of Operators, Technical report. 50

Sokullu, S. (2010), 'Nonparametric Analysis of Two-Sided Markets'. 87, 109

Stone, C. J. (2005), 'Nonparametric M-regression with free knot Splines', *Journal of Statistical Planning and Inference* **130**(1–2), 183 – 206. 103

Vogel, C. (2002), *Computational Methods for Inverse Problems*, Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics. 11, 87

Wahba, G. (1977), 'Practical Approximate Solutions to Linear Operator Equations when the Data are Noisy', *SIAM Journal on Numerical Analysis* **14**(4), 651–667. 11