THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l' Université Toulouse 1 Capitole
Discipline : Sciences Economiques

Présentée et soutenue par

Samuele CENTORRINO
Le 5 juillet 2013

Titre :

Causality, Endogeneity and Nonparametric Estimation

JURY

Stephane BONHOMME, professeur, CEMFI
Jean-Pierre FLORENS, professeur, Université Toulouse I
Pascal LAVERGNE, professeur, Université Toulouse I
Jeffrey S. RACINE, professeur, Mc-Master University
Eric RENAULT, professeur, Brown University

Ecole doctorale : Toulouse School of Economics
Unité de recherche : GREMAQ - TSE
Directeur de Thèse : Jean-Pierre FLORENS

L'Université n'entend ni approuver, ni désapprouver les opinions particulières du candidat.

Suppose for example that I see one billiard ball moving in a straight line towards another: even if the contact between them should happen to suggest to me the idea of motion in the second ball, aren't there a hundred different events that I can conceive might follow from that cause? May not both balls remain still? May not the first bounce straight back the way it came, or bounce off in some other direction? All these suppositions are consistent and conceivable. Why then should we prefer just one, which is no more consistent or conceivable than the rest? Our a priori reasonings will never reveal any basis for this preference. In short, every effect is a distinct event from its cause. So it can't be discovered in the cause, and the first invention or conception of it a priori must be wholly arbitrary. Also, even after it has been suggested, the linking of it with the cause must still appear as arbitrary, because plenty of other possible effects must seem just as consistent and natural from reason's point of view. So there isn't the slightest hope of reaching any conclusions about causes and effects without the help of experience.

(David Hume, **Enquiry Concerning Human Understanding**)

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

To my parents, Angela e Nando

Acknowledgments

Writing the acknowledgements for this thesis is the most wonderful and difficult task at the same time. It is not only about the people that have helped me during these last 5 years and have made this intellectual journey much more exciting; but also about all those that have taken me hand in hand until this turning point of my life.

I am delighted to be finally able to thank my supervisor, Jean-Pierre Florens, for his patience, guidance and support. More than anybody else, he has transmitted to me the passion and the curiosity that are necessary to be a good researcher. All the hours spent in his office remain very precious to me and have allowed me to improve considerably my knowledge and understanding.

I am particularly grateful to Jeffrey S. Racine for the enormous support I received from him, all the interesting conversations about research and all the delicious lunches and dinners I enjoyed in his company. Not to forget his wife's delicious banana bread. Becoming a doctor would finally allow me to pay you a meal.

A special thank goes to Eric Renault, for being a great host during my visiting period at Brown. I appreciate the time he has devoted to be my mentor and my sponsor. I would also like to thank him for having accepted to referee my work and to be present as a member of my thesis committee.

I would equally like to thank Frank Kleibergen, Adam McCloskey, and Blaise Melly, for having made my stay at Brown very exciting and enjoyable. I hope I am going to deserve the trust they have given to me, and I wish them luck with all their future endeavours.

I would like to express my gratitude to all friends, coauthors and colleagues that have contributed during these five long years to my improvement as a researcher and as a man. In no particular order: Giuseppe Attanasi, Christophe Bontemps, Fortuna Casoria, Roberta Dessì, Elodie Djemai, Fréderique and Patrick Fève, Astrid Hopfensitz, Thibaut Laurent, Pascal Lavergne, Thierry Magnac, Maxime Marty, Nour Meddahi, Manfred Milinsky, Ivan Moscati, Nicolas Pistolesi, Paul Seabright, Guillaume Simon, Christine Thomas, and Giulia Urso.

I would finally like to thank Stephane Bonhomme, for having accepted to be part of my thesis committee.

This thesis is a personal achievement, but I would not have got here without the constant help of my family and my friends.

My first thank goes to Nicoletta, whose encouragement and enthusiasm have been essential for me to kick off this journey. She has seen something I could not see at the time, and I am very grateful she has taken the burden of guiding me towards the beginning of my PhD.

Inside and outside the courtyard of the Manufacture, I have shared my lunch breaks, my cigarettes, coffees and afternoons along the Garonne with my friends Kyriacos, Paulo, Antonio R., Anna and Racha.

I am grateful to Olivier Faugeras and Olivier Perrin (mieux connus comme *les deux Oliviers*), for all the very amusing and interesting conversations about research, politics and life.

A special thank goes to all my friends in Toulouse, who have shared with me many joyful meals, parties and nights out, and have always been beside me, even in the darkest moments: Antonio P., Beatrice, Flavia, Laura, Nico, Simone B. and Viviana.

I would also like to express my gratidute to Anaïs, Brigitte et Philippe, that have being great hosts when I first arrived here and helped me settle down in Toulouse; and to Gaël, Isa and Gigi, for cheering up my dinners with their herring, fajitas and various *delicatessen*.

Foremost, no words can express my immense gratitude to my everlasting friends that have remained loyal to me, despite all the time spent apart. Since the last years of high school, I have enjoyed their company and their affection. This thesis is an achievement I would like to share with Agata, Angelo, Ciccio, Filippo, Giovanni, Giuseppe, Sonia and Tiziana. A very particular thank goes to Marco, my friend, room-mate, wingman, guitar teacher and more.

This work is dedicated to my parents, Angela and Nando, and to my sisters, Micol and Clizia, whose unconditional love and support has been my main engine during all these years. I would also like to thank my brother-in-law, Antonio, for having so far patiently taken care of my sister. In a very Sicilian fashion, I am grateful to my godparents, Angelo and Angela, who have been a constant presence in my life and have followed closely my progresses and achievements.

Last but not least, I would like to thank you, Maria, for standing beside me everyday, beyond my moody and nervous temper, especially in these last months. I hope we will have many more years and precious moments to enjoy together.

Abstract

This thesis deals with the broad problem of causality and endogeneity in econometrics when the function of interest is estimated nonparametrically. It explores this problem in two separate frameworks.

In the cross sectional, iid setting, it considers the estimation of a nonlinear additively separable model, in which the regression function depends on an endogenous explanatory variable. 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 the time series context, it studies the implications of the assumption of exogeneity in a regression type model in continuous time. In this model, the state variable depends on its past values, but also on some external covariates and the researcher is interested in the nonparametric estimation of both the conditional mean and the conditional variance functions.

This first chapter deals with the latter topic. In particular, we give sufficient conditions under which the researcher can make meaningful inference in such a model. It shows that noncausality is a sufficient condition for exogeneity if the researcher is not willing to make any assumption on the dynamics of the covariate process. However, if the researcher is willing to assume that the covariate process follows a simple stochastic differential equation, then the assumption of noncausality becomes irrelevant.

Chapters two to four are instead completely devoted to the simple iid model. 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 second 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 regulariza-

tion). We derive the properties of a leave-one-out cross validation criterion in order to choose the regularization parameter.

In the third 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 four, 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.

Résumé

Cette thèse porte sur les problèmes de causalité et d'endogénéité avec estimation non-paramétrique de la fonction d'intérêt. On explore ces problèmes dans deux modèles différents.

Dans le cas de données en coupe transversale et iid, on considère l'estimation d'un modèle additif séparable, dans lequel la fonction de régression dépend d'une variable endogène. L'endogénéité est définie, dans ce cas, de manière très générale : elle peut être liée à une causalité inverse (la variable dépendante peut aussi intervenir dans la réalisation des régresseurs), ou à la simultanéité (les résidus contiennent de l'information qui peut influencer la variable indépendante). L'identification et l'estimation de la fonction de régression se font par variables instrumentales.

Dans le cas de séries temporelles, on étudie les effets de l'hypothèse d'exogénéité dans un modèle de régression en temps continu. Dans un tel modèle, la variable d'état est fonction de son passé, mais aussi du passé d'autres variables et on s'intéresse à l'estimation nonparamétrique de la moyenne et de la variance conditionnelle.

Le premier chapitre traite de ce dernier cas. En particulier, on donne des conditions suffisantes pour qu'on puisse faire de l'inférence statistique dans un tel modèle. On montre que la non-causalité est une condition suffisante pour l'exogénéité, quand on ne veut pas faire d'hypothèses sur les dynamiques du processus des covariables. Cependant, si on est prêt à supposer que le processus des covariables suit une simple équation différentielle stochastique, l'hypothèse de non-causalité devient immatérielle.

Les chapitres de deux à quatre se concentrent sur le modèle iid simple. Etant donné que la fonction de régression est solution d'un problème mal-posé, on s'intéresse aux méthodes d'estimation par régularisation.

Dans le deuxième chapitre, on considère ce modèle dans le cas d'un régularisation sur la norme $\mathbb{L}^2$ de la fonction (régularisation de type Tikhonov). On dérive les propriétés d'un critère de validation croisée pour définir le choix du paramètre de régularisation.

Dans le chapitre trois, coécrit avec Jean-Pierre Florens, on étend ce modèle au cas oú la variable dépendante n'est pas directement observée mais où on observe seulement une transformation binaire de cette dernière. On montre que le modèle peut être identifié en utilisant la décomposition de la variable dépendante dans l'espace des variables instrumentales et en supposant que les résidus de ce modèle réduit ont une distribution connue. On démontre alors, sous ces hypothèses, qu'on préserve les propriétés de convergence de l'estimateur non-paramétrique.

Enfin, le chapitre quatre, coécrit avec Frédérique Fève et Jean-Pierre Florens, décrit une étude numérique, qui compare les propriétés de diverses méthodes de régularisation. En particulier, on discute des critères pour le choix adaptatif des paramètres de lissage et de régularisation et on teste la validité du *bootstrap sauvage* dans le cas des modèles de régression non-paramétrique avec variables instrumentales.

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.

The time dimension offers often a solution to this problem. A cause-effect relation can unfold in time and give us further information about how to glue the puzzle. However, this brings in a new spectrum of problems related to the assumptions we can meaningfully make on the dynamics of the processes of interest.

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 and causality (or noncausality) 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 second part of this thesis to this literature (chapters two to four) 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 second contribution of this thesis is to link the concepts of causality and exogeneity in continuous time models. In this kind of models, which were borrowed from mathematics, the dependent state variable, that can be univariate or multivariate, follows a nonlinear stochastic differential equation, driven by a Brownian noise and it is only affected by its past values. However, in economics, we would like to be a little more general than that. We are interested in a state variable that can depend on its past and on other covariates. This modelling device has been used in several contributions to the theoretical and applied literature, but, to the best of our knowledge, no existing work has dug into the main assumption that underlies meaningful inference. Exogeneity is often assumed in a *naïve* way by considering mean independence conditions of the Brownian component with respect to the covariate process. The question we ask, in the first chapter of this thesis, is whether this assumption can stand still by itself or if it needs to be supported by further hypotheses. The answer we provide is double edged. This assumption is meaningful only if we are willing to completely specify the dynamics of the covariate process. In particular, if the covariate process follows a simple stochastic differential equation, then the exogeneity assumption is valid. However, if the researcher does not want to specify any particular dynamics for the covariate process, then the assumption of noncausality of the state variable onto the covariate process is needed to back exogeneity. Therefore, it is shown that noncausality is a sufficient condition for exogeneity in continuous time regression models and that, in some particular cases, it can allow to consider the covariate process to be long memory.

The exposition of the results of this research has privileged a temporal unfolding. Chapter 1 has been the first to be written by the author and presents the latter contribution of this thesis.

Chapters 2, 3 and 4 present instead the contribution to the nonparametric instrumental variable literature in the iid setting. Chapter 2 discusses the results about the data-driven selection of the regularization parameter in nonpametric instrumental regressions and it proves its optimality. Chapter 3, coauthored with Jean-Pierre Florens, extends the nonparametric instrumental variable framework to binary response models. Finally, Chapter 4, 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 2 has been the last one to be started, as it was inspired by some empirical observation when working on chapters 3 and 4, its results are used in the latter part of this work and are therefore presented first.

Chapter 1

Nonparametric Nonstationary Regressions in Continuous Time

Abstract

This paper extends nonparametric estimation to time homogeneous nonstationary diffusion processes where the drift and the diffusion coefficients are function of a multivariate exogenous time dependent variable Z. We base our estimation framework on a discrete sampling of data, following a recent stream of literature. We prove almost sure convergence and normal asymptotic distribution using the concept of multivariate occupation densities, in order to make the multivariate kernel estimation meaningful in the context of nonstationary time processes. We widely discuss the noncausality assumption in such a context and provide an extension in which Z is a long memory process of dimension 1.

1.1 Introduction

In economics, a time homogeneous diffusion process in dimension one is often used to characterize the behaviour of a given variable Y_t, called the state variable (e.g., a stock price, the interest or the exchange rate). The structural model is written under the form:

$$dY_t = \mu(Y_t)dt + \sigma(Y_t)dB_t^* \tag{1.1.1}$$

where dB_t^* is the time increment of a standard Brownian motion, that is normally distributed with zero mean and variance equal to the time increment dt[1]. The two functions $\mu(Y_t)$ and $\sigma(Y_t)$ are called the drift and the diffusion coefficient, respectively.

This paper copes with a more general structural form of the model, where the drift and the diffusion coefficients can possibly be function of a time dependent variable Z_t. Our data generating process (DGP) can therefore be written in the following way:

$$dY_t = \mu(Y_t, Z_t)dt + \sigma(Y_t, Z_t)dB_t \tag{1.1.2}$$

where dB_t is the Brownian motion associated to the covariate depending process. This model can

[1] For a review of the properties of a standard Brownian motion, see Karatzas and Shreve (1991) and Øksendal (2003)

be interpreted as a general location scale model in continuous time. In particular, in this regression model, the objects of interest are both the location and the scale function.

This structural model is interesting in different respects. First of all, it generalizes to continuous Markov processes the economic idea that a given phenomenon may not be *self-explanatory*. Other factors may intervene in determining the outcome of the state today. This may be summarized in the concept of causality, which is central in econometrics but which has not yet been extensively studied, to the best of our knowledge, in the case of continuous time diffusions. Furthermore, Z_t may be thought as a set of parameters which varies over time. The latent stochastic volatility model can be therefore encompassed in this more general framework (e.g. see Bandi and Reno, 2009). Finally, this model uses higher level assumption than a simple univariate diffusion, as the state variable needs to be only conditionally Markov; and it is not reducible to a multivariate diffusion, as we are not making any assumption about the structure of the covariate process Z_t, which is allowed to be any continuous Feller process. In that sense, we also allow for greater flexibility and we discuss a particular case in which Z_t exhibits long memory.

The approach of this paper is not completely new either to theoretical or to applied literature. It belongs, in fact, to the more general class of semimartingale regression models, as defined in Aalen (1980). Some authors have considered the estimation of the drift term in (1.1.2). A recent nonparametric approach is presented in Stone and Huang (2003), who use a free knots regression splines estimator, when continuous realizations of the process are observed and the diffusion term is assumed to be known. Park (2008) proposes a parametric minimum distance estimator of the drift, under a time change approach.

Applications of this model also counts several contributions, both in macroeconomics and in finance. Creedy and Martin (1994) and Creedy et al. (1996) develop a framework in which the variable Z represents market fundamentals that influence the behaviour of prices and US/UK exchange rate respectively[2]. These papers however use parametric methods (i.e. maximum likelihood) and maintain the assumption of stationarity. In a more recent paper, Fernandes (2006) generalizes the same framework in order to supply a model for forecasting financial crashes, under the assumption of a constant diffusion term, where Y denotes market indexes or long-term interest rate and Z

[2]For a more recent application see also Jäger and Kostina (2005)

represents market fundamentals (i.e. dividends, short-term interest rate).

Beside ergodic stationarity, the existing theoretical and applied literature overlooks the assumption of strict exogeneity in such models. While it is easy to interpret exogeneity in discrete time, when it comes to continuous time models, exogeneity strictly relates to the causality of the state variable Y onto the covariate process Z. In this paper, we show that noncausality is a sufficient but not necessary condition for correct statistical inference in model (1.1.2). We give explicit examples in which the failure of noncausality does not harm our nonparametric estimators and other examples in which it does.

For instance, in a monetarist model, one may reasonably expect exchange rate dynamics to affect money demand and supply if the country under study is big enough. This would lead to a two-way causality between exchange rate and its covariates. Therefore, the underlying assumptions about the dynamics of money demand and supply and the type of causality arising between these covariates and the exchange rate become essential to prove the goodness of our inference.

The novelty of this work is thus twofold. On the one hand, it clearly defines the assumption of *strict exogeneity* in such a continuous time context. On the other hand, it focuses on nonparametric estimation of both the location and the scale parameter while relaxing the assumption of stationarity, following a recent stream of literature (Bandi and Phillips, 2003; Bandi and Nguyen, 2003, among others)[3]. Finally, it presents and discusses a very simple approach to the uncovered interest parity of such a nonparametric approach.

Nonparametric estimation of stochastic diffusion processes hinges on a considerably rich literature. The main objects of interest being the drift and the diffusion coefficients, it may be difficult to identify them without further assumptions when the data are discretely sampled, because of the so-called *aliasing problem* (Phillips, 1973; Hansen and Sargent, 1983). Furthermore, while the drift term is of order dt, the diffusion term is of order $\sqrt{dt}$, which means that much of the infinitesimal variation in the process reflects the latter more than the former. This entails the impossibility to show consistency of the drift estimator as the sample frequency increases, i.e. $dt \to 0$ (so-called *infill* asymptotics).

[3]Interested readers are referred to Bandi and Phillips (2010), for a complete review of the existing econometric literature on Nonparametric Estimation for Nonstationary Processes in Continuous Time.

A possible way to correctly identify both the diffusion and the drift coefficient is to assume that the process is time stationary, so that a time invariant density $\pi(y)$ exists. The *backward* and the *forward* Kolmogorov equations allow then to specify a relation between this density, the drift and the diffusion coefficients.

Nevertheless, the assumption of stationarity seems somehow too restrictive and it does not take into account many interesting phenomena in economics. Relaxing the assumption of stationarity requires careful handling of kernel estimators, which is not meaningful any more as an estimator of the invariant density. An interpretation of the kernel estimator in time series, both in the univariate and multivariate case, may be given in terms of occupation densities (Geman and Horowitz, 1980). Namely, in the univariate case, Phillips and Park (1998) show the convergence of the nonparametric kernel estimator to the *chronological* local time of the stochastic process (see, e.g. Revuz and Yor, 1999, Ch. VI, for a review of the properties of local time).

Bandi and Phillips (2003) are then able to overcome the identification issues without assuming stationarity. *Harris recurrence*, which is a substantially milder assumption, is required instead. To ensure consistency of the drift term, they couple *infill* asymptotics with lengthening time span of observations, i.e. $T \to \infty$ (so-called *long span* asymptotics).

In related papers, Locherbach and Loukianova (2008) and Bandi and Moloche (2008) use the same framework under the assumption of Harris recurrence for the joint process to prove convergence of such an estimator in the multivariate case.

In this paper, we show that their convergence results can be extended to the nonparametric estimator of the drift and the diffusion in model (1.1.2).

However, while we show the properties of our estimation for any dimension d of the covariate process, we run simulations for the case in which $d = 1$. As pointed out by Schienle (2011), Harris recurrence is a property which is rarely satisfied when the dimension of the process increases. We do not tackle this question here, as it goes beyond the scope of the present paper. We therefore acknowledge the limited applicability of this framework that may be a topic for further research.

The paper is structured as follows. Section 1.2 set up the general framework. Section 1.3 overviews the theoretical foundations on which this work is based upon. Section 1.4 provides the main

estimation framework and the asymptotic properties. Section 1.5 discusses an extension to long memory processes. Section 1.6 includes a simulation study which draws the finite sample properties of the estimator. Finally, section 1.6 outlines the practical relevance of our approach by discussing an application to the Uncovered Interest Parity.

1.2 Motivations and theoretical foundations

The possibility to meaningfully define conditional moments for continuous time processes is a necessary condition to perform statistical inference based on sample analogues. Diffusion type processes are extremely convenient in this respect, as the definition of conditional moments is straightforward under the Markov property. The goal of this section is therefore to show that, under suitable assumptions on the conditional and the marginal process, we can make our data generating process being a diffusion process.

We suppose here to observe a multivariate Markov continuous time process $\{Z_t : t \geq 0\}$ of given dimension d; and a scalar process $\{Y_t : t \geq 0\}$ which is Markov conditionally on Z_t. We denote by X_t the joint process $\{Y_t, Z_t\}$ which takes value in a Polish space $(E, \mathcal{E})$.

Define $(\Omega_z, \mathcal{Z}, \mathcal{P}_z)$ and $\{\mathcal{Z}_t\}_{t\geq 0}$ the probability space and the natural filtration associated to the process Z_t, respectively.

We further consider a univariate Brownian motion $\{B_t : t \geq 0\}$ defined on the probability space $(\Omega_B, \mathcal{F}^B, \mathcal{P}_B)$ and adapted to a filtration $\{\mathcal{F}_t^B\}_{t\geq 0}$. We assume B_t to be a $\mathcal{Z}_t-$adapted martingale, so that $\mathbb{E}[dB_t|\mathcal{Z}_t] = 0$.

The joint filtration, generated by the process $\{X_t : t \geq 0\}$ is set as follows:

$$\mathcal{X}_t := \mathcal{Y}_t \vee \mathcal{Z}_t = \sigma(y_0) \vee \mathcal{Z}_t \vee \mathcal{F}_t^B = \sigma\left(y, Z_s, B_s; 0 \leq s \leq t\right) \tag{1.2.1}$$

where y_0 is the starting value of the process Y, which is assumed to be independent of Z. We assume all filtrations satisfy the *usual conditions* (or hypotheses), i.e. they contain all the sets of zero measure for $t = 0$ and they are *right-continuous*.

In our framework, the filtration generated by the process Z_t enters the construction of the filtration

under which the process Y_t is defined. To ensure exogeneity of the joint process, we need to impose some conditions on the marginal process Z.

Definition 1.2.1 (Strong global noncausality, Florens and Fougere, 1996)**.** $\mathcal{X}_t$ does not strongly cause Z_t given $\mathcal{Z}_s$ if:

$$\mathcal{Z}_t \perp\!\!\!\perp \mathcal{X}_s | \mathcal{Z}_s \qquad \forall s, t \in [0, T] \qquad \blacksquare$$

This properties is trivially satisfied if $t \leq s$. Nevertheless, if $\mathcal{X}_t$ does not strongly cause Z_t, every $\mathcal{Z}_t$-adapted martingale is also a $\{\mathcal{X}_t\}$-martingale (Florens and Fougere, 1996, Theorem 2.2).

The assumption of strong global noncausality is simply stating that, conditionally on the observation of the process Z at time s, the joint process is not delivering any additional information about the marginal process Z_t, $\forall t$. However, the most important implication of this hypothesis is that it immediately entails the preservation of the martingale property of B_t under the joint filtration.

It is also important to notice that, in this context, the assumption of global noncausality is equivalent to the assumption of instantaneous noncausality (in a Granger sense) and to any other noncausality assumption, as $\mathcal{Z}$ is also the conditioning filtration (see, Comte and Renault, 1996; Florens and Fougere, 1996). Therefore, using the most restrictive assumption of noncausality only serves maintaining the martingale property.

Under the conditional markovianity of Y_t and noncausality, we can give to our regression model the attribute of a stochastic differential equation (Karatzas and Shreve, 1991). The *conditional* diffusion process is thus defined as::

$$dY_t = \mu(Y_t, Z_t)dt + \sigma(Y_t, Z_t)dB_t \tag{1.2.2}$$

where $\mu(\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ and $\sigma(\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$, which are our objects of interest in what follows. This model can be considered as an extension of the conditional mean model studied in Park (2008). We extend his model in two respects. First of all, we allow the volatility term to also depend on Z_t. Second, we allow for any (possibly nonlinear) specification of the drift and the diffusion term[4].

[4] Park (2008) considers as an error term in his model any continuous martingale with bounded variations. How-

Remark 1. Noncausality is a sufficient but *not* necessary condition for correct inference in a conditional continuous time model. For instance, consider Z_t, a Brownian motion; $\mathcal{Z}_t$ its natural filtration; and W_t another Brownian motion, adapted to $\mathcal{Z}_t$. Then, we can generate B_t so that:

$$B_t = \rho Z_t + \sqrt{1-\rho^2} W_t$$

where $\rho \in (0,1)$. In this case, we generate instantaneous causality in the sense of Comte and Renault (1996). However, since dB_t are iid increments independent of any filtration, instantaneous causality does not harm inference in model (1.2.3). ■

Remark 2 (Simultaneous equations in continuous time). Consider the previous example and our conditional diffusion process. Suppose, Z_t and B_t are correlated Brownian motions and we are interested in the estimation of the drift term. However the true DGP writes as:

$$dY_t = \mu(Y_t, Z_t)dt + \sigma(Y_t, dZ_t)dB_t$$

where dZ_t is the infinitesimal increment of Z_t. In this particular example, instantaneous causality is coupled with predictability of Z_t with respect to the joint filtration $\mathcal{X}_t$, i.e. dB_t is not a martingale on $\mathcal{X}_t$, as a part of it can be predicted through dZ_t. In this case, our approach cannot deliver a consistent estimation of the drift. We can consider this model as an extention of simultaneous equations in continuous time.

For ease of notations, we write our DGP as follows:

$$dY_t = \mu(X_t)dt + \sigma(X_t)dB_t \tag{1.2.3}$$

where X_t denotes the joint process.

We assume the following conditions to hold in studying (1.2.3).

Assumption 1. *The functions $\mu(\cdot)$ and $\sigma(\cdot)$ satisfy the following assumptions:*

(i) They are measurable on the σ-field generated by all the Borel sets on $\mathcal{E}$ and they are at least

ever, up to a time change, any continuous martingale can be rewritten as a *Dambis-Dubins-Schwarz* Brownian motion.

twice continuously differentiable with respect to both their arguments;

(ii) They satisfy local Lipschitz and growth conditions in X_t, i.e. for every compact set $B \in \mathcal{E}$, there exists a constant, C, such that, for any realization x_1 and x_2 in B,

$$\|\mu(x_1) - \mu(x_2)\| + \|\sigma(x_1) - \sigma(x_2)\| \leq C\|x_1 - x_2\| \tag{1.2.4}$$

and

$$\|\mu(x_1)\|^2 + \|\sigma(x_1)\|^2 \leq C^2\left(1 + \|x_1\|^2\right) \tag{1.2.5}$$

(iii) Nondegeneracy (ND) - $\sigma^2(\cdot) > 0$ on $\mathcal{E}$

(iv) Local Integrability (LI) with respect to Y_t, for any realization of the process $Z_t = z$:

$$\forall (y_1, z) \in E, \exists \delta > 0 \quad \textit{such that} \quad \int_{y_1-\delta}^{y_1+\delta} \frac{|\mu(\zeta, z)|d\zeta}{\sigma^2(\zeta, z)} < \infty \qquad \blacksquare \tag{1.2.6}$$

Conditions (ii) and (iii) (Karatzas and Shreve, 1991, Theorem 2.2, p. 289) ensure the existence of a strong solution to equation (1.2.3). We can therefore write the usual Itô's stochastic differential equation, which is the solution of our *DGP* in the following form:

$$Y_t = y + \int_0^t \mu(X_s)ds + \int_0^t \sigma(X_s)dB_s \tag{1.2.7}$$

where y is an initial condition independent of the Brownian motion B_t and Y_t is adapted to the filtration $\mathcal{Y}_t \vee \mathcal{Z}_t$.

The drift and the diffusion coefficients can be thus defined as in the standard framework. Take any function $f \in \mathbb{C}^2$ of Y_t, so to preserve the semimartingale properties of our solution (see Protter, 2003, Theorem 32, p. 174). Using Itô's lemma and taking expectation over any couple of realizations (y, z), the infinitesimal generator $\mathcal{L}$ of equation (1.1.2) can be defined as:

$$\lim_{t \to 0} \frac{1}{t}\mathbb{E}^x \left[f(Y_t) - f(y)\right] = (\mathcal{L}f)(y) \tag{1.2.8}$$

Taking $f(Y_t) = Y_t$, we obtain the drift coefficient as the conditional instantaneous change in the

process:

$$\mathbb{E}^x [Y_t - y] = t\mu(x) + o(t) \tag{1.2.9}$$

while, taking $f(Y_t) = (Y_t - y)^2$, we obtain the diffusion coefficient as the conditional instantaneous change in the volatility of the process,

$$\mathbb{E}^x \left[(Y_t - y)^2\right] = t\sigma^2(x) + o(t) \tag{1.2.10}$$

We can then proceed as in any standard nonparametric inference problem for conditional moments, using sample analogues to identify conditional expectations over infinitesimal time distances. In practise, the exogenous case is encompassed in the existing literature for stochastic processes. In the next sessions, we show that the asymptotic properties of the drift and the diffusion term are equivalent to those of a multivariate diffusion when the dimension is equal to $d + 1$.

1.3 Additive Functionals and Occupation Density

Before to explicitly derive the nonparametric estimators of the drift and the diffusion coefficient, we need to set up the main definitions and theorems which allow us to meaningfully define a standard kernel estimator in such a nonstationary context.

We assume the following conditions about the joint process to hold.

Assumption 2. *(i) X_t is Harris recurrent;*

(ii) Under $\mathcal{X}_t$, X_t is a special semi-martingale and it admits a Doob-Meyer decomposition of the type:

$$X_t = H_t + M_t \qquad \forall t \in (0, T]$$

where H_t is a $\mathcal{X}_t$-predictable process and M_t is a $\mathcal{X}_t$-local martingale such that $\mathbb{E}(M_t|\mathcal{X}_s) = 0, \forall s < t$. ■

In particular, since every $\mathcal{X}_t$-martingale can be written as a time changed *Dambis-Dubins-Schwarz* Brownian motion (Revuz and Yor, 1999, Ch. V, Theorem 1.6), X_t is a Brownian semimartingale.

Condition (i) is the minimal requirement to perform nonparametric inference on the joint process. It is possible to show that conditional stationarity of Y given Z and Harris recurrence of Z are sufficient conditions to obtain Harris recurrence of the joint process (see the Appendix). However, it is not possible to assume a more general structure on the conditional process and still to obtain condition (i).

Example 1. Consider a conditional Ornstein-Uhlembeck process, where the drift function μ is linear and the diffusion function is a constant:

$$dY_t = (\theta_1(Z_t) - \theta_2 Y_t)\, dt + \theta_3 dB_t$$

where $\theta_2 > 0$ (so that the process is mean reverting) and any function θ_1 of Z_t. This process has a stationary distribution given $Z = z$. Therefore, for Z being Harris recurrent, the joint process (Y, Z) will also be Harris recurrent. ■

For any measurable Borel set $B \subset \mathcal{E}$, we choose a measure m. This measure is invariant if and only if,

$$m(B) = \int_E \mathbb{P}\left(X_t^{(x)} \in B\right) m(dx) \tag{1.3.1}$$

where $X_t^{(x)}$ denotes the realization of the joint process at time t for a given initial condition x. In particular, Harris recurrence is a sufficient condition for the existence of an invariant measure, unique up to multiplication by a constant and absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^{d+1}$ (i.e. $m \ll \lambda$). The absolute continuity of m further implies that m admits a density with respect to the Lebesgue measure, i.e. a random function $p_t(\cdot)$ such that $m(dx) = p_t(x)\lambda(dx)$.

Definition 1.3.1 (Höpfner and Löcherbach, 2003)**.** An additive functional of X is a process $A = (A_t)_{t\geq 0}$, such that:

(i) A is $\mathcal{X}$−adapted, $A_0 = 0$;

(ii) All paths of A are nondecreasing and right-continuous;

(iii) For all $s, t \geq 0$, we have $A_{t+s} = A_t + A_s * \theta_t$, where θ_t is a family of shift operators for X. ■

We focus our attention here to integrable additive functionals. For every Borel set B, the measure ν defined by the functional A for each t is equal to:

$$\nu_A(B) = \mathbb{E}_m\left(\int_0^1 \mathbb{1}_B(X_s)dA_s\right) = \frac{1}{t}\mathbb{E}_m\left(\int_0^t \mathbb{1}_B(X_s)dA_s\right)$$

A functional is termed *integrable* when:

$$\| \nu_A \| = \nu_A(E) = \mathbb{E}_m(A_1) < \infty$$

In particular, when the functional $A_t = t$, for each Borel set B, we can define:

$$\eta_t^B = \int_0^t \mathbb{1}_B(X_s)ds \quad , \quad t \geq 0$$

which heuristically counts the amount of times for which X_s belong to B, for $T \to \infty$. In this particular case, we obtain that:

$$\mathbb{E}_m\left(\int_0^1 \mathbb{1}_B(X_s)ds\right) = m(B)$$

which defines the *occupation measure* for the set B (Geman and Horowitz, 1980), i.e. the time spent by the process in the set B up to time t. Therefore, the measure defined by the constant functional on each subset of $\mathcal{E}$ is equivalent to the invariant measure of X_t. Since the invariant measure admits a density with respect to the Lebesgue measure, there exists a random function $p_t(\cdot)$, such that:

$$m(B) = \int_B p_t(x)\lambda(dx)$$

We define , following this terminology, $p_t(\cdot)$ to be the *occupation density* of X. In dimension 1, the invariant measure is defined to be the *sojourn* time of a given process X (Park, 2005), while the random function $p_t(x)$ corresponds to the local time of the process (Borodin, 1989). This is formally defined as the Radon-Nykodim derivative of the sojourn time with respect to the Lebesgue measure. Our approach can be thus considered a generalization of the univariate case.

Remark 3. For the stationary case we have that:

$$\int p_t(x)\lambda(dx) = 1$$

so that $p_t(x) = \pi(x)$ is the invariant stationary density of X_t. ■

The following theorem gives the condition for weak convergence of additive functionals of a Harris recurrent process X:

Theorem 1.3.2 (Höpfner and Löcherbach, 2003)**.** *For a given constant $\alpha \in (0,1]$ and a function $l(\cdot)$ slowly varying at infinity*[5]*, the following are equivalent:*

(i) For every nonnegative measurable function $g(\cdot)$ with $0 < m(g) < \infty$, one has regular variation at 0 of resolvants[6] *in X if*

$$(R_{1/t}g)(x) = \mathbb{E}_x\left(\int_0^\infty e^{-\frac{1}{t}s} g(X_s)ds\right) \quad \sim \quad \frac{t^\alpha}{l(t)} m(g) \quad , \quad t \to \infty \tag{1.3.2}$$

(ii) every additive functional A of X with $0 < \mathbb{E}_m(A_1) < \infty$, one has:

$$\frac{(A_t)_{t\geq 0}}{t^\alpha / l(t)} \to \mathbb{E}_m(A_1) W^\alpha \qquad as \quad t \to \infty \tag{1.3.3}$$

under the Skorohod topology, where W^α is the Mittag-Leffler process of index α[7]*.* ■

Remark 4. Equation 1.3.2 simply states that we are restricting our attention to null recurrent diffusions with regular variation of the resolvent at 0. In the more general case, one should define the kernel estimator for any function $v_t = \mathbb{E}_m\left[\int_0^t g(X_s)ds\right]$ (Locherbach and Loukianova, 2008). In our case we take $v_t = t^\alpha / l(t)$. Moreover, equation 1.3.2 is equivalent to the condition given by Bandi and Moloche (2008, Theorem 2), where $C_X = m(g) < \infty$. ■

Remark 5. For stationary processes, we simply set $\alpha = 1$, $l(t) = 1$ and the Mittag-Leffler process $W^1 = Id$ (the deterministic process) by definition. Thus, for any measurable bounded function

[5] A function $f : [a,\infty) \to (0,\infty)$, $a > 0$ is said to be slowly-varying at infinity in the sense of Karamata if $\lim_{x\to\infty} f(\lambda x)/f(x) \to 1$, for $\lambda > 0$.

[6] For $\alpha > 0$ and a continuously differentiable function $g(\cdot)$, we define the resolvent operator R_α, by $(R_\alpha g)(x) = \mathbb{E}_x\left(\int_0^\infty e^{-\alpha s} g(X_s)ds\right)$. $R_\alpha g$ is a bounded continuous function (Øksendal, 2003, Definition 8.1.2 and Lemma 8.1.3, pg. 135).

[7] Interested readers are referred to Höpfner and Löcherbach (2003), for general definition and properties of Mittag-Leffler processes.

$f(\cdot)$, we obtain convergence by equation 1.3.3, i.e.:

$$\frac{1}{T}\int_0^T f(X_s)ds \xrightarrow{p} \int f(x)\pi(x)dx = \mathbb{E}(f(x))$$

where $\pi(\cdot)$ is the invariant stationary probability density. ■

1.4 Estimation and Asymptotic Properties

For simplicity, we suppose that the process $\{X_t, t \geq 0\}$ is sampled at equispaced times in the interval $[0,T]$, where T is a strictly positive number. If n is the sample size in $[0,T]$, we obtain that the time lag between two observations is equal to $\Delta_{n,T} = \frac{T}{n}$. The observed sample is therefore denoted as $X_{i\Delta_{n,T}}$ for all $i = 1, \cdots, n$.

Under these hypotheses and following the definitions given in equations (1.2.9) and (1.2.10), we can estimate the drift and the diffusion coefficients as follows:

$$\hat{\mu}_{n,T}(x) = \frac{1}{\Delta_{n,T}} \frac{\frac{1}{h_{n,T}^{d+1}} \sum_{i=1}^{n-1} \mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)\left(Y_{(i+1)\Delta_{n,T}} - Y_{i\Delta_{n,T}}\right)}{\frac{1}{h_{n,T}^{d+1}} \sum_{i=1}^{n} \mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)} \tag{1.4.1}$$

$$\hat{\sigma}^2_{n,T}(x) = \frac{1}{\Delta_{n,T}} \frac{\frac{1}{h_{n,T}^{d+1}} \sum_{i=1}^{n-1} \mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)\left(Y_{(i+1)\Delta_{n,T}} - Y_{i\Delta_{n,T}}\right)^2}{\frac{1}{h_{n,T}^{d+1}} \sum_{i=1}^{n} \mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)} \tag{1.4.2}$$

where,

$$\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right) = K\left(\frac{Y_{i\Delta_{n,T}} - y}{h_{n,T}^{(y)}}\right) \prod_{j=1}^{d} K\left(\frac{Z_{j,i\Delta_{n,T}} - z_j}{h_{n,T}^{(z)}}\right)$$

where $h_{n,T}^{(y)}$ and $h_{n,T}^{(z)}$ are two bandwidths parameters for the process Y_t and Z_t respectively. For notational brevity, we also suppose that $h_{n,T}^{(y)} = h_{n,T}^{(z)}$. For further ease of notations, we denote $x = (y, z)$.

The kernel functions $\mathbf{K}(\cdot)$ and $K(\cdot)$ satisfy the following conditions.

Assumption 3. - *(Pagan and Ullah, 1999; Bandi and Moloche, 2008; Ruppert and Wand, 1994)*

(i) The function $K(\cdot)$ is a non negative, bounded, continuous, and symmetric function such that:

$$\int_{-\infty}^{\infty} K(u)du = 1 \qquad \int_{-\infty}^{\infty} K^2(u)du < \infty \quad \text{and} \quad \int_{-\infty}^{\infty} u^2 K(u)du < \infty$$

(ii) The function $\mathbf{K}(\cdot)$ is a bounded kernel, such that $\int uu'\mathbf{K}(u)du = \rho_2(\mathbf{K})\mathbb{I}$, where $\rho_2(\mathbf{K}) \neq 0$ is a scalar and $\mathbb{I}$ is the identity matrix of dimension $d+1$.

(iii) The function $\mathbf{K}(\cdot)$ is locally Lipschitz, i.e.

$$|\mathbf{K}(x) - \mathbf{K}(v)| \leq D(v,\varepsilon)\|x - v\| \tag{1.4.3}$$

where:

$$D(v,\varepsilon) := \sup\left\{\frac{|\mathbf{K}(x) - \mathbf{K}(v)|}{\|x - v\|}, \quad s.t. \quad \|x - v\| \leq \varepsilon\right\} \tag{1.4.4}$$

is the non negative local-Lipschitz constant function, such that:

$$\lim_{\varepsilon \to 0} \mathbb{E}_m\left(D(v,\varepsilon)\right) < \infty \qquad \blacksquare \tag{1.4.5}$$

While many of these assumptions are standard in the nonparametric literature, assumption *(iii)* deserves some additional discussion. The multivariate kernel function is often supposed to satisfy some global regularity condition, e.g. some Hölder type of continuity. However, in the nonstationary case, any function which satisfies such a kind of global uniform continuity will explode as $T \to \infty$, when it is integrated with respect to time. Therefore, we require the kernel function to satisfy this uniform condition only locally in an open ball of radius ε. In particular,we suppose that *local-Lipschitz constant function* (as defined e.g. in Borwein et al., 2003) is itself a random variable and that it is integrable with respect to the invariant measure [8].

Under assumption (3), we can thus define the kernel estimator of the occupation density of X:

$$\hat{L}^X(T,x) = \frac{\Delta_{n,T}}{h_{n,T}^{d+1}} \sum_{i=1}^{n} \mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right) \tag{1.4.6}$$

[8] This assumption can be considered a stronger version of the joint Hölder continuity of the occupation density for Gaussian field. For a review on this topic see, e.g., Dozzi (2003, p. 146).

Using theorem 1.3.2, it is possible to show the weak convergence of this estimator towards the Radon-Nikodym derivative of m with respect to the Lebesgue measure on R^{d+1}.

Corollary 1.4.1. *Consider the following additive functional of X_s:*

$$\Phi_t = \int_0^t \frac{1}{h_{n,T}^{d+1}} \mathbf{K}_{h_{n,T}}(X_s - x)\, ds$$

which is strictly positive and integrable $\forall t \geq 0$. The kernel estimator (1.4.6) converges almost surely to Φ_t for $n, T \to \infty$, provided that:

$$\frac{\hat{L}^X(T,x)}{h_{n,T}^{d+1}} \left(\Delta_{n,T} \log(1/\Delta_{n,T})\right)^{1/2} \xrightarrow{a.s.} 0$$

Moreover, when $h_{n,T} \to 0$, we obtain:

$$\frac{\Phi_t}{t^\alpha / l(t)} \to C p_\infty(x) W^\alpha \qquad as \quad t \to \infty$$

by theorem 1.3.2, where C is a process specific constant.

Proof. See the Appendix. ■

Remark 6. Under stationarity, (1.4.6) is a well defined estimator of the stationary density, as $\frac{\hat{L}^X(T,x)}{T} \xrightarrow{p} \pi(x)$. ■

Remark 7. The estimator presented here has been firstly proposed by Bandi and Moloche (2008) and it is a generalization to multivariate processes of the local time estimator for scalar diffusion process presented in Florens-Zmirou (1993).

1.4.1 Estimation and asymptotic distribution of the drift coefficient

In this section we report the convergence properties of the drift estimator.

Theorem 1.4.2. Almost sure convergence of the drift estimator.

Suppose that:

$$\frac{\hat{L}^X(T,x)}{h_{n,T}^{d+1}} \left(\Delta_{n,T} \log(1/\Delta_{n,T})\right)^{1/2} \xrightarrow{a.s.} 0$$

with $\hat{L}_X(T,x)h_{n,T}^{d+1} \to \infty$ *with* $\Delta_{n,T} \to 0$, $h_{n,T} \to 0$ *and* $n, T \to \infty$, *then the estimator of equation (1.4.1) converges almost surely to the drift coefficient. I.e.:*

$$\hat{\mu}_{n,T}(x) \xrightarrow{a.s.} \mu(x) \tag{1.4.7}$$

Proof. See the Appendix. ∎

Theorem 1.4.3. Asymptotic distribution of the drift estimator.

Suppose that:

$$\frac{\hat{L}^X(T,x)}{h_{n,T}^{d+1}} \left(\Delta_{n,T} \log(1/\Delta_{n,T})\right)^{1/2} \xrightarrow{a.s.} 0$$

$$\hat{L}^X(T,x)h_{n,T}^{d+1} \xrightarrow{a.s.} \infty$$

with $h_{n,T} = O_{a.s}\left(\hat{L}^X(T,x)^{-\frac{1}{d+1}}\right)$, $\Delta_{n,T} \to 0$, $h_{n,T} \to 0$ *and* $n, T \to \infty$, *then the estimator described in equation (1.4.1) converges in distribution to a Gaussian random variable.*

$$\begin{aligned}\sqrt{\hat{L}^X(T,x)h_{n,T}^{d+1}}\left(\hat{\mu}_{n,T}(x) - \mu(x) - \Gamma^{\mu}(x)\right) \\ \xrightarrow{d} \sigma(x)\mathcal{N}\left(0, \left(\int \mathbf{K}^2(u)du\right)\right)\end{aligned} \tag{1.4.8}$$

where $\Gamma^{\mu}(x)$ *is a bias term, equal to:*

$$\Gamma^{\mu}(x) = h_{n,T}^2 \rho_2(\mathbf{K}) \left(tr\left\{\mathcal{D}_{\mu,p}(x)\right\} + \frac{1}{2} tr\left\{\mathcal{H}_{\mu}(x)\right\}\right) \tag{1.4.9}$$

where,

$$\mathcal{H}_{\mu}(x) = \left(\frac{\partial^2 \mu(x)}{\partial x_j \partial x_l}\right)_{j,l=1}^{d+1} \qquad \mathcal{D}_{\mu,p}(x) = \left(\frac{\partial \mu(x)}{\partial x_j}\frac{\partial p_t(x)}{\partial x_l}\right)_{j,l=1}^{d+1}$$

Instead if, everything being equal:

$$\hat{L}^X(T,x)h_{n,T}^{d+5} \xrightarrow{a.s.} 0$$

the bias term disappears asymptotically.

Proof. See the Appendix. ∎

Remark 8. The random speed of covergence of the drift estimator depends on the occupation density of the joint process. This is a natural consequence of considering the occupation density as

the number of visits of the process in a small set which diverges to infinity as the time span grows. Therefore, the higher the dimension d of the covariate process, the slower the speed of convergence. Together with the standard dimensionality problem in nonparametric statistics, Bandi and Moloche (2008) refer to it as *double curse of dimensionality*.

Remark 9 (Bandwidth choice). The asymptotic mean squared error (AMSE) is equal to:

$$O\left(h_{n,T}^4\right) + O\left(\frac{1}{h_{n,T}^{d+1}\hat{L}^X(T,x)}\right)$$

This suggests the bandwidth parameter for the drift term being set proportionally to $\hat{L}^X(T,x)^{-\frac{1}{d+5}}$. As already pointed out in related papers, drift bandwidth selection is locally adapted in order to account for the number of visits to the point in which the estimation is performed.

Remark 10 (Stationary case). In the stationary case, we showed that $\hat{L}^X(T,x) \xrightarrow{p} T\pi(x)$. Therefore, our result can be restated as follows:

$$\sqrt{Th_{n,T}^{d+1}}\left(\hat{\mu}_{n,T}(x) - \mu(x) - \Gamma^{\mu}(x)\right) \xrightarrow{d} \sigma(x)\mathcal{N}\left(0, \left(\frac{\int \mathbf{K}^2(u)du}{\pi(x)}\right)\right)$$

as $Th_{n,T}^{d+1} \to \infty$. The bias term is now equal to:

$$\frac{h_{n,T}^2}{\pi(x)}\rho_2(\mathbf{K})\left(tr\left\{\left(\frac{\partial\mu(x)}{\partial x_j}\frac{\partial\pi(x)}{\partial x_l}\right)_{j,l=1}^{d+1}\right\} + \frac{1}{2}tr\left\{\left(\frac{\partial^2\mu(x)}{\partial x_j\partial x_l}\right)_{j,l=1}^{d+1}\right\}\right)$$

This is a standard results in conditional moments estimation (see, e.g. Pagan and Ullah, 1999, p. 101).

1.4.2 Estimation and asymptotic distribution of the diffusion coefficient

In this section we report the convergence properties of the diffusion estimator.

Theorem 1.4.4. Almost sure convergence of the diffusion estimator.

Suppose that:

$$\frac{\hat{L}^X(T,x)}{h_{n,T}^{d+1}}\left(\Delta_{n,T}\log(1/\Delta_{n,T})\right)^{1/2} \xrightarrow{a.s.} 0$$

with $\Delta_{n,T} \to 0$, $h_{n,T} \to 0$ *and* $n, T \to \infty$, *then the estimator of equation (1.4.2) converges almost*

surely to the diffusion coefficient. I.e.:

$$\hat{\sigma}^2_{n,T}(x) \xrightarrow{a.s.} \sigma^2(x) \tag{1.4.10}$$

Proof. See the Appendix. ■

Theorem 1.4.5. Asymptotic distribution of the diffusion estimator.

Suppose that:

$$\frac{\hat{L}^X(T,x)}{h^{d+1}_{n,T}} \left(\Delta_{n,T} \log(1/\Delta_{n,T})\right)^{1/2} \xrightarrow{a.s.} 0$$

$$\hat{L}^X(T,x) h^{d+1}_{n,T} \xrightarrow{a.s.} \infty$$

with $\Delta_{n,T} \to 0$, $h_{n,T} \to 0$ *and* $n, T \to \infty$, *so that:*

$$\sqrt{\frac{h^{d+5}_{n,T} \hat{L}^X(T,x)}{\Delta_{n,T}}} \xrightarrow{a.s.} 0$$

then the estimator described in equation (1.4.2) converges in distribution to a Gaussian random variable.

$$\begin{aligned} \sqrt{\frac{\hat{L}^X(T,x) h^{d+1}_{n,T}}{\Delta_{n,T}}} &\left(\hat{\sigma}^2_{n,T}(x) - \sigma^2(x)\right) \\ &\xrightarrow{d} 2\sigma^2(x) \mathcal{N}\left(0, \left(\int \mathbf{K}^2(u) du\right)\right) \end{aligned} \tag{1.4.11}$$

If, instead,

$$\sqrt{\frac{h^{d+5}_{n,T} \hat{L}^X(T,x)}{\Delta_{n,T}}} = O_{a.s.}(1)$$

then, there is an asymptotic bias term $\Gamma^{\sigma^2}(x)$, *equal to:*

$$\Gamma^{\sigma^2}(x) = h^2_{n,T} \rho_2(\mathbf{K}) \left(tr\left\{\mathcal{D}_{\sigma^2,p}(x)\right\} + \frac{1}{2} tr\left\{\mathcal{H}_{\sigma^2}(x)\right\}\right) \tag{1.4.12}$$

where,

$$\mathcal{H}_{\sigma^2}(x) = \left(\frac{\partial^2 \sigma^2(x)}{\partial x_j \partial x_l}\right)^d_{j,l=1} \qquad \mathcal{D}_{\sigma^2,p}(x) = \left(\frac{\partial \sigma^2(x)}{\partial x_j} \frac{\partial p_t(x)}{\partial x_l}\right)^d_{j,l=1}$$

Proof. See the Appendix. ■

Remark 11. It is also possible to identify the diffusion term for any fixed time horizon T. This has been already pointed out in Bandi and Moloche (2008) and goes back to a result first shown in Brugière (1993). The general results can also be applied to our setting. In the fixed T case, if one is ready to assume that:

$$h_{n,T}^{d+1} = O_{a.s.}\left(\sqrt{\Delta_{n,T}\log(1/\Delta_{n,T})}\right)$$

it is possible to show the consistency and asymptotic normality of the diffusion estimator. In particular, for $\Delta_{n,T}, h_{n,T}^{d+1} \to 0$ and $n \to \infty$, it is possible to show that:

$$\sqrt{\frac{h_{n,T}^{d+1}}{\Delta_{n,T}}}\left(\hat{\sigma}_{n,T}^2(x) - \sigma^2(x)\right) \sim MN\left(0, \frac{4\sigma^4(x)}{\hat{L}^X(T,x)}\right)$$

where, MN denotes a mixed normal distribution, with mixing factor $\hat{L}^X(T,x)$. ■

Remark 12. The asymptotic mean squared error (AMSE) is equal to:

$$O\left(h_{n,T}^4\right) + O\left(\frac{\Delta_{n,T}}{h_{n,T}^{d+1}\hat{L}^X(T,x)}\right)$$

This suggests to use again an adaptive scheme to set the bandwidth for the diffusion term. In particular, we oversmooth in areas that are less visited by the process and undersmooth in areas that are often visited. The diffusion bandwidth is therefore set proportionally to $\left(\frac{\hat{L}^X(T,x)}{\Delta_{n,T}}\right)^{-\frac{1}{d+5}}$. However, as long as the diffusion term can be identified for fixed T, we can also choose a constant bandwidth which is going to be proportional to $n^{-1/(d+5)}$. ■

1.5 An extension to long memory processes

The results presented so far are obtained under the assumption that the joint process X_t is a Markov process. However, it is possible to extend this model to allow for the marginal process Z_t to be a long memory process (e.g. fractional Brownian motion, fBM, or stochastic differential equations driven by a fBM), at least when Z_t is defined on the real line.

The problem which arises in this case is that processes driven by fBM are not semi-martingales and are not Markov[9]. Therefore our assumption 2 would completely fail.

Let $\left\{B_t^H, t \geq 0\right\}$ to be a fBM, with Hurst parameter equal to $H \in (0,1)$ and suppose that Z_t follows a stochastic differential equation driven by a B_t^H,

$$Z_t = \int_0^t \psi(s)ds + \int_0^t \xi(s)dB_t^H$$

where $\{\psi(t), t \geq 0\}$ is a $\mathcal{Z}_t$-adapted process and $\xi(t)$ is a non-vanishing deterministic function. Although Z_t is not a semimartingale in this case, one can associate to it a semi-martingale $\{J_t, t \geq 0\}$, called the *fundamental semi-martingale* such that the natural filtration $\mathcal{J}_t$ of the process J coincides with $\mathcal{Z}_t$ (Kleptsyna et al., 2000). Therefore, one can perform inference on Y_t in model 1.2.3 using J_t instead of Z_t without losing any information.

Define, for $0 < s < t$:

$$\begin{aligned} k_H(t,s) &= \kappa_H^{-1} s^{\frac{1}{2}-H}(t-s)^{\frac{1}{2}-H}, \qquad \kappa_H = 2H\Gamma\left(\frac{3}{2}-H\right)\Gamma\left(H+\frac{1}{2}\right) \\ w_t^H &= \lambda_H^{-1} t^{2-2H}, \qquad \lambda_H = \frac{2H\Gamma(3-2H)\,\Gamma\left(H+\frac{1}{2}\right)}{\Gamma\left(\frac{3}{2}-H\right)} \\ M_t^H &= \int_0^t k_H(t,s)\, dB_s^H \end{aligned}$$

where M_t^H is referred to as the *fundamental martingale* associated to the fBM B_t^H, whose quadratic variation is nothing but the function w_t^H (Norros et al., 1999).

Finally suppose that the sample paths of the function $\xi^{-1}(t)\psi(t)$ are smooth enough and define:

$$Q_t^H = \frac{d}{dw_t^H}\int_0^t k_H(t,s)\,\xi^{-1}(s)\psi(s)ds, \quad t \in [0,T]$$

We can therefore define the process J_t as:

$$J_t = \int_0^t k_H(t,s)\,\xi^{-1}(s)dZ_s$$

[9]For an extensive review of the properties of fBM and stochastic diffusions driven by fBM (see, e.g. Biagini et al., 2008; Rao, 2010)

such that (see Kleptsyna et al., 2000):

(i) J_t is a semi-martingale which admits the following decomposition:

$$J_t = \int_0^t Q_t^H(s) dw_s^H + M_t^H$$

(ii) Z_t admits a representation as a stochastic integral with respect to J_t.

(iii) the natural filtrations $\mathcal{Z}_t$ and $\mathcal{J}_t$ coincide.

We can therefore define the joint process $X_t^* = (Y_t, J_t)$ onto the natural filtration $\mathcal{X}_t^*$. Under the *fundamental semi-martingale* result and definition 1.2.1 of noncausality, the filtrations $\mathcal{X}_t$ and $\mathcal{X}_t^*$ coincide.

This equivalence between the two filtrations allows us to perform inference on Y_t by means of the process X_t^*, as long as the information carried by Z_t and J_t is the same. We can therefore restate assumption 2 as follows:

Assumption 2a. *(i)* $X_t^* \in \mathbb{R}^2$ *is Harris recurrent.*

(ii) Under $\mathcal{X}_t^*$, X_t^* *is a special semi-martingale and it admits a Doob-Meyer decomposition of the type:*

$$X_t^* = H_t^* + M_t^* \qquad \forall t \in (0, T]$$

where H_t^* *is a* $\mathcal{X}_t^*$*-predictable process and* M_t^* *is a* $\mathcal{X}_t^*$*-local martingale such that* $\mathbb{E}(M_t^* | \mathcal{X}_s^*) = 0, \forall s < t$. ■

Under this assumption, our inference results can be used to deal with the case of Z_t being a long memory process in $\mathbb{R}$.

The two following equations would be used to theoretically identify the drift and the diffusion coefficient:

$$\mathbb{E}^{x^*}\left[Y_t - y\right] = t\mu(x) + o(t) \tag{1.5.1}$$

$$\mathbb{E}^{x^*}\left[(Y_t - y)^2\right] = t\sigma^2(x) + o(t) \tag{1.5.2}$$

where $x^\star = (y, j)$. Under assumption 2a, we can apply the same estimation technique and asymptotic theory presented in previous sections.

Example 2 (Instantaneous noncausality when Z_t is a long memory process)**.** Consider Z_t a fBM of given Hurst index H, and the fundamental martingale M_t^Z, associated to Z_t. It is possible to show that (see, Norros et al., 1999):

$$W_t^Z = \frac{2H}{\sqrt{w^H}} \int_0^t s^{H-\frac{1}{2}} dM_s^Z$$

is a standard Brownian motion. We set:

$$dY_t = \mu(Y_t, Z_t)dt + \sigma(Y_t, Z_t)dB_t$$

with:

$$dB_t = \rho dW_t^Z + \sqrt{1-\rho^2} dW_t$$

where W_t is another Brownian motion, independent of W_t^Z. Using the fundamental martingale result, our inference results extend verbatim. ■

1.6 Simulations

Notwithstanding the curse of dimensionality problem which is common to nonparametric inference and which can be even more severe in the case of nonstationary diffusion processes, because of the random divergence of the occupation density, we provide here a simulation study in which the diffusion process is a function of a scalar covariate Z. This is the minimal framework that can be use to prove the reliability of our estimation procedure in finite samples. Programming has been conducted in Matlab and codes are available upon request.

We consider the following true data generating processes:

$$dY_t^{(1)} = \left(\theta_1(Z_t) - \theta_2 Y_t^{(1)}\right) dt + dB_t^{(1)} \tag{1.6.1a}$$

$$dY_t^{(2)} = \left(\theta_1(Z_t) - \theta_2 Y_t^{(2)}\right) dt + \zeta \left(Y_t^{(2)} + Z_t\right) dB_t^{(2)} \tag{1.6.1b}$$

where $\theta_2 = 2$ and $\zeta = 0.4$. The former process is a generalization of a Ornstein-Uhlenbeck process, where the drift only is function of Z and the diffusion is a constant (taken equal to one for simplicity); while the latter is a CKLS model (Chan et al., 1992), generalized to encompass the dependence on the covariate. The process Z has been taken as follows:

$$Z_t^{(1)} = W_t \tag{1.6.2a}$$

$$Z_t^{(2)} = B_t^{H=0.2} \tag{1.6.2b}$$

$$Z_t^{(3)} = B_t^{H=0.7} \tag{1.6.2c}$$

where $\{W_t\}_{t\geq 0}$ is a standard Wiener process and $\{B_t^H\}_{t\geq 0}$ is a fractional Brownian motion, with Hurst index equal to 0.2 and 0.7, respectively. Namely, the latter numerical schemes have been chosen to assess the performance of our estimate where Z is a long memory process. For the sake of simplicity, we consider $\theta_1(Z_t) = Z_t^2$ in all replications. We draw 250 paths of the processes in (1.6.1a) and (1.6.1b), using a Milstein scheme which reaches an order of approximation equal to one (Iacus, 2008).

Remark 13. Following Phillips (1973), because of the *aliasing problem* in the estimation of stochastic diffusions, when data are discretely sampled, it is not possible to identify a nonlinear drift without imposing any structural restrictions on the model. In our simulating equations, structural restrictions are coming both from the additive form of the drift and from the dependence on Z. ■

The goal of this exercise is to recover an estimate of the functional form of $\theta_1(\cdot)$.

If we hope to correctly identify both the drift and the diffusion term, we have to construct a finite sample in which dt is sufficiently small and T is sufficiently large. We therefore set $\Delta_{n,T} = 1/52$ and $n = 4800$. In practical application, this would imply weekly observations over roughly 100 years time span. However, the scope of this exercise is to check that our estimators have desirable properties. Research on the applicability of this method is in progress.

To the best of our knowledge, there is not a general theory for choosing a bandwidth parameter to estimate the occupation density of multidimensional nonstationary processes in continuous time. Moreover, the bandwidth parameter depends on the recurrence properties of the underlying

stochastic process which are difficult to assess. Following Schienle (2011), we set the bandwidth according to an adaptive scheme. For each evaluation point, we count the number of neighbours in a small interval around that point. That is, for a fixed interval I_j around the point x_j:

$$h_{n,T}(x_j) = \left(\sum_{i=1}^{n} \mathbb{1}(X_{i\Delta_{n,T}} \in I_j) \right)^{-\frac{1}{d+5}} \tag{1.6.3}$$

The estimators for the drift and the diffusion coefficient have been computed using (1.2.9) and (1.2.10), respectively. In order to recover the functional form of $\theta_1(\cdot)$, a semiparametric method has been applied. In particular, we first project the estimated drift on Y_t and Z_t using a simple linear regression model. We obtain a first estimate of θ_2, say $\hat{\theta}_2^{(1)}$. We then use this estimate to compute:

$$\hat{\theta}_1^{(1)}(z) = \frac{\sum_{i=1}^{n-1} K_h\left(Z_{i\Delta_{n,T}} - z\right)\left(\hat{\mu}(Z_{i\Delta_{n,T}}, Y_{i\Delta_{n,T}}) - \hat{\theta}_2^{(1)} Y_{i\Delta_{n,T}}\right)}{\sum_{i=1}^{n} K_h\left(Z_{i\Delta_{n,T}} - z\right)}$$

We then plug the nonparametric estimate into the first step regression in order to get a new value of θ_2, say $\hat{\theta}_2^{(2)}$, and we iterate until convergence.

The drift bandwidth parameter has been set according to the theoretical proportionality rule i.e.:

$$h_{n,T}^{dr} = c_{drift} \hat{L}^X(T,x)^{-\frac{1}{d+5}}$$

for a given constant c_{drift}.

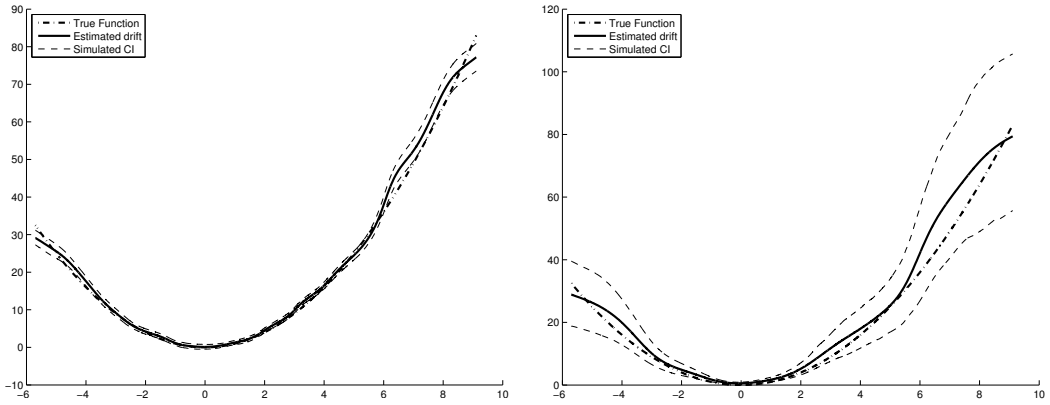
Remark 14. Bandi and Moloche (2008) suggest applying a correction factor in order to undersmooth and center at zero the asymptotic distribution. However, we do not find this correction factor having any impact in our simulation study. ■

The diffusion bandwidth has instead been taken constant and proportional to the sample size. That is:

$$h_{n,T}^{df} = n^{-\frac{1}{d+5}}$$

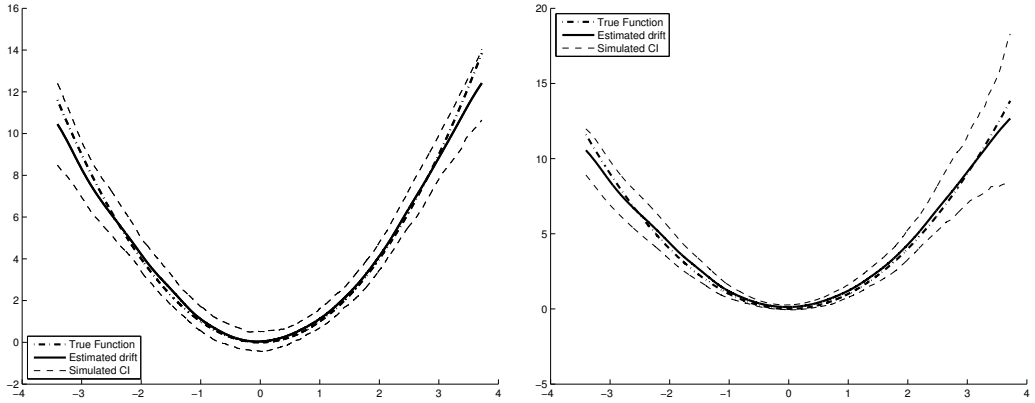
We report separately the results for the estimation of the drift, for models 1.6.1a and 1.6.1b. We also draw simulated confidence bands over the interval $2.5\% - 97.5\%$.

Figure 1.1: Estimation of $\theta_1(\cdot)$ when Z_t is drawn from 1.6.2a, with 250 simulated paths.



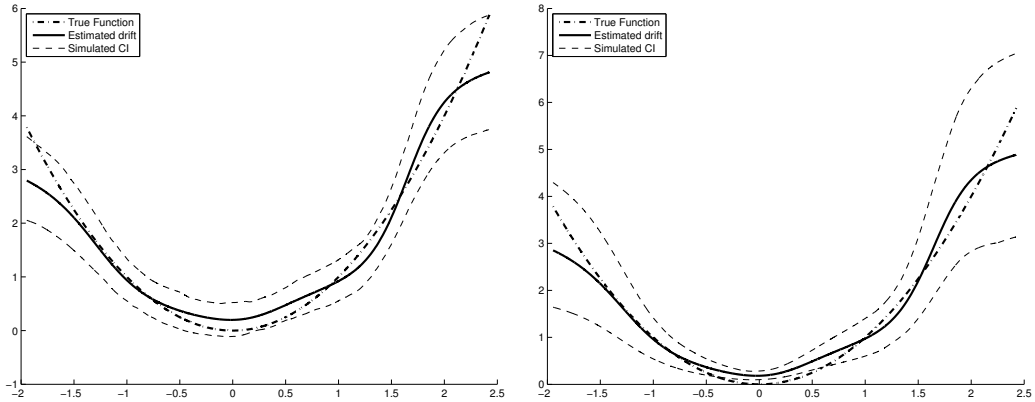
(a) Model 1.6.1a (b) Model 1.6.1b

Figure 1.2: Estimation of $\theta_1(\cdot)$ when Z_t is drawn from 1.6.2b, with 250 simulated paths.



(a) Model 1.6.1a (b) Model 1.6.1b

Figure 1.3: Estimation of $\theta_1(\cdot)$ when Z_t is drawn from 1.6.2c, with 250 simulated paths.



(a) Model 1.6.1a (b) Model 1.6.1b

As it can be seen from figures 1.1 and 1.2, estimation of the drift is rather satisfactory, despite a poorer behaviour at the boundaries.

In order to complete our simulation study, we analyse the case in which the assumption of non-causality does not hold and our inference procedure fails. For simplicity, we only consider the case in which Y_t is generated according to (1.6.1*b*); and Z_t is a plain brownian motion. Bandwidths are chosen as before.

Consider the example of simultaneous equation models in continuous time:

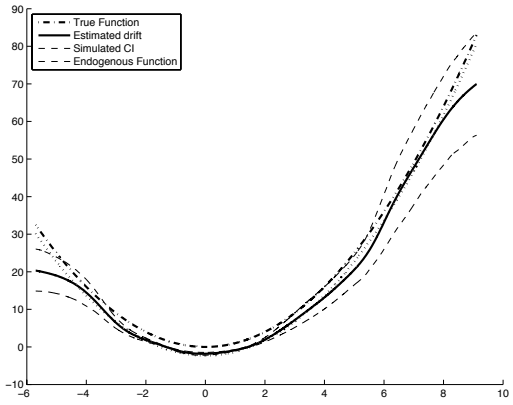
$$dY_t^{(2)} = \left(\theta_1(Z_t) - \theta_2 Y_t^{(2)}\right) dt + \zeta \left(Y_t^{(2)} + \frac{dZ_t}{\sqrt{dt}}\right) dB_t^{(2)}$$

where Z_t a standard Brownian motion, and

$$dB_t^{(2)} = \rho dZ_t + \sqrt{1-\rho^2} dW_t$$

with $\rho = -0.8$ and W_t another standard Brownian motion independent of Z_t[10]. In this case, Z_t is predictable in $\mathcal{X}_t$, so that $B_t^{(2)}$ is not a martingale on the joint filtration. Results are reported in figure (1.4) and we can clearly see that there is a sort of endogeneity bias in the estimation. For completeness, we have also plotted the function which is actually estimated (with improper terminology we call it endogenous function). The bias in the estimation is exactly equal to $\zeta\rho/\sqrt{dt}$.

Figure 1.4: Estimation of $\theta_1(\cdot)$ when Z_t is a predictable BM correlated with the brownian increments, with 250 simulated paths.



[10] dZ_t has been rescaled by $\sqrt{dt}$ only to make the effect more visible in the figure. This does not alter our result.

1.7 An Application to Uncovered Interest Parity

In continuous time, the Uncovered Interest Parity (UIP) may be expressed as the first order stochastic differential equation:

$$\mathbb{E}\left(ds_t|\mathcal{S}_t\right) = r_t dt$$

where ds_t is the instantaneous change in the log exchange rate, $\mathcal{S}_t$ is the filtration of s up to time t, and r_t is the yield differential between domestic and foreign currency denominated debt. We can use our model to test for UIP to hold by using the generic specification:

$$ds_t = \mu(r_t)dt + \sigma(r_t)dB_t \tag{1.7.1}$$

It is often standard to assume that the interest rate differential follows a random-walk. However, there is no consensus in the literature about it being $I(0)$ or $I(1)$[11]. Here, we do not make any assumption about the *DGP* followed by r_t. Instead, we assume that the interest rate differential is globally not caused by the exchange rate. Notice that this is a higher level assumption, as it encompasses the case in which r_t is a random walk; and that our inference is robust to the case in which r_t has long memory. Finally, we assume that the joint process (s_t, r_t) is Harris recurrent.

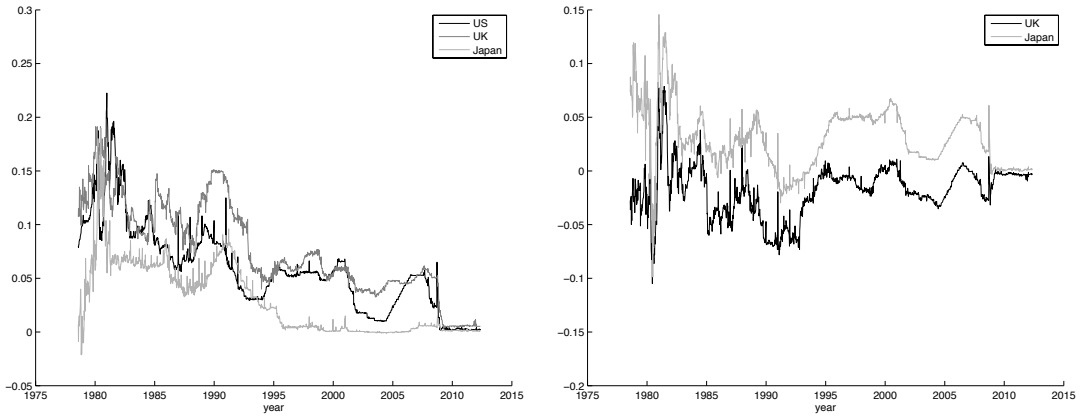
We collected data about the one-week Eurocurrency rates in the US, the UK and Japan. The exchange rate are collected weekly, and denominated in dollars per unit of foreign currency (British Pound or Japanese Yen). Data spans from August 3^{rd} 1978 to May 10^{th} 2012. All series have been downloaded from *Datastream*.

The bandwidths for the drift and the diffusion estimation in equation (1.7.1) have been chosen adaptively, as in section (1.6), using a preliminary estimator of the local time of the process r_t.

Figure 1.6 depicts the results of our estimation. The estimator of the drift coefficient (left panel) clearly rejects the UIP, both for the UK and Japan, as the curves are negatively sloped. This result is consistent with the so-called *forward premium anomaly*, which has been widely reported by the existing literature (see, e.g. Backus et al., 2001), i.e. the tendency of high interest rate currency to

[11] This property is usually tested by verifying that the spot and forward exchange rate are cointegrated. Evans and Lewis (1995) cannot reject that the interest rate differential is $I(1)$, while, e.g., Zivot (2000) do reject. Baillie and Bollerslev (1994) conclude that the interest rate differential has long memory properties, with Hurst parameter between 0.5 and 1.

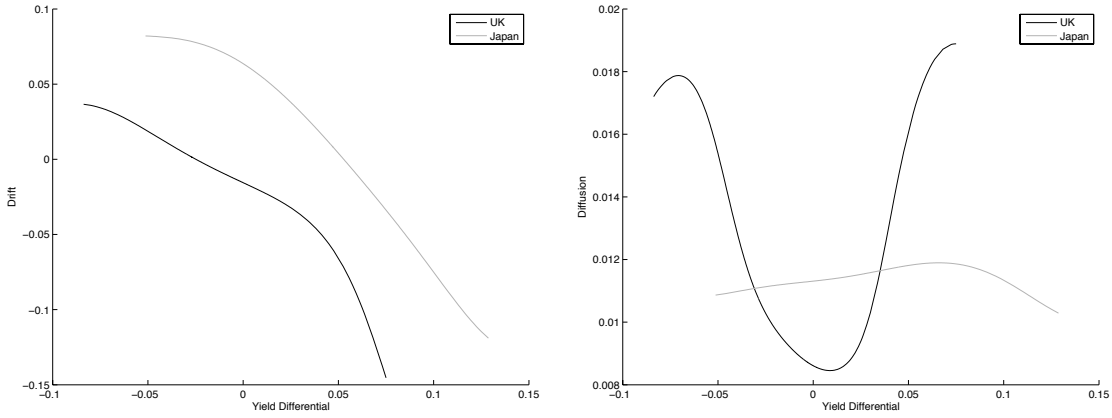
Figure 1.5: Data on Eurocurrency rates for the US, the UK and Japan.



(a) Rates (b) Yield differential with respect to the US

appreciate, when the UIP predicts instead that such currencies should depreciate. The estimators of the diffusion coefficient (right panel) are instead substantially different. The black curve for the UK suggests a linear diffusion coefficient; while the grey line for Japan suggests the diffusion being constant. This difference might be explained as a consequence of central bank interventions in the foreign exchange market, as argued in Mark and Moh (2007). However, given the historically low level of interest rates in Japan, the conditional volatility of the exchange rate can be related to different factors but the yield differential.

Figure 1.6: Nonparametric Estimation of 1.7.1 for UK and Japan.



(a) Drift $\hat{\mu}(r_t)$ (b) Diffusion squared $\hat{\sigma}^2(r_t)$

1.8 Conclusions

We propose in this paper a methodological approach to conditional nonstationary diffusion models in continuous time. Our goal is to provide a wider set of hypothesis on the conditional and marginal process such that a simple nonparametric inference can be applied. In particular, we argue that our approach is flexible as it allows the marginal process Z_t to be any Harris recurrent Feller process and, in some particular case, also a long memory process.

We also believe that our theoretical results improve what has been done so far in the literature on Harris recurrent stochastic processes, by tuning some of the underlying assumptions.

Finally, we stress that this framework can be of interest both in finance and macroeconomics. Our final application on UIP briefly depicts how our approach can be relevant in practice.

1.9 Appendix

1.9.1 General Definitions, Corollaries and Theorems.

Definition 1.9.1 (Harris Recurrence Azéma et al., 1969)**.** A strongly Markov process X taking values in a Polish space $(E, \mathcal{E})$ is Harris recurrent, if there exists some $\sigma-$finite measure m on $(E, \mathcal{E})$, such that:

$$m(A) > 0 \Rightarrow \forall x \in E \quad : \qquad P_x\left(\int_0^\infty \mathbb{1}_A(X_s)ds = \infty\right) = 1$$

This process is also called $m-irreducible$. ■

Definition 1.9.2 (Höpfner and Löcherbach, 2003)**.** A Harris recurrent process X, taking values in a Polish space $(E, \mathcal{E})$, with invariant measure m is called positive recurrent (or ergodic) if $m(E) < \infty$, null recurrent if $m(E) = \infty$. ■

Theorem 1.9.3 (Ratio Limit Theorem Azéma et al., 1969)**.** *If a process X is Harris recurrent with invariant measure m and A and B are two integrable additive functionals and if $\| \nu_B \| > 0$, then:*

(i) $\lim_{t\to\infty} \frac{\mathbb{E}_x(A_t)}{\mathbb{E}_x(B_t)} = \frac{\|\nu_A\|}{\|\nu_B\|} \qquad m-a.s.,$

(ii) $\lim_{t\to\infty} \frac{A_t}{B_t} = \frac{\|\nu_A\|}{\|\nu_B\|} \qquad P_x-a.s.,\ \forall x.$

■

Definition 1.9.4 (Modulus of Continuity of Multivariate Brownian Semimartingales)**.** Suppose X is a special multivariate Brownian semimartingale, and denote:

$$\kappa_{n,T} = \sup_{|t-s|<\Delta_{n,T},[0\le s<t\le T]} |X_t - X_s|$$

to be its modulus of continuity. We can then write (McKean, 1969):

$$\mathbb{P}\left[\lim \sup_{\Delta_{n,T}\to 0} \frac{\kappa_{n,T}}{\sqrt{\Delta_{n,T}\,(1/\Delta_{n,T})}} = \max_{t\le T} \sqrt{2\gamma(X_t)}\right] = 1$$

where $\gamma(X_t)$ is the biggest eigenvalue of the covariance matrix of the process X. ■

1.9.2 Proof of Lemma (1.4.1)

We want to prove that:

$$\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)\xrightarrow{a.s.}\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)ds$$

We start by writing:

$$\begin{aligned}
&\left|\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)-\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)ds\right|\\
\leq&\left|\frac{1}{h_{n,T}^{d+1}}\sum_{i=0}^{n-1}\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\left[\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)-\mathbf{K}_{h_{n,T}}\left(X_s-x\right)\right]ds\right.\\
&\left.-\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\mathbf{K}_{h_{n,T}}\left(X_{0\Delta_{n,T}}-x\right)+\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\mathbf{K}_{h_{n,T}}\left(X_{n\Delta_{n,T}}-x\right)\right|\\
\leq&\frac{1}{h_{n,T}^{d+1}}\left|\sum_{i=0}^{n-1}\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\left[\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)-\mathbf{K}_{h_{n,T}}\left(X_s-x\right)\right]ds\right|+O\left(\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\right)\\
\leq&\frac{1}{h_{n,T}^{d+1}}\sum_{i=0}^{n-1}\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}D\left(\frac{X_s-x}{h_{n,T}^{d+1}},\frac{\kappa_{n,T}}{h_{n,T}^{d+1}}\right)\left|\frac{X_{i\Delta_{n,T}}-X_s}{h_{n,T}^{d+1}}\right|ds\\
\leq&\frac{\kappa_{n,T}}{h_{n,T}^{d+1}}\int_0^T\frac{1}{h_{n,T}^{d+1}}D\left(\frac{X_s-x}{h_{n,T}^{d+1}},\frac{\kappa_{n,T}}{h_{n,T}^{d+1}}\right)ds
\end{aligned}$$

by the triangle inequality and assumption (3). Finally using the *Ratio Limit theorem*, we have that:

$$\int_0^T\frac{1}{h_{n,T}^{d+1}}D\left(\frac{X_s-x}{h_{n,T}^{d+1}},\frac{\kappa_{n,T}}{h_{n,T}^{d+1}}\right)ds=O_{a.s.}\left(\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)ds\right)$$

By theorem (1.3.2), we now have that, for $n,T\to\infty$:

$$\frac{\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)ds}{t^\alpha/l(t)}\to\mathbb{E}_m\left(\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)ds\right)W^\alpha$$

Therefore, to prove our final result, we only need to prove that:

$$\mathbb{E}_m\left(\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)ds\right)=Cp_t(x)\tag{1.9.1}$$

By the strong version of the Ratio Limit Theorem, for any couple of integrable functions $f(\cdot)$ and $g(\cdot)$, we have that:

$$\frac{\mathbb{E}_m(f)}{\mathbb{E}_m(g)} = \frac{m(f)}{m(g)}$$

which implies:

$$\mathbb{E}_m(f) = Cm(f) \qquad \text{where} \qquad C = \frac{m(g)}{\mathbb{E}_m(g)}$$

We can then write:

$$\begin{aligned}
&\mathbb{E}_m\left(\frac{1}{h_{n,T}^{d+1}}\int_0^T \mathbf{K}_{h_{n,T}}(X_s - x)\,ds\right) = C\int_{\mathcal{E}} \frac{1}{h_{n,T}^{d+1}}\mathbf{K}_{h_{n,T}}(X_s - x)\,m(dX_s)\\
&= \int_{\mathcal{E}} \frac{1}{h_{n,T}^{d+1}}\mathbf{K}_{h_{n,T}}(X_s - x)\,p_\infty(X_s)\lambda(dX_s) = \int_{\mathcal{E}} \frac{1}{h_{n,T}^{d+1}}\mathbf{K}(u)p_\infty(uh_{n,T} + x)\lambda(h_{n,T}du)\\
&= \int_{\mathcal{E}} \mathbf{K}(u)p_\infty(uh_{n,T} + x)\lambda(du)
\end{aligned}$$

where we use the continuity of m wrt λ and the properties of the Lebesgue measure (Billingsley, 1979, Theorem 12.2, p.172). Finally, as $h_{n,T}^{d+1} \to 0$:

$$\int_{\mathcal{E}} \mathbf{K}(u)p_t(uh_{n,T}^{d+1} + x)\lambda(du) \to p_\infty(x)\int_{\mathcal{E}} \mathbf{K}(u)\lambda(du) = p_\infty(x)$$

By the relation between Riemann and Lebesgue integration and assumption (3). This concludes the proof.

1.9.3 Proof of Theorem (1.4.2)

We want to prove that:

$$\hat{\mu}_{n,T}(x) \xrightarrow{a.s.} \mu(x)$$

We start by writing the drift estimator of equation (1.4.1) as follows:

$$\begin{aligned}
&\hat{\mu}_{n,T}(x)\\
&= \frac{\frac{1}{h_{n,T}^{d+1}}\sum_{i=1}^{n-1}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\mu(X_s)ds}{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)}
\end{aligned} \tag{1.9.2}$$

$$+\frac{\frac{1}{h_{n,T}^{d+1}}\sum_{i=1}^{n-1}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\sigma(X_s)dB_s}{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)} \tag{1.9.3}$$

We start with the numerator of equation (1.9.2). We want to prove that:

$$\begin{aligned}&\frac{1}{h_{n,T}^{d+1}}\sum_{i=1}^{n-1}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\mu(X_s)ds\\ \xrightarrow{a.s.}&\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)\mu(X_s)ds\end{aligned} \tag{1.9.4}$$

We start by writing:

$$\begin{aligned}&\left|\frac{1}{h_{n,T}^{d+1}}\sum_{i=1}^{n-1}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\mu(X_s)ds-\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)\mu(X_s)ds\right|\\
\leq&\left|\frac{1}{h_{n,T}^{d+1}}\sum_{i=0}^{n-1}\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\left[\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)-\mathbf{K}_{h_{n,T}}\left(X_s-x\right)\right]\mu(X_s)ds\right.\\
&\left.-\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\mathbf{K}_h\left(X_{0\Delta_{n,T}}-x\right)\mu(X_{0\Delta_{n,T}})\right|\\
\leq&\left|\frac{1}{h_{n,T}^{d+1}}\sum_{i=0}^{n-1}\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\left[\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)-\mathbf{K}_{h_{n,T}}\left(X_s-x\right)\right]\mu(X_s)ds\right|\\
&+\left|\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\mathbf{K}_h\left(X_{0\Delta_{n,T}}-x\right)\mu(X_{0\Delta_{n,T}})\right|\\
\leq&\frac{\kappa_{n,T}}{h_{n,T}^{d+1}}\left|\frac{1}{h_{n,T}^{d+1}}\int_0^T D\left(\frac{X_s-x}{h_{n,T}^{d+1}},\frac{\kappa_{n,T}}{h_{n,T}^{d+1}}\right)\mu(X_s)ds\right|+O_{a.s.}\left(\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\right)\\
\leq&\frac{\kappa_{n,T}}{h_{n,T}^{d+1}}\left(\frac{1}{h_{n,T}^{d+1}}\int_0^T D\left(\frac{X_s-x}{h_{n,T}^{d+1}},\frac{\kappa_{n,T}}{h_{n,T}^{d+1}}\right)|\mu(X_s)|\,ds\right)+O_{a.s.}\left(\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\right)\end{aligned}$$

by the triangle inequality, the continuity of $\mu(\cdot)$, and assumption (3). Finally using the *Ratio Limit theorem*, we have that:

$$\frac{1}{h_{n,T}^{d+1}}\int_0^T D\left(\frac{X_s-x}{h_{n,T}^{d+1}},\frac{\kappa_{n,T}}{h_{n,T}^{d+1}}\right)|\mu(X_s)|\,ds=O_{a.s.}\left(\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)ds\right)$$

We are now left with the following expression:

$$\frac{\frac{1}{h_{n,T}^{d+1}}\int_0^T \mathbf{K}_{h_{n,T}}(X_s - x)\,\mu(X_s)ds + O_{a.s.}\left(\frac{(\Delta_{n,T}\log(1/\Delta_{n,T}))^{1/2}\hat{L}^X(T,x)}{h_{n,T}^{d+1}}\right)}{\frac{1}{h_{n,T}^{d+1}}\int_0^T \mathbf{K}_{h_{n,T}}(X_s - x)\,ds + O_{a.s.}\left(\frac{(\Delta_{n,T}\log(1/\Delta_{n,T}))^{1/2}\hat{L}^X(T,x)}{h_{n,T}^{d+1}}\right)}$$

We have now to prove that this converges to the true functional form of the drift coefficient. We denote the true functional as $\mu(x)$ and write the following equation:

$$\frac{\frac{1}{h_{n,T}^{d+1}}\int_0^T \mathbf{K}_{h_{n,T}}(X_s - x)\,(\mu(X_s) - \mu(x))\,ds}{\frac{1}{h_{n,T}^{d+1}}\int_0^T \mathbf{K}_{h_{n,T}}(X_s - x)\,ds}$$

We want to show that the numerator converges almost surely to 0. To do so, we exploit the Lipschitz continuity property of the drift function. Write:

$$\begin{aligned}
&\left|\frac{1}{h_{n,T}^{d+1}}\int_0^T \mathbf{K}_{h_{n,T}}(X_s - x)\,(\mu(X_s) - \mu(x))\,ds\right| \\
&\leq \frac{1}{h_{n,T}^{d+1}}\int_0^T \left|\mathbf{K}_{h_{n,T}}(X_s - x)\right|\,|\mu(X_s) - \mu(x)|\,ds \\
&\leq \frac{C}{h_{n,T}^{d+1}}\int_0^T \left|\mathbf{K}_{h_{n,T}}(X_s - x)\right|\,|X_s - x|\,ds \leq C(\kappa_{n,T})\frac{1}{h_{n,T}^{d+1}}\int_0^T \mathbf{K}_{h_{n,T}}(X_s - x)\,ds \\
&= C(\kappa_{n,T})O_{a.s.}\left(\frac{1}{h_{n,T}^{d+1}}\int_0^T \mathbf{K}_{h_{n,T}}(X_s - x)\,ds\right)
\end{aligned}$$

which gives the desired result.

In order to prove that equation (1.9.3) converges to zero almost surely, we proceed as follows. We notice that, as in Bandi and Phillips (2003), the numerator of the equation can be embedded in a continuous time martingale for any value of $X_{i\Delta_{n,T}}$. As a matter of fact we have:

$$\beta_{(i+1)\Delta_{n,T}} = \int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}} \sigma(X_s)dB_s$$

is a stochastic integral which is $\mathcal{Y}_{(i+1)\Delta_{n,T}} \vee \mathcal{Z}_{(i+1)\Delta_{n,T}}$-measurable and such that $\mathbb{E}\left[\beta_{(i+1)\Delta_{n,T}}\right] = 0$. Moreover by Itô isometry (see Øksendal, 2003, Lemma 3.15, p. 26):

$$var(\beta_{(i+1)\Delta_{n,T}}) = \mathbb{E}\left[\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}} \sigma(X_s)dB_s\right]^2 = \mathbb{E}\left[\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}} \sigma^2(X_s)ds\right] < \infty$$

We can therefore construct the following continuous martingale:

$$\begin{aligned} M^{X_{i\Delta_{n,T}}}(r) =& \sqrt{h_{n,T}^{d+1}}\left(\frac{1}{h_{n,T}^{d+1}}\sum_{i=1}^{[(n-1)r]}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\sigma(X_s)dB_s\right) \\ =& \frac{1}{\sqrt{h_{n,T}^{d+1}}}\sum_{i=1}^{[(n-1)r]}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\sigma(X_s)dB_s \end{aligned} \tag{1.9.5}$$

whose quadratic variation is equal to:

$$\left[M^{X_{i\Delta_{n,T}}}(r)\right] = \frac{1}{h_{n,T}^{d+1}}\sum_{i=1}^{[(n-1)r]}\mathbf{K}_h^2\left(X_{i\Delta_{n,T}}-x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\sigma^2(X_s)ds \tag{1.9.6}$$

Using the same method applied for equation (1.9.2) and using the *Ratio Limit theorem*, we can show that:

$$\left[M^{X_{i\Delta_{n,T}}}(1)\right] = O_{a.s.}\left(\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)ds\right) \tag{1.9.7}$$

Finally, as in Phillips and Ploberger (1996), expanding the probability space as needed:

$$\left(M^{X_{i\Delta_{n,T}}}(1)\right)^2 / \left[M^{X_{i\Delta_{n,T}}}(1)\right] = O_{a.s.}(1)$$

which gives:

$$\sqrt{\hat{L}^X(T,x)h_{n,T}^{d+1}}\left(\frac{\frac{1}{\Delta_{n,T}}\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n-1}\mathbf{K}_{h_{n,T}}\left(X_s-x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\sigma(X_s)dB_s}{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_s-x\right)}\right) = O_{a.s.}(1)$$

Therefore, the term in equation (1.9.3) converges almost surely to zero, provided that $\hat{L}^X(T,x)h_{n,T}^{d+1}\xrightarrow{a.s.}\infty$. This completes the proof.

1.9.4 Proof of Theorem (1.4.3)

We start by decomposing the estimator into a bias and a variance component:

$$\underbrace{(1.9.2)-\mu(x)}_{\text{BIAS}}+\underbrace{(1.9.3)}_{\text{VARIANCE}}$$

We start by analyzing the variance term. We use again the fact that this term can be written as a sequence of martingale components. Namely, we know that every martingale array can be written as a time changed *Dambis, Dubins-Schwartz* Brownian motion. We call τ, the time change associated to $M^{X_{i\Delta_{n,T}}}(1)$. This implies:

$$\frac{M_{\tau}^{X_{i\Delta_{n,T}}}(1)}{\sqrt{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_s-x\right)}}\xrightarrow{d}\mathcal{N}\left(0,\frac{\left[M_{\tau}^{X_{i\Delta_{n,T}}}(1)\right]}{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_s-x\right)}\right)$$

Using dominated convergence and the *Ratio Limit Theorem*, we can show that the numerator of the variance of $M_{\tau}^{X_{i\Delta_{n,T}}}(1)$ converges to:

$$\begin{aligned}&\frac{1}{h_{n,T}^{d+1}}\sum_{i=1}^{n-1}\mathbf{K}_h^2\left(X_{i\Delta_{n,T}}-x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\sigma^2(X_s)ds\\&\xrightarrow{a.s.}\sigma^2(x)\left(\int\mathbf{K}^2(u)du\right)\end{aligned}\tag{1.9.8}$$

Now, we turn to the bias term. Write the bias term in the following way:

$$\begin{aligned}&\frac{\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)\left(\mu(X_s)-\mu(x)\right)ds}{\frac{1}{h_{n,T}^{d+1}}\int_0^T\mathbf{K}_{h_{n,T}}\left(X_s-x\right)ds}\\&\xrightarrow{a.s.}\frac{\frac{1}{h_{n,T}^{d+1}}\int\mathbf{K}(u)\left(\mu(x+uh_{n,T})-\mu(x)\right)p_t(x+uh_{n,T})\lambda(du)}{\frac{1}{h_{n,T}}\int\mathbf{K}(u)p(x+uh_{n,T}^{d+1})\lambda(du)}\end{aligned}$$

We therefore compute the Taylor expansion of this function around x.

$$\frac{\int\mathbf{K}(u)\left[h_{n,T}\sum_{j=1}^{d+1}\frac{\partial\mu(x)}{\partial x_j}u_j+\frac{h_{n,T}^2}{2}\sum_{j,l=1}^{d+1}\frac{\partial^2\mu(x)}{\partial x_j\partial x_l}u_ju_l\right]\left[p_t(x)+h_{n,T}\sum_{j=1}^{d+1}\frac{\partial p_t(x)}{\partial x_j}u_j\right]\lambda(du)}{\int\mathbf{K}(u)\left[p_t(x)+h_{n,T}\sum_{j=1}^{d+1}\frac{\partial p_t(x)}{\partial x_j}u_j+o(h_{n,T})\right]\lambda(du)}$$

Using the symmetry of kernels and neglecting terms of order higher than $h_{n,T}^2$ leads to:

$$\int\mathbf{K}(u)\left[h_{n,T}^2\left(\sum_{j,l=1}^{d+1}\frac{\partial\mu(x)}{\partial x_j}\frac{\frac{\partial p_t(x)}{\partial x_l}}{p_t(x)}u_ju_l\right)+\frac{h_{n,T}^2}{2}\left(\sum_{j,l=1}^{d+1}\frac{\partial^2\mu(x)}{\partial x_j\partial x_l}u_ju_l\right)\right]\lambda(du)$$

We define

$$\mathcal{H}_\mu(x) = \left(\frac{\partial^2 \mu(x)}{\partial x_j \partial x_l}\right)_{j,l=1}^{d} \qquad \mathcal{D}_{\mu,p}(x) = \left(\frac{\partial \mu(x)}{\partial x_j}\frac{\partial p_t(x)}{\partial x_l}\right)_{j,l=1}^{d}$$

where $\mathcal{H}_\mu(x)$ is the symmetric hessian matrix of the function μ and we rewrite the bias term as follows:

$$\begin{aligned}
&h_{n,T}^2 tr\left\{\int \mathbf{K}(u)u'\left(\mathcal{D}_{\mu,p}(x) + \frac{1}{2}\mathcal{H}_\mu(x)\right)u\lambda(du)\right\}\\
=&h_{n,T}^2 tr\left\{\left(\mathcal{D}_{\mu,p}(x) + \frac{1}{2}\mathcal{H}_\mu(x)\right)\int \mathbf{K}(u)uu'\lambda(du)\right\}\\
=&h_{n,T}^2 \rho_2(\mathbf{K})\left(tr\left\{\mathcal{D}_{\mu,\lambda(du)p}(x)\right\} + \frac{1}{2}tr\left\{\mathcal{H}_\mu(x)\right\}\right)
\end{aligned}$$

using the properties of the trace operator, the relation between Lebesgue and Riemann integration and assumption (3).

1.9.5 Proof of Theorem (1.4.4)

We want to prove that:

$$\hat{\sigma}_{n,T}^2(x) \xrightarrow{a.s.} \sigma^2(x)$$

Using Itô's lemma, we can show that $(Y_{(i+1)\Delta_{n,T}} - Y_{i\Delta_{n,T}})^2$ satisfies the following SDP:

$$\begin{aligned}
(Y_{(i+1)\Delta_{n,T}} - Y_{i\Delta_{n,T}})^2 =& \int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}} \left(2(Y_s - Y_{i\Delta_{n,T}})\mu(X_s) + \sigma^2(X_s)\right)ds\\
&+ \int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}} 2(Y_s - Y_{i\Delta_{n,T}})\sigma(X_s)dB_s
\end{aligned}$$

This leads us to decompose equation (1.4.2) as follows:

$$\begin{aligned}
&\hat{\sigma}_{n,T}^2(x)\\
&= \frac{1}{\Delta_{n,T}} \frac{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n-1}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}\sigma^2(X_s)ds}{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)} \qquad (1.9.9)
\end{aligned}$$

$$+ \frac{1}{\Delta_{n,T}} \frac{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n-1}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}2(Y_s - Y_{i\Delta_{n,T}})\sigma(X_s)dB_s}{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}} - x\right)} \qquad (1.9.10)$$

$$+\frac{1}{\Delta_{n,T}}\frac{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n-1}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)\int_{i\Delta_{n,T}}^{(i+1)\Delta_{n,T}}2(Y_s-Y_{i\Delta_{n,T}})\mu(X_s)ds}{\frac{\Delta_{n,T}}{h_{n,T}^{d+1}}\sum_{i=1}^{n}\mathbf{K}_{h_{n,T}}\left(X_{i\Delta_{n,T}}-x\right)} \tag{1.9.11}$$

In order to prove consistency of the diffusion term, we treat the drift as a nuisance parameter. As in the proof of theorem (1.4.2), using dominated convergence, the properties of the diffusion function and the *Ratio Limit Theorem*, we can prove that equation (1.9.9) almost surely converges to the true value of the diffusion term, as long as $\frac{\hat{L}^X(T,x)}{h_{n,T}^{d+1}}\left(\Delta_{n,T}\log(1/\Delta_{n,T})\right)^{1/2}=O_{a.s.}(1)$.

For equation (1.9.10) and equation (1.9.11), we follow Florens-Zmirou (1993) and Bandi and Phillips (2003). The term in $(Y_s-Y_{i\Delta_{n,T}})$ is a semi-martingale, so that we can use Burkholder-Davis-Gundy inequality (see,e.g. Protter, 2003, Theorem 48, p. 193) to show that its expectation can be bounded by the square root of its quadratic variation which converges at a rate equal to $\sqrt{\Delta_{n,T}}$. Therefore, following the proof of theorem (1.4.2), the component in equation (1.9.10) can be embedded in a continuous martingale whose expectation converges to zero as long as $\sqrt{\frac{\hat{L}^X(T,x)h_{n,T}^{d+1}}{\Delta_{n,T}}}$ diverges to infinity. In the same way, the term in (1.9.11) is bounded as long as $\sqrt{\frac{\hat{L}^X(T,x)h_{n,T}^{d+1}}{\Delta_{n,T}}}$ diverges (Bandi and Phillips, 2003; Bandi and Moloche, 2008).

1.9.6 Proof of theorem (1.4.5)

Using the same procedure as in theorem (1.4.3), we decompose our estimator into a bias and a variance component:

$$\underbrace{(1.9.9)-\sigma^2(x)}_{\text{BIAS}}+\underbrace{(1.9.10)+(1.9.11)}_{\text{VARIANCE}}$$

For the variance, the component in equation (1.9.11) converges to zero almost surely as noted in the previous proof. Using the Ratio Limit theorem we can prove that equation (1.9.10) converges in distribution to a normal with variance equal to:

$$4\sigma^4(x)\left(\int\mathbf{K}^2(u)du\right) \tag{1.9.12}$$

We then turn to the bias term. We can follow the same procedure that for theorem (1.4.3). Define:

$$\mathcal{H}_{\sigma^2}(x) = \left(\frac{\partial^2\sigma^2(x)}{\partial x_j \partial x_l}\right)_{j,l=1}^{d} \qquad \mathcal{D}_{\sigma^2,p}(x) = \left(\frac{\partial\sigma^2(x)}{\partial x_j}\frac{\partial p_t(x)}{\partial x_l}\right)_{j,l=1}^{d}$$

where $\mathcal{H}_\sigma(x)$ is the symmetric hessian matrix of the function σ. Then the bias term is equal to:

$$h_{n,T}^2 \rho_2(\mathbf{K})\left(tr\left\{\mathcal{D}_{\sigma^2,p}(x)\right\} + \frac{1}{2}tr\left\{\mathcal{H}_{\sigma^2}(x)\right\}\right)$$

1.9.7 Additional Proofs

Theorem 1.9.5. *Suppose Y_t is a stationary process conditionally on Z_t and Z_t is Harris Recurrent. Then $X_t = (Y_t, Z_t)$ is a joint Harris Recurrent process.*

Proof. Remember that X_t lies in a Polish space $(E, \mathcal{E})$. We have to show that there exists a measure m, such that:

$$0 < m(A) < \infty \qquad \forall A \subset \mathcal{E}$$

i.e. a $\sigma-$finite measure on E, such that X is m-irreducible (see Definition 1.9.1).

We start to show that, for every set A and $t \to \infty$, if a measure exists, it is $\sigma-$finite. Take any set $A \subset \mathcal{E}$, such that $A = B \times C$, where B and C are compact, with $Z_{s+1} \in B$ and $Y_{s+1} \in C$. We denote by ϕ_z the invariant measure of the process Z_t and by $\pi(y|z)$ the stationary probability measure of Y given Z. We can write down the transition probability for the joint process, under the markovianity of X, as:

$$\begin{aligned}
& \int_0^\infty \mathbb{P}\left(X_{s+1} \in A | X_s\right) ds \\
= & \int_0^\infty \mathbb{P}\left(Z_{s+1} \in B, Y_{s+1} \in C | Z_s, Y_s\right) ds \\
= & \int_0^\infty \mathbb{P}\left(Z_{s+1} \in B | Z_s\right) \mathbb{P}\left(Y_{s+1} \in C | Z_s, Y_s, Z_{s+1} \in B\right) ds \\
\leq & \left(\int_0^\infty \mathbb{P}\left(Z_{s+1} \in B | Z_s\right) ds\right)\left(\int_0^\infty \mathbb{P}\left(Y_{s+1} \in C | Z_s, Y_s, Z_{s+1} \in B\right) ds\right) \\
= & \left(\int \mathbb{P}\left(Z_{s+1} \in B\right) \phi_z(dz)\right)\left(\int_0^\infty \mathbb{P}\left(Y_{s+1} \in C | Z_s, Y_s, Z_{s+1} \in B\right) ds\right) \\
= & \left(\int \mathbb{P}\left(Z_{s+1} \in B\right) \phi_z(dz)\right)\left(\int \mathbb{P}\left(Y_{s+1} \in C | Z_{s+1} \in B\right) \pi(dy|z)\right)
\end{aligned}$$

with a straightforward application of Bayes' theorem. Finally:

$$\phi_z(B) = \int \mathbb{P}\left(Z_{s+1} \in B\right) \phi_z(dz) < \infty$$

since A is bounded, and:

$$\pi(y \in C | z \in B) = \int \mathbb{P}\left(Y_{s+1} \in C | Z_{s+1} \in B\right) \pi(dy|z) \in (0,1]$$

This implies:

$$\int_0^\infty \mathbb{P}\left(X_{s+1} \in A | X_s\right) ds < \infty \tag{1.9.13}$$

Therefore, for every set A, there exists a $\sigma-$finite measure for X. This concludes the first part of the proof.

Now, denote $\tau_A = inf\{t \geq 0, X_t \in A\}$, the hitting time of set A, for a given realization of X_t, $x = (z, y) \notin A$. For any arbitrary measure m:

$$\mathbb{P}^x(\tau_A < \infty) = 1 \tag{1.9.14}$$

implies $m(A) > 0$ (Revuz, 1984). We set $\tau_B^z = inf\{t \geq 0, Z_t \in B\}$ and $\tau_C^y = inf\{t \geq 0, Y_t \in C\}$. Then define:

$$\begin{aligned} \mathbb{P}^x(\tau_A < \infty) &= \mathbb{P}^x(\tau_B^z < \infty, \tau_C^y < \infty) \\ = \quad \mathbb{P}^x(\tau_B^z < \infty) &\mathbb{P}^x(\tau_C^y < \infty | \tau_B^z < \infty) \end{aligned}$$

where the conditional probability is well defined since τ_B^z is a stopping time and $\{\tau_B^z < \infty\} \in \mathcal{Z}_\infty$ (Protter, 2003). Since Y is stationary conditional on Z, we have that:

$$\mathbb{E}^x(\tau_C^y | \tau_B^z < \infty) < \infty$$

which implies:

$$\{\sup_{t \geq 0, \tau_B^z < \infty} \tau_C^y\} < \infty \quad \rightarrow \quad \mathbb{P}^x(\tau_C^y < \infty | \tau_B^z < \infty) = 1$$

We then obtain (1.9.14), from the Harris recurrence of Z.

Therefore, for every set A, X is m-irreducible and m is a $\sigma-$finite measure by (1.9.13). By definition (1.9.1), X is Harris recurrent. This concludes the proof. ■

Chapter 2

On the Choice of the Regularization Parameter in Nonparametric Instrumental Regressions

Abstract

This paper discusses in details the implementation of nonparametric instrumental regressions with adaptive choice of the regularization parameter when a Tikhonov scheme is used to estimate the unknown function of the endogenous variable. A leave-one-out cross validation criterion is proposed which is rate optimal in mean squared error, upon some regularity conditions on the regression function. This result is further extended to the general case of the estimation of functional derivatives of any order. A numerical simulation shows that this selection 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 type of goods. This application emphasizes the properties, the flexibility and the simplicity of the methodology presented in this work, irrespective of the nonparametric approach chosen to estimate the conditional mean functions.

2.1 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 an unobservable random error term, which, according to the setting under study, is supposed to satisfy some independence condition with respect to the predictors. These independence conditions enable 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. Because, for instance, the error term contains unobservable regressors that are likely to be correlated with the observed independent variables; or because the causality structure between the response and the predictors is reversed, i.e. the dependent variable is somehow affecting the regressors. In econometrics, this problem is usually referred as endogeneity of the predictors, i.e., the dependent and the independent variables are simultaneously determined by the unobservables. This endogeneity issue does not allow 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.

Suppose, for instance, that the relation between the response variable Y, the predictors Z and a random error U could be defined by the following additively separable model:

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

with φ being a smooth function. In the standard setting, when Z is exogenous, the mean independence condition $\mathbb{E}(U|Z = z) = 0$, implies that:

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

Hence, φ is the conditional expectation of the response variable Y given the predictor Z. However, if the mean independence condition does not hold anymore, the unknown function φ cannot be defined as such.

Instrumental variables are a standard approach in econometrics to identify and estimate functional dependency in the presence of endogenous regressors. The main idea is to suppose to observe a set of variables, defined as W and called *instruments*, such that they enjoy some correlation with the endogenous predictors and they satisfy the independence condition with respect to the random component. In the example of the separable model (4.2.1a), one has:

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

i.e., the error term in (4.2.1a) 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, among others).

This assumption allows to eliminate the noise term in (4.2.1a), by taking the expectation with respect to W. Hence, our object of interest, the function φ, is now implicitly defined by the equation:

$$\mathbb{E}(\varphi(Z)|W) = r \tag{2.1.2}$$

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

As an example of an application of this framework, 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; Hoderlein and Holzmann, 2011; Horowitz, 2011). The Engel curve describes the expansion path for commodity demands as the household's budget increases. Therefore, to estimate its shape, 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 Z. However, the latter is likely to be jointly determined with individual demands, and hence it has to be considered as an endogenous regressor in the estimation of consumer expansion paths. Therefore, empirical studies that aim at obtaining meaningful results about the *structural* shape of the Engel curve shall take this endogeneity problem into account for identification.

As discussed in Blundell et al. (2007), the allocation model of income to individual consumption goods and savings suggests exogenous sources of income to 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. Hence, the shape of the Engel curve can be identified by using gross income as an instrument for total expenditure.

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 (2SLS) regressions. However, this imposes several restrictions on the shape of φ, that may or may not be justified by the economic theory.[1] For instance, the recent empirical study by Blundell et al. (2007) shows that 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). Therefore, a parametric specification might not be appropriate for the empirical application discussed above. More generally, the researcher would like to maintain some flexibility in the specification of the function φ. Hence, this paper focuses on the fully nonparametric estimation of the regression function (Hall and Horowitz, 2005; Darolles et al., 2011*a*).

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

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 (3.2.2), a direct estimation of φ is not feasible as it requires to disentangle φ from its conditional expectation with respect to W. Namely, equation (3.2.2) can be rewritten as:

$$\int \varphi(z) f(z|w) dz = r \tag{2.1.3}$$

where $f(z|w)$ is the conditional distribution of Z given W and it defines a Fredholm integral equation of the first kind (Kress, 1999). 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 *naif* way to look at the *ill-posedness* of the inverse problem is to imagine the integral operator in equation (2.1.3) as an infinite dimensional matrix. This matrix is one-to-one and therefore invertible, 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 choose a constant parameter which transform the ill-posed into a well-posed inverse problem.

Therefore, in the application to nonparametric instrumental variable regressions, the implementation of these regularization methods requires, beside the usual issues related to the nonparametric estimation (e.g., selection of the smoothing parameters), also the selection of this regularization parameter. A sound criterion for the choice of this parameter is extremely important, as an erroneous alternative can lead to misleading conclusions about the shape of the function of interest. In particular, it would be necessary to provide data-driven procedures for this choice, which, in many cases, remains arbitrary.

[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*.

The aim of this article is to discuss the selection of the regularization parameter in nonparametric instrumental variable regressions when the so-called Tikhonov regularization is applied (Darolles et al., 2011*a*). In particular, a *leave-one-out* cross validation criterion is proposed here and its properties discussed. Moreover, its advantages in relation to existing procedures are examined (see, e.g. Fève and Florens, 2010). Finally, the article provides an application to the estimation of the Engel curve for food, fuel and leisure, using a sample of UK households.

Under a different regularization technique (Galerkin), Marteau and Loubes (2012) discuss the properties of the adaptive selection of the regularization parameter when the conditional expectation operator in (2.1.3) is known. They prove an Oracle inequality for their minimization criterion. Horowitz (2012) extends their framework in the case 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 function φ.

The closest in spirit to this work is Fève and Florens (2010). They 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 estimated counterpart of equation (3.2.2), 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. 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.

Cross Validation (CV) has been already advocated as a viable solution to choose the regularization parameter in case of penalized Ridge regressions (Wahba, 1977), and for ill-posed solutions of integral equations of the first kind (Vogel, 2002). Similarly, Golub et al. (1979); Lukas (1993, 2006) discuss the application of Generalized Cross Validation (GCV) to Ridge regressions and to the linear inverse problem in mathematical statistics respectively. GCV is generally preferred to CV as it does not require the computation of the estimator at each sample point and, therefore, reduces computation time tremendously. However, it 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.

To the best of our knowledge, there is not a theoretical work that discusses the properties of CV

in the case of nonparametric instrumental regressions. This paper fills this gap.

In particular, it provides a detailed discussion about the selection of the regularization parameter and its relation to the so-called *source condition*. Finally, it presents a numerical simulation in which the robustness of the cross validation procedure is shown with respect to the smoothness properties of the function φ for a given joint distribution of Z and W.

2.2 The main framework

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

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

For simplicity, the assumption that W and Z are defined on the unit hypercube of dimension $p+q$ is maintained. Suppose further that $\varphi \in \mathbb{L}^2_Z$, the space of square integrable functions of Z. Define T, the conditional expectation operator which maps $\mathbb{L}^2_Z$ into $\mathbb{L}^2_W$, and its adjoint T^*. Further denote by $\{\varphi_i, \psi_i, i \geq 0\}$, two orthonormal sequences in $\mathbb{L}^2_Z$ and $\mathbb{L}^2_W$, respectively. In the following, Y is supposed to be observed, although the results of this paper applies also to the case in which Y is latent and the researcher observes $\tilde{Y} = \mathbb{1}(Y > 0)$, a binary transformation of it (see Centorrino and Florens, 2013).

Our framework needs the following high level assumption.

Assumption 4. *The joint distribution of the instruments W and the endogenous variable Z is dominated by the product of the marginal distributions and its density, $f_{Z,W}(z,w)$, 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 2.2.1. *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) $T\varphi_i = \lambda_i \psi_i$

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

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

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

Using operator's notations, equation (3.2.2) can be rewritten as follows:

$$T\varphi = r \tag{2.2.2}$$

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:

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

As stressed in Darolles et al. (2011*a*), Assumption (4) *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 3 (The Normal Case)**.** Suppose that $(Z, W) \in \mathbb{R}^2$ 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 Z 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 $(\sqrt{\rho^2})^j$. Notice that, as $j \to \infty$, the eigenvalues are converging to 0, which causes the *ill-posedness* of the problem.

Finally assume that all other necessary identification conditions are satisfied (Andrews, 2011;

Darolles et al., 2011*a*; D'Haultfoeuille, 2011). In particular, the following *completeness condition* is supposed to hold throughout the paper:

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

This condition is related to the concept of completeness in statistics. In particular, this condition implies that every non-constant and square integrable function of Z is correlated with some square integrable function of W.

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

$$\varphi^\alpha = \underset{\varphi \in \mathbb{L}_Z^2}{\arg\min} \|T\varphi - r\|^2 + \alpha\|\varphi\|^2 \tag{2.2.3}$$

where α is called the penalization (or regularization) 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 (Z, 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 5 (Source condition)**.** *For some real $\beta > 0$, and a set of functions $g \in \mathbb{L}_Z^2$ and $h \in \mathbb{L}_W^2$, one has:*

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

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

$$\varphi = (T^*T)^{\frac{\beta}{2}} v \qquad \text{, i.e.} \qquad \varphi \in \mathcal{R}\left((T^*T)^{\frac{\beta}{2}}\right)$$

which clearly links the properties of the function φ with the ones of the joint distribution of (Z, W) (see also Chen and Reiss, 2011).

Under this assumption, one obtains that the rate of convergence of the regularization bias is the 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 to exploit higher orders of regularity of the function φ. In fact:

$$\|\varphi^\alpha_{(k)} - \varphi\|^2 = O_p\left(\alpha^{\min(\beta,2k)}\right) \qquad \forall k \geq 1 \tag{2.2.4}$$

In the following, $\varphi^\alpha_{(1)} = \varphi^\alpha$ and it is referred to as the non-iterated Tikhonov solution of (2.2.3).

2.3 Nonparametric estimation and the choice of α

Suppose to observe $\{(y_i, z_i, w_i), i = 1, \ldots, N\}$, an iid realization of the random variables (Y, Z, 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 extend the framework to stationary mixing time series, see Hansen (2008)

parameter h.[4] For simplicity, the same bandwidth h_N is used for both Z 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 Z, i.e.:

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

(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 2.3.1 (Darolles et al. 2011*a*)**.** *Under assumptions (4) and (5), and the convergence of the regularization bias given in (2.2.4):*

$$\|\hat{\varphi}^{\alpha_N} - \varphi\|^2 = O_P\left[\frac{1}{\alpha^2}\left(\frac{1}{N} + h_N^{2\rho}\right) + \left(\frac{1}{N h_N^{p+q}} + h_N^{2\rho}\right)\alpha_N^{\min(\beta-1,0)} + \alpha_N^{\min(\beta,2)}\right] \tag{2.3.1}$$

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

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

$N^{-1/2\rho}$, to exploit the parametric rate of convergence of the first term in (2.3.1). 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 and the instrument, p, q, and of 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 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 (2.3.1) asymptotically 0 can therefore be expressed in terms of β and γ.

Corollary 2.3.2 (Convergence of the upper bound to 0 and rate optimal α_N)**.** *The rate optimal value of α_N, for which (2.3.1) $\xrightarrow{a.s.}$ 0, is such that:*

(i) If $\beta \geq 1$ and $0 < \gamma \leq 1$, so that $N^\gamma \alpha_N^2 \to \infty$, and $N h_N^{p+q} \to \infty$, 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 a such a way that $N^\gamma \alpha_N^2 \to \infty$, and $N h_N^{p+q} \to \infty$, 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 a 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 a 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. (i) If $\beta \geq 1$, the second term of the upper bound in (2.3.1) 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.

■

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 bias of the nonparametric estimator. Therefore:

$$\gamma = \frac{2\rho}{2\rho+p+q} \tag{2.3.2}$$

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 2.3.1). 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 $\gamma = 1$:

$$\rho > \frac{p+q}{2} \tag{2.3.3}$$

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 seldomly 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 instrument used. A similar reasoning applies if the value of γ is chosen too small. In this case, the bias in the nonparametric estimation is going to play the role of further slowing down the convergence of (2.3.1) to 0.

When $\beta < 1$, the choice of the bandwidths impacts 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 (5). If γ is taken equal to 1, point *iii* of Corollary (2.3.2) shows again that one needs condition (2.3.3) 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 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 (2.3.1) is designed in a 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 (2.3.2) 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 (2.3.1) and for a given dimension of the endogenous variable Z, constraint (2.3.3) 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 (2.2.2) 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 fixing γ as in (2.3.2). 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., 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 paper considers criteria of the form:

$$P(\alpha_N)\|\hat{T}\hat{\varphi}^{\alpha_N} - \hat{r}\|^2 \tag{2.3.4}$$

where $P(\alpha_N)$ is a penalization function. These criteria selects α_N as the minimizer of the sum of squared residuals in (2.2.2).

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:

$$SSR(\alpha_N) = \frac{1}{\alpha_N}\|\hat{T}\hat{\varphi}_{(2)}^{\alpha_N} - \hat{r}\|^2 \tag{2.3.5}$$

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 (2.3.4), 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 2.3.3. *The $SSR(\alpha_N)$ criterion 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. 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 A4 in Darolles et al. (2011*a*), 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:

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

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:

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

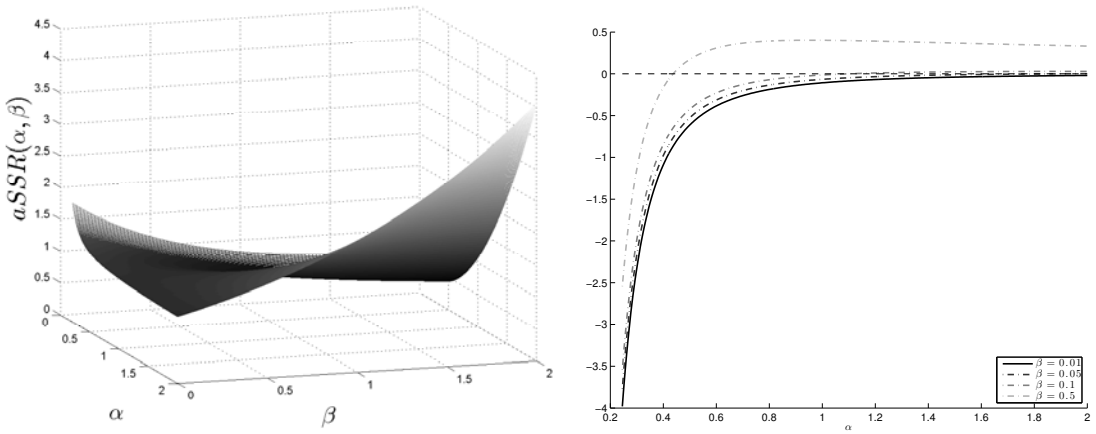
which gives:

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

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

This criterion has the same speed of convergence as the original MSE in (2.3.1). Therefore, upon the optimal choice of the bandwidth, theoretically, α is selected in such a way that the variance and the bias term converges 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 0 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 the parameter.

Therefore, for $\beta < 1$, the SSR criterion may lead to *over-regularize* the solution of the inverse



(c) (d)

Figure 2.1: A 3 dimensional plot of $aSSR(\alpha_N, \beta)$ (left), and its derivative wrt α_N for several values of β (right).

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 (2.1) graphically illustrates the issue. On the left panel, the function $aSSR(\cdot,\cdot)$ is plotted for $N = 1000$, $\rho = 2$, $p = 2$, $q = 1$, and for a reasonable range of values for the two parameters α_N and β, with γ as in (2.3.2).

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 a 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 paper discusses a *leave-one-out* procedure for the selection of the regularization parameter. Define the cross validation function:

$$CV(\alpha_N) = \|\hat{T}\hat{\varphi}_{(-i)}^{\alpha_N} - \hat{r}\|^2 \tag{2.3.6}$$

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 2.3.4. *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. First notice that minimizing the cross validation function (2.3.6) 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:

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

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 (see the Appendix). 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:

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

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:

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

■

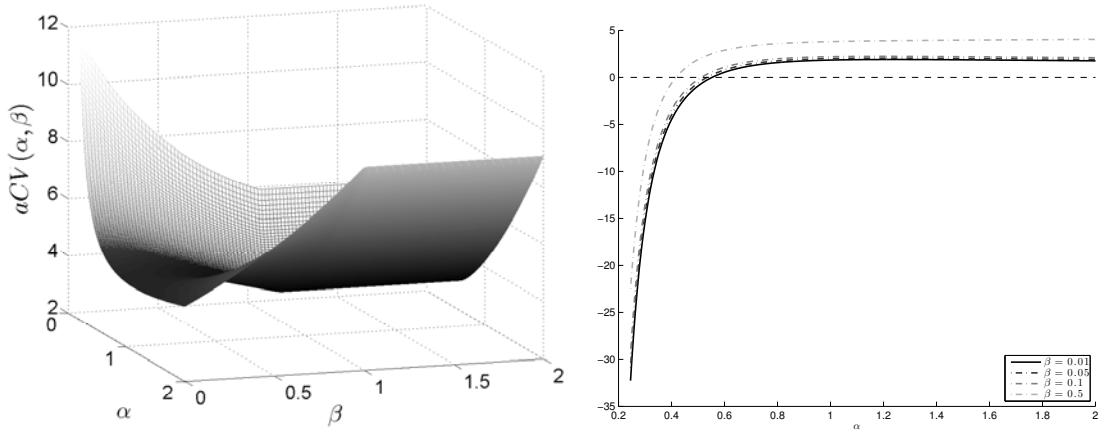
An example about the behavior of this criterion function is reported in figure (2.2). Consider, as before, a case in which $N = 1000$, $\rho = 2$, $p = 2$, $q = 1$, and the bandwidth is chosen such that:

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

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 belong to the family (2.3.4). The penalizing factor is tantamount to:

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



(a) (b)

Figure 2.2: A 3 dimensional plot of $aCV(\alpha_N, \beta)$ (left), and its derivative wrt α_N for several values of β (right).

which also contains the penalizing factor $1/\alpha_N$. However, it has also two additional terms: 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 2.3.5. *For an optimal choice of the smoothing parameter h, the minimization of the cross validation function (2.3.6) 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).[5] 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 (2.2.2) (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 more slowly 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).

However, for higher values of β, it is feasible to achieve the optimal rate of using the same idea as in the SSR method of Fève and Florens (2010), i.e., to increase the qualification of the regularization procedure with an iterated Tikhonov approach. 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

[5] 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.

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.

2.4 A more general approach to the Regularization in Hilbert Scale

Following the result in the previous section, it can be actually shown that the cross validation procedure of this paper 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:

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

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:

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

where $B = TL^{-s}$ and $\varphi^{(s),\alpha}$ is the regularized s^{th} derivative. A detailed explanation on how to approximate L^s, at least when s is equal to 1, is given in Centorrino et al. (2013a) and Florens and Racine (2012).[6] This section explores the extension of the CV criterion of theorem (2.3.4) to this more general case.

The assumptions stated in section (2.2) are maintained here. In particular, the operator T is assumed to be one to one and the solution φ exists.[7] 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 6. *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 (Z, W). Then for B defined as above, $|\nu| \leq 1$ and $s \geq 1$, Assumption (6) implies the following inequality (see Engl et al., 2000, Corollary 8.22, p. 214):

$$\underline{c}(\nu)\|g\|_{-\nu(a+s)} \leq \|\left(B^*B\right)^{\nu/2} g\| \leq \overline{c}(\nu)\|g\|_{-\nu(a+s)} \tag{2.4.1}$$

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 (2.4.1) implies that:

$$\mathcal{D}\left((B^*B)^{\nu/2}\right) = \mathcal{D}\left(L^{\nu(a+s)}\right) \tag{2.4.2}$$

Furthermore, Assumption (6), together with the fact that $\varphi \in \mathcal{D}(L^u)$ implies the source condition (5), with $\beta = u/a$ (Carrasco et al., 2013; Florens et al., 2011). Heuristically, this can be explained

[6] 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).

[7] See Florens et al. (2011) for the non identified case in penalized Tikhonov regularization.

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 (2.4.2) implies $\varphi \in \mathcal{D}\left((B^*B)^{\frac{u-s}{2(a+s)}}\right)$. Therefore, there exists a vector $v \in \mathbb{L}_Z^2$, 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).

Under these assumptions, the main result of this section follows.

Theorem 2.4.1. *Suppose that φ is u times differentiable, and that assumption (6) holds. Suppose further that φ is estimated by penalization of its s^{th} derivative, where $u \leq a + 2s$. Then, the cross validation criterion (2.3.6) 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. Following the proof of Theorem (2.3.4), 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 $\|\varphi\|_u < \infty$.

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

Throughout the proof, I use the following inequalities (see Engl et al., 2000):

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

together with Assumption (6) and inequality (2.4.1), with:

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

which explains why one needs to assume that $u \leq a + 2s$. 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:

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

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:

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

and

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

Finally:

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

which gives the desired result. ■

For $s = 0$, the result of Theorem (2.4.1) is just a generalization of Theorem (2.3.4). Note further that, following Florens et al. (2011), the penalization by derivatives increases the qualification of the

Tihkonov regularization, upon the assumption that T is one-to-one. Finally, when the bandwidth is chosen optimally, i.e. $h \approx N^{-1/(2\rho+p+q)}$, the second term of the asymptotic expansion is dominated by the first one, given the constraints on u and a. This finally implies that the optimal α is chosen in such a way that:

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

Again, this selection of the optimal parameter attains the same rate as the discrepancy principle of Morozov (see Engl et al., 2000). Moreover, it embeds the case presented in corollary (2.3.5), when $s = 0$.

2.5 A Numerical Illustration

In order to illustrate the small sample properties of our cross validation procedure and to compare it to existing methods, a simulation scheme similar to the one employed in Hall and Horowitz (2005) is considered.

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

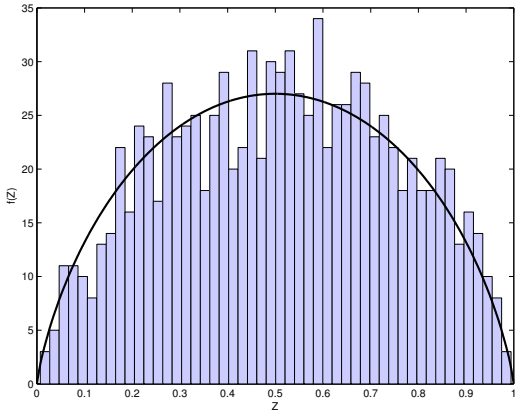
$$f_{ZW}(z,w) = 2C_f \sum_{i=1}^{\infty} (-1)^{i+1} i^{-b/2} \sin(i\pi z) \sin(i\pi w)$$
$$\varphi(z) = \sqrt{2} \sum_{i=1}^{\infty} (-1)^{i+1} i^{-a} \sin(i\pi z)$$
$$Y = \mathbb{E}(\varphi(Z)|W = w) + V$$

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

Note that the value of a and b respectively controls 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 condition has to hold:

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

Figure 2.3: Marginal density of Z and W, with one draw using slice sampling.



with $a > 1/2$ and $b > 1$ (see Hall and Horowitz, 2005; Darolles et al., 2011*a*).[8]

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, Z and W have the same marginal distribution, which is depicted in figure (2.3). Note that in the former numerical study $\beta < 0.75$, while in the latter $\beta < 1.75$. 1000 paths of the endogenous variable Z, the instrument W and the error V are simulated. Epanechnikov kernels of order 2 are employed. The conditional expectation operators T and T^* are estimated as the matrix of kernel weights from the nonparametric regressions of Y on W, and of $\hat{r} = \mathbb{E}(Y|W)$ on Z (see also Fève and Florens, 2010; Centorrino et al., 2013a). 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:

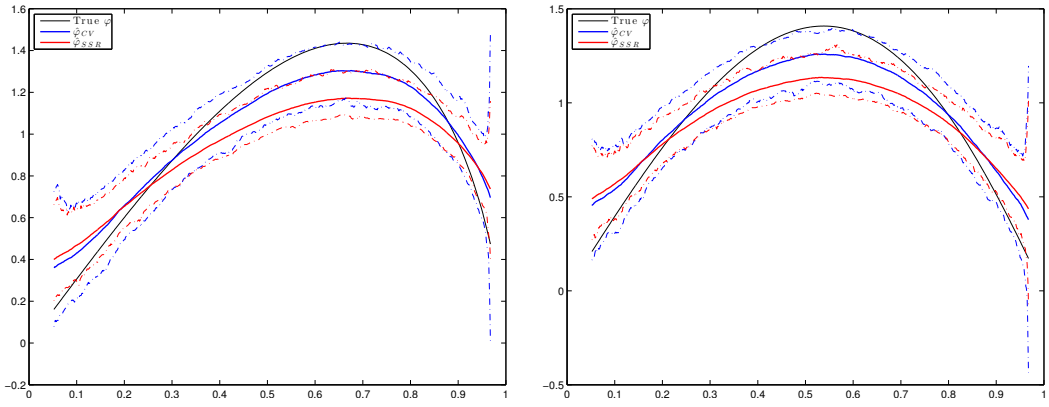
$$\alpha^{OPT} = \arg\min_{\alpha>0} \|\hat{\varphi}^{\alpha} - \varphi\|^2$$

Notice that this criterion produces the optimal value of α, given the estimation error.

Results of the numerical study are reported in Figure (2.4). 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

[8] Note that 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 relevant for this paper.

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



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

Figure 2.4: Estimation of the function φ using the CV and the SSR criterion respectively, with penalization of the function.

the φ^{SSR} estimator in terms of fitting. This implies a lower bias and a higher variance of the former estimator. The simulated pointwise 95% 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 function φ.

Another comparison between the two vectors of α's is reported in table (2.1). Summary statistics for the vector of α^{CV}, α^{SSR} 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, although the distribution of both α^{CV} and of α^{SSR} are shifted on the right, compared to the one of α^{OPT}.

		Mean	Median	St.Dev	Min	Max
$a = 2$	α^{CV}	0.0426	0.0399	0.0110	0.0229	0.1252
	α^{SSR}	0.1214	0.1222	0.0184	0.0263	0.1734
	α^{OPT}	0.0263	0.0250	0.0074	0.0099	0.0612
$a = 4$	α^{CV}	0.0475	0.0446	0.0121	0.0210	0.1177
	α^{SSR}	0.1207	0.1220	0.0181	0.0238	0.1792
	α^{OPT}	0.0270	0.0256	0.0075	0.0119	0.0592

Table 2.1: Summary statistics for the regularization parameter, with penalization of the function.

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

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

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:

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

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). Also, recentering and multiplying both sides by the inverse of the pdf function of Z is immaterial in our case. Thus, one can obtain B^* simply as:

$$\left(B^*\lambda\right)(u) = \int \lambda(w)S(u,w)dw$$

This can be approximated by the matrix of survivor weights of Z. Denote by $K_h(\cdot)$ a positive and

[10] For a detailed description of the implementation the reader is referred to Florens and Racine (2012) and Centorrino et al. (2013a).

symmetric kernel with (possibly) unbounded support, and define:

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

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

$$\hat{S}_z = \left[1 - \mathcal{K}_h\left(\frac{z - z_i}{h_z}\right)\right]_{i=1}^{N}$$

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

$$\hat{B}^* = \hat{S}_z$$

Hence the Tihkonov regularized estimator with penalized first derivative is defined as:

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

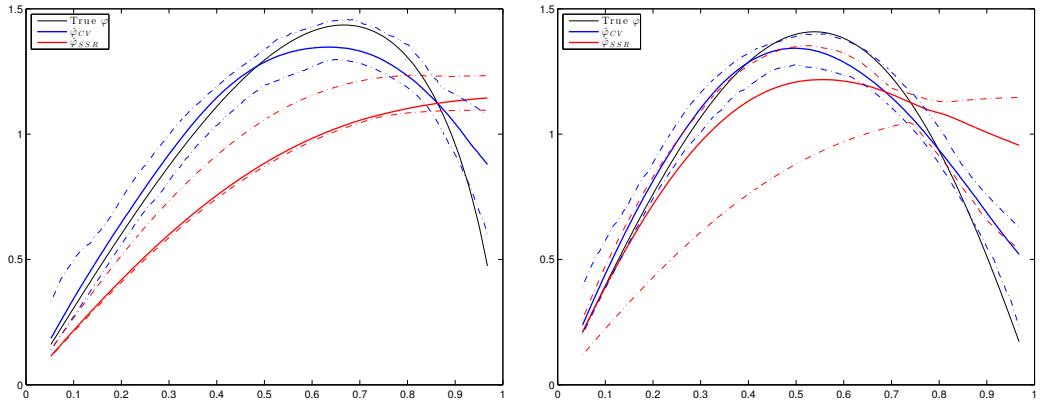
The SSR criterion of Fève and Florens (2010) has been extended to this case by Fève and Florens (2013). They generalize the SSR criterion by taking as penalizing term the squared norm of the estimator $\hat{\varphi}^\alpha$, i.e.,

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

The implementation of the CV criterion remains instead unchanged. Results of this numerical simulations are reported in figure (2.5), both for the case in which $a = b = 2$ (left panel), and for the case $a = 4$ and $b = 2$ (right panel).

It is evident from the figures that the cross validation criterion outperforms the modified SSR criterion. Also, it fulfills our theoretical predictions. As the qualification of Tikhonov regularization increases by penalizing the first derivative and the function of interest is infinitely smooth, the estimator clearly improves. Moreover, it improves more when the function is relatively less smooth ($a = 2$), which is again consistent with theoretical findings. Coverage of both functions also improves in this case.

Finally table (2.2) reports the summary statistics for the two vectors of α's. Once again the α^{CV}



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

Figure 2.5: Estimation of the function φ using the CV and the SSR criterion respectively, with penalization of the first derivative of the function.

has a substantially smaller mean than the α^{SSR} and a very small variance, which indicates its good properties with respect to sample selection. 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 values are once again consistently greater than the optimal ones.

		Mean	Median	St.Dev	Min	Max
$a = 2$	α^{CV}	0.00020	0.00021	0.00013	0.00004	0.00091
	α^{SSR}	0.10883	0.11146	0.01154	0.00583	0.11146
	α^{OPT}	0.00008	0.00005	0.00005	0.00005	0.00047
$a = 4$	α^{CV}	0.00032	0.00031	0.00014	0.00003	0.00095
	α^{SSR}	0.02217	0.00717	0.03265	0.00045	0.10766
	α^{OPT}	0.00010	0.00008	0.00006	0.00005	0.00049

Table 2.2: Summary statistics for the regularization parameter, with penalization of the first derivative of the function.

2.6 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 adaptive 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 (4.2.1a) is used to estimate the structural shape of the Engel curve, where Y is the budget share for each individual commodity; Z is the logarithm of total expenditure; and W is the logarithm of gross total income. That is:

$$Y = \varphi(Z) + U \tag{2.6.1}$$

$$\mathbb{E}(U|W) = 0 \tag{2.6.2}$$

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 paper 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.[11] This paper focuses on the estimation of the Engel curve for three categories of nondurables and services: food, fuel, and leisure. Table (2.3) reports some summary statistics for these data.

In order to show the flexibility of the approach of this paper, the application is presented under several estimation of the conditional expectation functions. In particular, both local constant and

[11] 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.

	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 2.3: Summary statistics UK Family Expenditure Survey.

local linear kernels and cubic B-spline bases 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 Centorrino et al. (2013a). For comparison, the estimator of the simple nonparametric regression of Y on Z is considered. Notice that, in the spirit of Blundell and Horowitz (2007), if the function obtained with the simple nonparametric regression, i.e. 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.[12]

Figures (2.6), (2.7) and (2.8) present the result of such an application for food, fuel and leisure respectively. 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 servces 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 increasing toward higher total expenditure. Finally, leisure is, as expected, a luxury service as the Engel curve is nondecreasing in total expenditure.

Another important aspect to notice is that the local linear and the B-spline specification for leisure seem to indicate that there is not any endogeneity problem in such a case. As a matter of fact, the simple curve obtained from the nonparametric regression of the share of expenditure on leisure and total expenditure is fully included in the 95% confidence interval obtained from bootstrapping

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

the nonparametric instrumental regression estimator. This can be due to expenditure on leisure not systematically planned by the household.

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

A final assessment of the performance of this estimator is reported in figure (2.9), (2.10) and (2.11). 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 Z is treated as exogenous is also reported for completeness.[13]

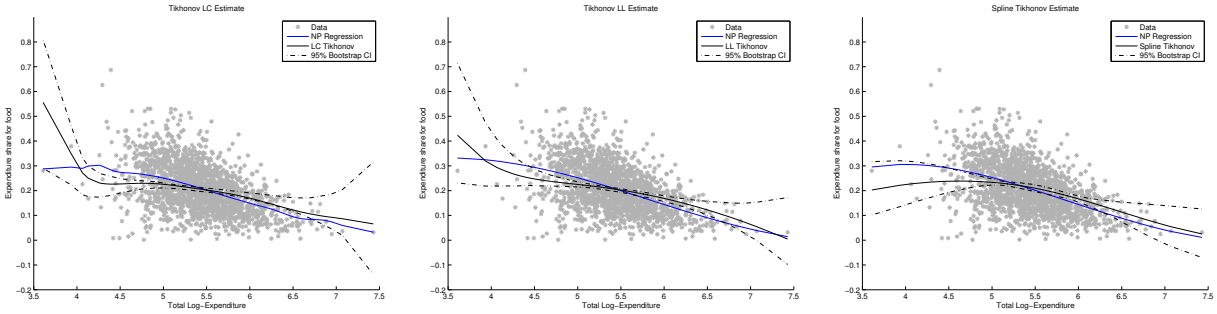
Results are consistent with those previously discussed. The estimators of each derivative are roughly constant, with indicates the Engel curve to be linearly decreasing (increasing).

2.7 Conclusions

This paper discusses the theoretical properties of a leave-one-out cross validation criterion for the selection of the regularization parameter in nonparametric instrumental regressions, when the Tikhonov scheme is used in order to estimate the function of interest. It is shown that this criterion is rate optimal in mean squared error, i.e., it delivers a regularization constant which possesses the same rate as the theoretical one, depending on the value of the regularity index β. The method proposed here outperforms in a simulation study existing data-driven criteria and can be easily extended to the case in which penalization is on the derivatives of the function rather than on the function itself. Hence, 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. Finally, an empirical application to the estimation of the Engel curve in a sample of

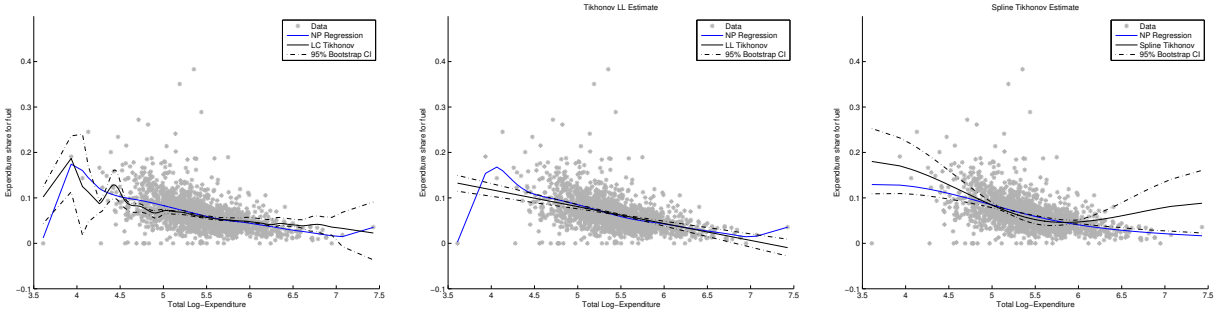
[13]However, as already pointed out in related work (Florens and Racine, 2012), the two are not directly comparable. As a matter of fact, in stardand nonparametric regression, the estimation of the nonparametric derivative is *self-consistent*, i.e. it is obtained as derivative of the conditional mean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

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.



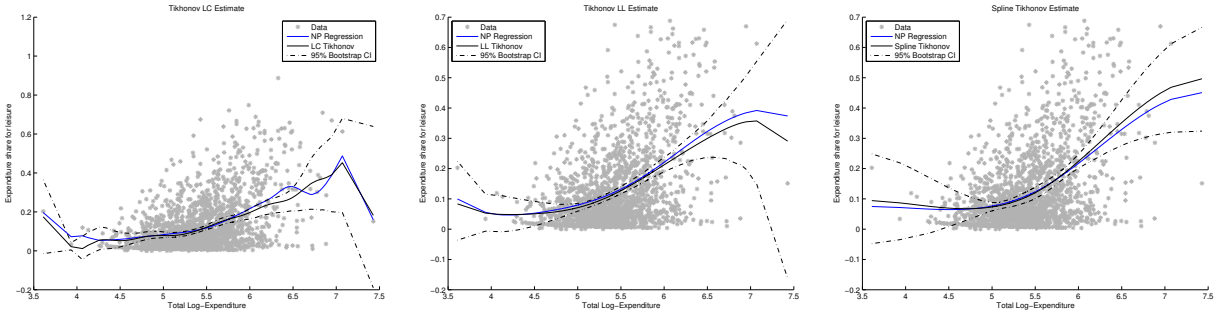
(a) Local Constant (b) Local Linear (c) Cubic B-splines

Figure 2.6: Engel Curve for food



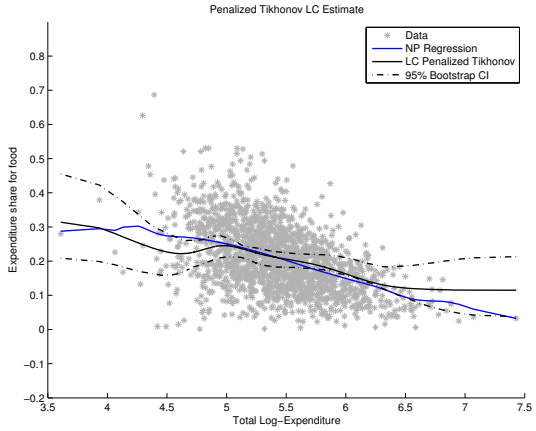
(a) Local Constant (b) Local Linear (c) Cubic B-splines

Figure 2.7: Engel Curve for fuel

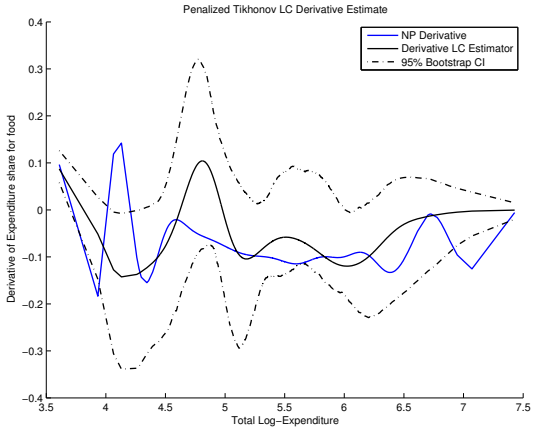


(a) Local Constant (b) Local Linear (c) Cubic B-splines

Figure 2.8: Engel Curve for leisure

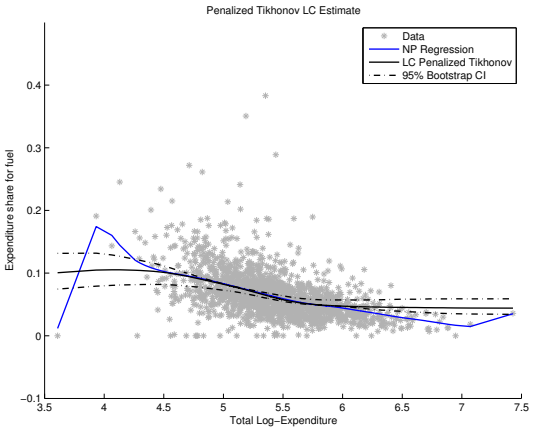


(a) Local Constant Penalized

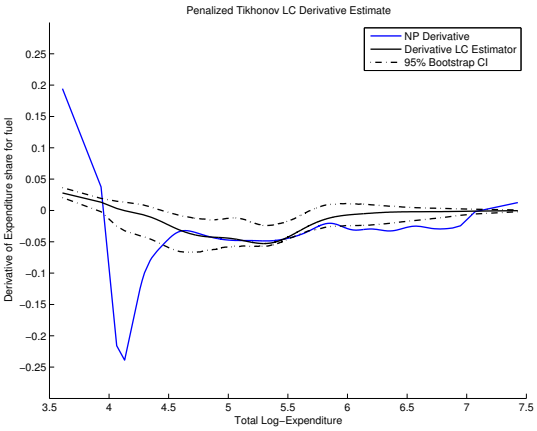


(b) First derivative

Figure 2.9: Engel Curve for food and its derivative

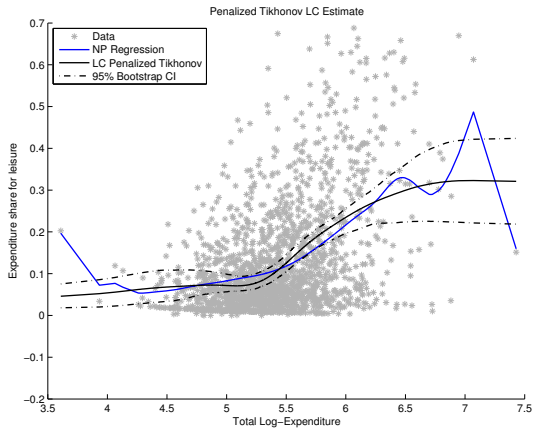


(a) Local Constant Penalized

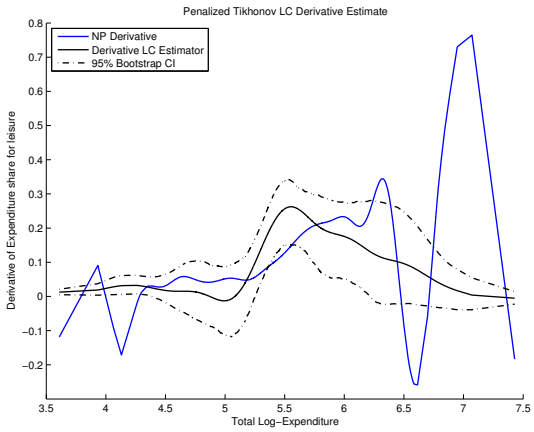


(b) First derivative

Figure 2.10: Engel Curve for fuel and its derivative



(a) Local Constant Penalized



(b) First derivative

Figure 2.11: Engel Curve for leisure and its derivative

Chapter 3

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.

3.1 Introduction

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

$$Y = \varphi(Z) + U \tag{3.1.1}$$

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, 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 paper, 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.

3.2 The Model

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

$$Y = \varphi(Z) + U \quad \text{with} \quad \mathbb{E}(U|W) = 0 \tag{3.2.1}$$

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 (3.2.1) is equivalent to:

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

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 paper, 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 (3.2.1) 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 (3.2.2) if T is one to one, or, equivalently, if:

$$T\varphi \overset{a.s.}{=} 0 \quad \Rightarrow \quad \varphi \overset{a.s.}{=} 0 \tag{3.2.3}$$

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

Let us remind that model (3.2.1) 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:

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

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 paper, 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 (4.2.5), 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 paper 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:

$$Var\left(Y|W = w\right) = Var\left[\left(\varphi(Z) + U\right)|W = w\right] = \sigma^2 \tag{3.2.4}$$

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 4 (Linear simultaneous equations)**.** Assume for simplicity that $p = q = 1$, so that $(Z, W) \in \mathbb{R}^2$, and consider model (3.2.1) 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:

$$\hat{r}\left(w\right) = G_{\varepsilon|w}^{-1}\left[\hat{\mathbb{P}}\left(\tilde{Y} = 1|W = w\right)\right] \tag{3.2.5}$$

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):

$$T\varphi = r \tag{3.2.6}$$

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).

3.3 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 (3.2.6), we need to estimate the operator T, its adjoint T^*, and r.

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 $\beta > 0$.

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

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:

$$\hat{\varphi}^\alpha = \left(\alpha I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\hat{r} \tag{3.3.1}$$

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 7. *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 paper. 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 (7),

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 8. *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.:

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

3.4 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 (3.2.5).

(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 paper 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 use the quantile function associated to the distribution G_ε, and the estimator of the conditional probability obtained in step (i) (see equation 3.2.5).

Step (iii)

We finally obtain the nonparametric instrumental regression function by solving (3.2.6), using a Tikhonov regularization method (see equation 3.3.1).

[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).

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, developed in Centorrino (2013). It consists of the minimization of the following function:

$$CV(\alpha) = \left\| \hat{T}\hat{\varphi}^{\alpha}_{(-i)} - \hat{r} \right\|^2 \tag{3.4.1}$$

where $\hat{\varphi}^{\alpha}_{(-i)}$ is the estimator of φ where the i^{th} observation has been removed. This function corresponds to the minimization of the norm of the residuals from the integral equation (3.2.6).

Using the optimal selection criterion, we obtain the first step Tikhonov estimator of the regression function as described in (3.3.1).

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.

3.5 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. $\varepsilon|w \sim 0.8\mathcal{N}(-1, 0.05) + 0.2\mathcal{N}(4, 0.15)$. 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.

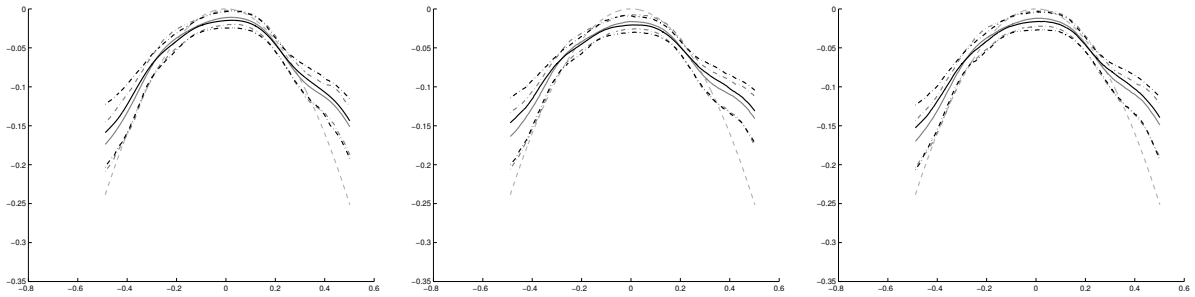
We use standard Gaussian kernels. The regularization parameters is computed as explained in section (3.4). The bandwidth parameters are obtained using leave-one-out cross validation[3].

Figures (3.1) and (3.3) report the estimation results when using a Probit specification of the model. Figures (3.2) and (3.4) 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.

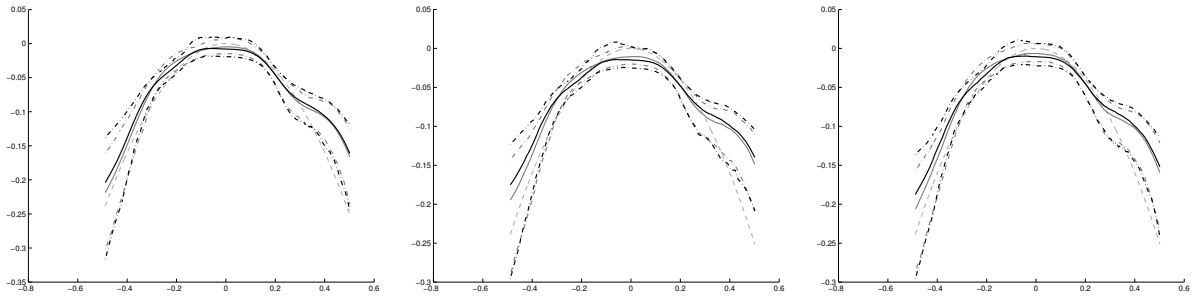
Furthermore, the regularity of the function of interest does change the quality of our results. As a matter of fact, our estimator performs much better in the case where we take a very regular function (z^2) compared to the case where the function is highly irregular ($e^{-|z|}$). This is particularly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

[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 3.1: 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

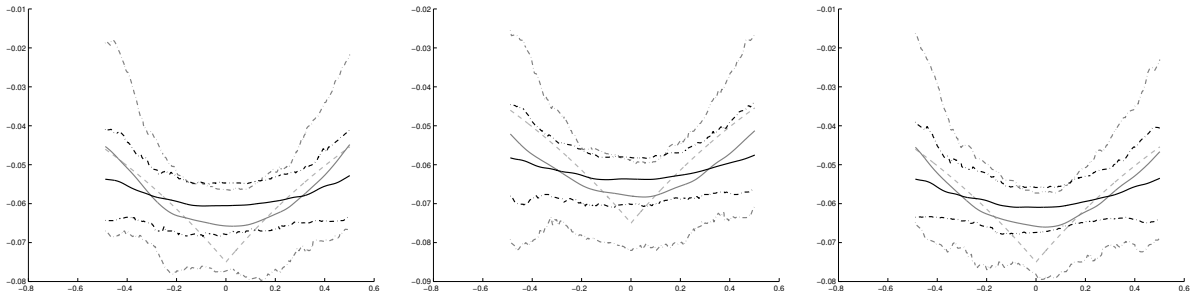
Figure 3.2: Estimation of the regression function $\varphi(z) = -z^2$ using a Logit specification. The true function (dashed light grey line) is plotted against the median of the first step (dark grey line) and the second step (black line) Tikhonov estimators, and their simulated confidence intervals.

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

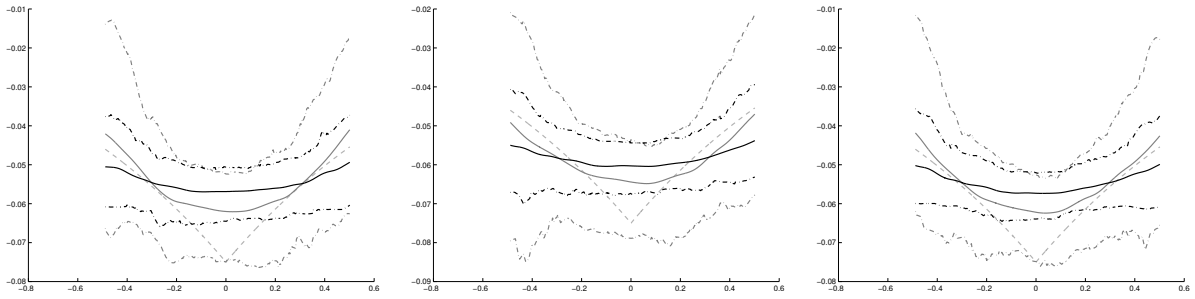
The choice to move to another US state may be related to higher expected income in the new state of residence. However, income is expected to increase, if and only if the individual decides to move. This makes income a potentially endogenous dependent variable.

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



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

Figure 3.3: 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 3.4: 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).

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 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 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 3.1: 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:

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

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.

[5]Some descriptive statistics for these variables are given in Table (4.6).

[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.

(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 (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 3.5), 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:

$$\tilde{Y} = \mathbb{1}\left(Z\gamma + X_1 + X_2\beta_2 + U \geq 0\right) \tag{3.6.1}$$

$$Z = \zeta(W) + V \tag{3.6.2}$$

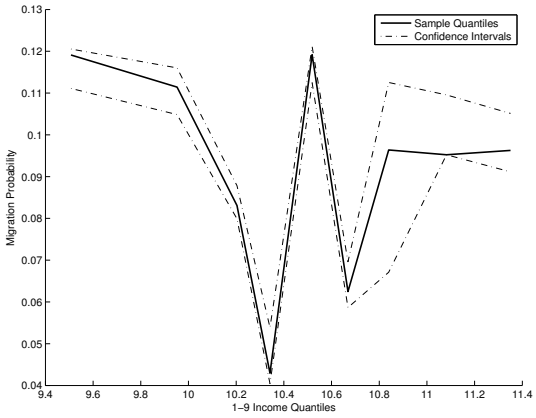


Figure 3.5: Average probability of migration by income quantile.

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 (3.6.2). 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]:

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

where $\left\{\hat{\mathbb{P}}\left(U|z_i\gamma + x_{2i}\beta_2,\hat{v}_i\right), i = 1,...,N\right\}$ 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.

Table (3.2) 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

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

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

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 (3.6) draws the nonparametric instrumental variable estimator of the impact of income on migration probabilities. Bootstrap confidence intervals are obtained using the method developed in Centorrino et al. (2013a). 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.

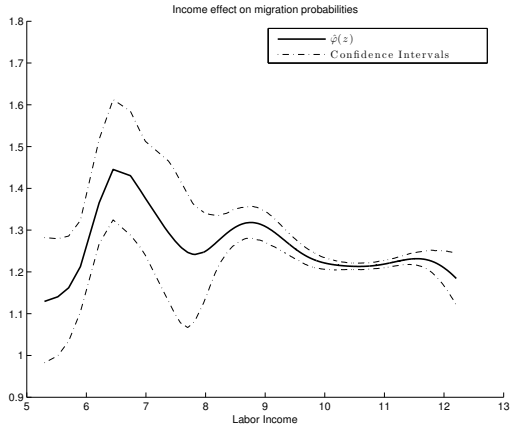


Figure 3.6: Functional estimator of the impact of income on migration decisions.

3.7 Conclusions

We propose in this paper 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.

3.8 Appendix

3.8.1 Proof of Assumption 8

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 7. Then:

$$\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\varphi = \hat{T}^*\hat{r}^* + \hat{T}^*Q_\varepsilon'(G_\varepsilon(\hat{r}^*))(\hat{p}(w) - G_\varepsilon(\hat{r}^*)) \tag{3.8.1}$$

where, for simplicity, we omit higher order terms.

We consider the Hilbert-Schmidt norm of the term on the lhs of 3.8.1:

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

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 4

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.

4.1 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 nonparamet-

ric 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 Centorrino (2013); 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 variables. 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); 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 paper is structured as follows. In section (4.2), we present the main framework. We review carefully each regularization scheme, and we discuss its practical implementation in section (4.3). In sections (4.4) and (4.5), we describe the structure of the Monte-Carlo experiment, and expose the bootstrap procedure and its validity. In section (4.6), we present an application to the estimation of the Engel curve for food using a cross section database of Pakistan households. Finally, section (4.7) concludes.

4.2 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:

$$Y = \varphi(Z) + U \tag{4.2.1a}$$

$$\mathbb{E}(U|W) = 0 \tag{4.2.1b}$$

This model is a regression type model, where the usual mean independence condition $\mathbb{E}(U|Z) = 0$ is replaced by condition (4.2.1b). 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 (4.2.1a) (see,e.g. Newey and Powell, 2003; Hall and Horowitz, 2005; Carrasco et al., 2007; Darolles et al.,

2011*a*; Chen and Pouzo, 2012, 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.

$$\mathbb{E}(\varphi(Z)|W) = \mathbb{E}(Y|W) \tag{4.2.2}$$

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 (4.2.2) now writes:

$$T\varphi = r \tag{4.2.3}$$

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 (4.2.3), 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 (4.2.3) 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:

$$T\varphi \overset{a.s.}{=} 0 \quad \Rightarrow \quad \varphi \overset{a.s.}{=} 0 \tag{4.2.5}$$

(see Newey and Powell, 2003; Darolles et al., 2011*a*; Andrews, 2011; D'Haultfoeuille, 2011). This *completeness condition* is assumed to hold throughout the paper, 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:

$$\varphi = \sum_{j=1}^{\infty} \frac{r_j}{\lambda_j} \psi_j \tag{4.2.6}$$

The ill-posedness in (4.2.3) 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 need 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 (4.2.3) 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 (4.2.6) 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 (4.2.6) 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.

4.3 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 parameters 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 dispose of 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 Li and Racine, 2007). Hence, here we

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

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 dispose of 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 (4.2.3), 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 paper, and discusses, for each of them, a criterion for the data-driven choice of the regularization parameter.

4.3.1 Tikhonov Regularization

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

$$\|T\varphi - r\|^2 + \alpha\|\varphi\|^2 \tag{4.3.1}$$

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

$$\alpha\varphi + T^*T\varphi = T^*r \tag{4.3.2}$$

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 (4.3.2), i.e.,

$$\hat{\varphi}^{\alpha} = \left(\alpha I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^* r \tag{4.3.3}$$

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

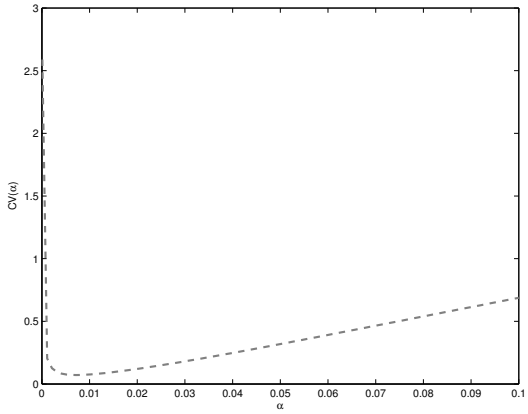


Figure 4.1: Criterion function for the optimal choice of α in Tikhonov regularization

In order to choose the regularization parameter α, we adopt the cross validation approach developed in Centorrino (2013). This method consists of minimizing the following sum of squares:

$$CV(\alpha) = \left\|\hat{T}\hat{\varphi}^{\alpha}_{(-i)} - \hat{r}\right\|^2$$

where $\hat{\varphi}^{\alpha}_{(-i)}$ is the estimator of φ obtained by removing the i^{th} observation from the sample. Centorrino (2013) proves that this criterion is rate optimal in mean squared error and shows its superior finite sample performances compared to other existing numerical methods. A typical shape of this criterion function can be found in figure (4.3.1).

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 (4.3.2). 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 paper, we only consider the performance of the first step TK estimator.

4.3.2 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 (4.2.3) 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:

$$\varphi_{j+1} = \varphi_j + cT^*\left(r - T\varphi_j\right), \quad \forall j = 0, 1, \ldots \tag{4.3.4}$$

or equivalently:

$$\varphi^M = c\sum_{j=0}^{M-1}\left(I - cT^*T\right)^j T^*r \tag{4.3.5}$$

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 (4.3.5) 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 (4.3.4). 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 (4.3.5) 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 (4.3.4), 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 (4.2). 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.

[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.

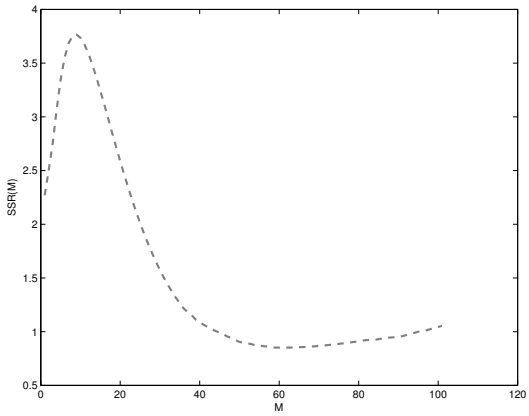


Figure 4.2: Stopping function for Landweber-Fridman regularization

4.3.3 Galerkin Regularization

The Galerkin type of regularization (GK henceforth) consists on truncating the infinite sum in (4.2.6), 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:

$$\varphi^{J_n} = \sum_{j=1}^{J_n} \beta_j \phi_j \tag{4.3.6}$$

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 (4.3.6), 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:

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

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:

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

Blundell et al. (2007) call ρ_{J_n} the sieve measure of ill-posedness. As $n \to \infty$, to obtain consistency of the estimator, 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 (4.3).

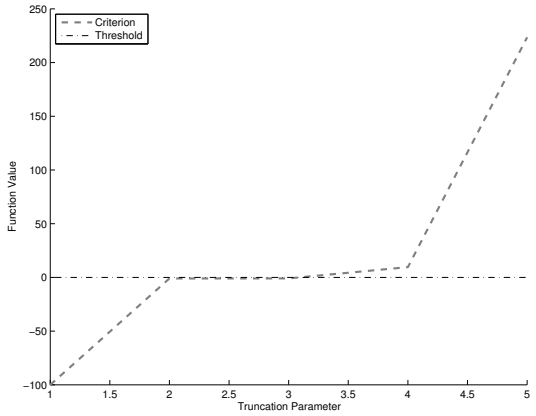


Figure 4.3: Choice of $\hat{J}_n$ for Galerkin regularization.

A final remark on GK regularization is about the variance of the estimator in finite samples. The GK estimation procedure is a nonparametric generalization of the 2SLS estimator. Mariano (1972), in an influential paper, shows that the 2SLS estimator only possesses moments of order $p - q + 1$, where p is the dimension of the endogenous variable and q the dimension of the instruments. Therefore, if one uses the same dimension for the matrices $\mathcal{W}_n$ and $\mathcal{Z}_n$, our GK would 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]

4.3.4 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 (4.2.3) to obtain φ as the integral of its derivatives of any order. Therefore, we can regularize the derivative

[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.

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. $\varphi' \in \mathbb{L}_z^2$. Define the first order differential operator L. We can rewrite equation (4.2.3) 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 φ' 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 a generic function ψ, 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:

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

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 $\bar{K}_h(a) = 1 - \int_{-\infty}^{a} K_h(b)db$. We then have:

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

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 (4.2.1a), 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:

$$\hat{B}^*\hat{B}\varphi' = \hat{B}^* r \tag{4.3.7}$$

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 (4.3.7) 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 Centorrino, 2013; Fève and Florens, 2013, for an application).

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

$$\varphi'_{j+1} = \varphi'_j + cT^*\left(r - T\varphi'_j\right), \quad \forall j = 0, 1, \ldots \tag{4.3.8}$$

where:

$$\varphi_j = L^{-1}\varphi'_j - \mathbb{E}\left(L^{-1}\varphi'_j\right)$$

with the initial condition:

$$\varphi'_0 = c\frac{1}{\hat{f}_Z}\left[\hat{S}r - \hat{S}_Z\hat{\mathbb{E}}_N(r)\right]$$

Finally:

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

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.

4.4 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 techniques presented in this paper and test the validity of the bootstrap, discussed in the next section. By no means, we would try to drive the empirical researcher towards one of these methods. On the contrary, 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 endogenous 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 paper 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 (4.2.1*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 paper 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 $(0, 1)$.

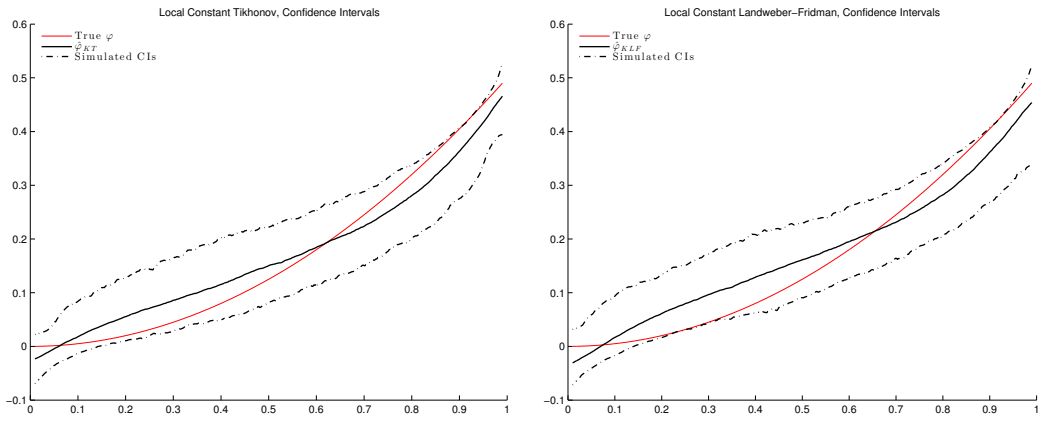
When using B-splines, we fix the order of the basis to 4 (cubic splines), and we compute the optimal number of knots using either least squares cross validation (TK and LF) or the method developed in Horowitz (2012) (GK). An important remark about the B-spline estimation is about the choice of knots. 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, this also seems appropriate in the case of nonparametric instrumental regressions (see Centorrino, 2013; Fève and Florens, 2013, for a further discussion on the topic).

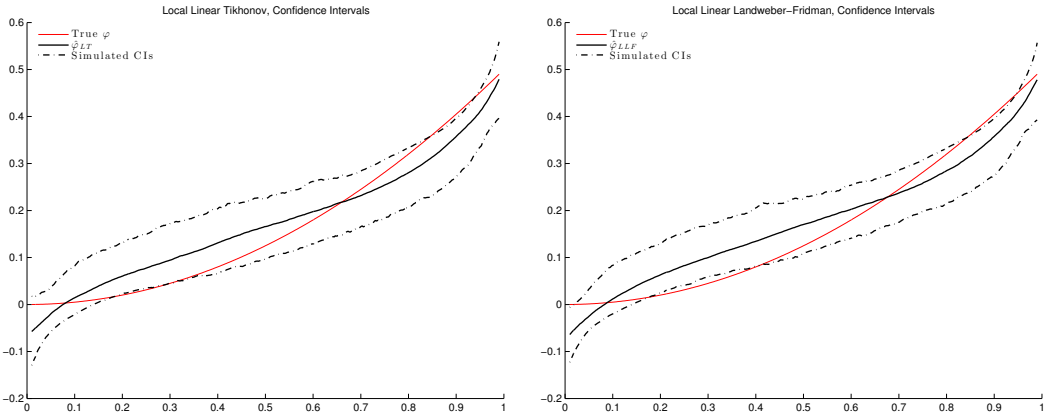
Figures (4.4), (4.5), (4.6) and (4.7) report the results of our simulations for the local constant, local linear, B-splines and penalized first derivative local constant estimators. On the left panel of each figure, we draw the TK regularized solution; the LF solution is instead on the right panel. Figure (4.8) presents the same results for GK with B-splines. The light gray line in each figure is the true function φ. The thick black line is the median value of the regression function at each evaluation point from the simulation and the dashed lines give the 95% confidence intervals.

[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).



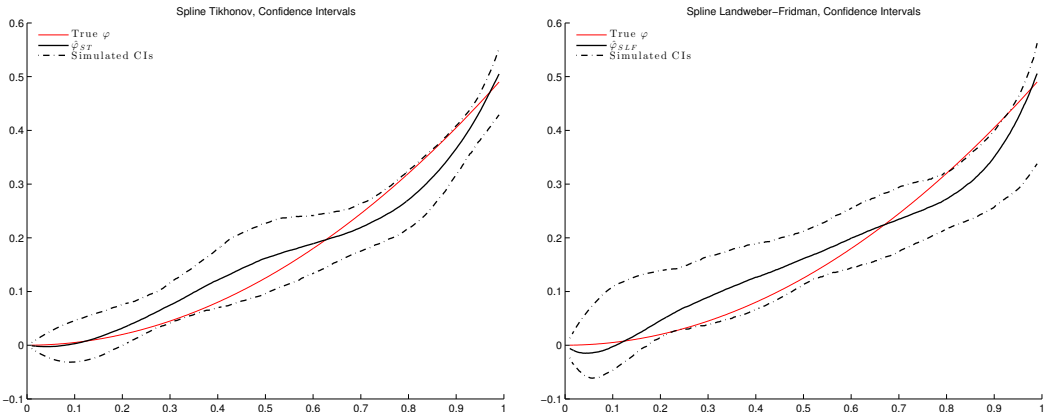
(a) Tikhonov (b) Landweber-Fridman

Figure 4.4: Simulations results using Local Constant Kernels



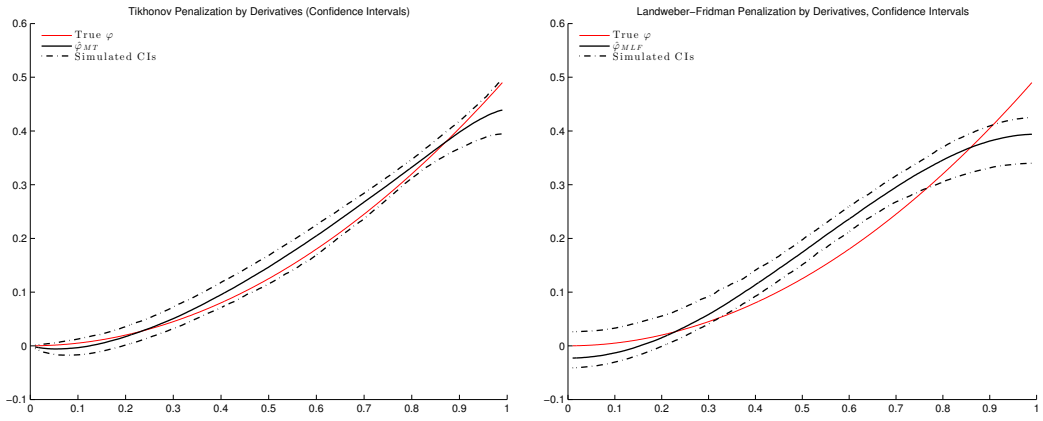
(a) Tikhonov (b) Landweber-Fridman

Figure 4.5: Simulations results using Local Linear Kernels



(a) Tikhonov (b) Landweber-Fridman

Figure 4.6: Simulations results using B-Splines



(a) Tikhonov (b) Landweber-Fridman

Figure 4.7: Simulations results using Local Constant Kernel with penalized first derivative

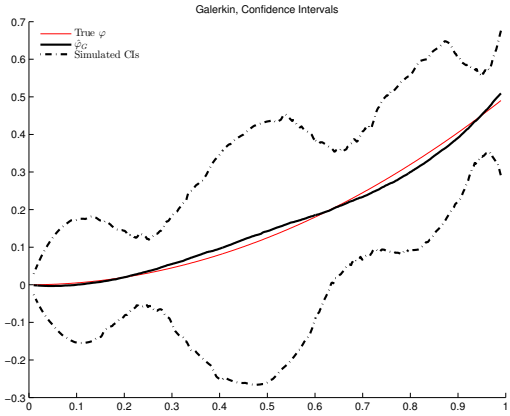


Figure 4.8: Simulations results using Galerkin with B-splines

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 (4.1). All estimators have roughly comparable performances. A comparison of the MISE shows that the Penalized Local Constant TK and the B-spline 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 bigger compared to the one of other estimators.

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

	MISE	MSE	Bias	Variance
Local Constant TH	0.00214	0.00219	0.01079	0.00119
Local Linear TH	0.00253	0.00260	0.01703	0.00104
Spline TH	0.00148	0.00129	0.00329	0.00057
Penalized TH	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 4.1: MISE and Median MSE, Bias and Variance for each estimator.

In order to explore further the differences between bias and variance for the different estimators, we report, in Table (4.2), 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, 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 (4.1).

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 (4.2), 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 knots and, in some particular cases, 4 knots. However, this is sufficient to have a huge effect on the total variance of the GK estimator.

	Mean	Median	St.Dev	Min	Max
Local Constant TH ($\times 10^5$)	1288.3	1082.2	724.5	122.5	6043.2
Local Linear TH ($\times 10^5$)	716.0	391.8	959.4	3.9	11251.9
Spline TH ($\times 10^5$)	872.7	866.3	439.9	0.0	2794.7
Penalized TH ($\times 10^5$)	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 4.2: Summary statistics for the regularization parameter.

Finally, we also report in Table (4.3) 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 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]

[8] An additional comment about LF is that, although updating the regularization parameter at each iteration may

	Mean	Median	St.Dev	Min	Max
Local Constant TH	16.64	15.20	5.00	10.27	43.35
Local Linear TH	35.96	31.16	18.11	14.20	227.92
Spline TH	16.49	16.57	2.08	4.03	24.72
Penalized TH	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 4.3: CPU time for each estimator (in seconds).

4.5 Wild Bootstrap in Nonparametric IV

4.5.1 Resampling from sample residuals in Nonparametric Regression Models

In standard nonparametric regressions without endogeneity, the general theory of bootstrap in is presented in Härdle and Bowman (1988) and Härdle and Marron (1991). To present briefly their approach, 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:

$$y^* = \hat{m}(z) + u^*$$

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.

For simplicity, we refer to this technique in the following as *naïf 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 $b = \hat{u}_i(1+\sqrt{5})/2$.

4.5.2 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), 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 paper. 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 paper 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:

$$Y = \varphi(Z) + U \tag{4.5.1}$$

$$Z = g(W, V) \tag{4.5.2}$$

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 (4.5.1) 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:

$$\eta = r - T\varphi \tag{4.5.3}$$

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 (4.5.3) is a functional residual. To consistently bootstrap from it, we can write its Fourier decomposition as follows:

$$\eta = \sum_{j=0}^{\infty} \frac{\langle \eta, \phi_j \rangle}{\lambda_j} \lambda_j \phi_j$$

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.

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.:

$$\varepsilon = Y - \mathbb{E}(\varphi(Z)|W) \tag{4.5.4}$$

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:

$$T_N : \mathbb{L}_Z^2 \to \mathbb{R}^N \tag{4.5.5}$$

$$T_N^* : \mathbb{R}^N \to \mathbb{L}_Z^2 \tag{4.5.6}$$

and the inverse problem would be the one defined by the sample counterpart of equation (4.5.4). Notice that this approach is much simpler than the direct bootstrap from equation (4.5.3). A potential criticism is that, resampling from (4.5.4), leads to bootstrap only the dependent variable Y and not the endogenous component Z. However, by the definition of the error term ε in (4.5.4), 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 (4.5.1), at least in some particular cases.

Example 5 (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 and standard bootstrap techniques can be applied. However, the estimated residual from (4.5.4) is, by the 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]

Call $\hat{T}$ the estimated conditional expectation operator, acting onto the space spanned by W. The estimated residuals are defined as follows:

$$\hat{\varepsilon}_i(w) = y_i - \hat{T}\hat{\varphi}(z_i) \quad \forall i = 1, \dots, N$$

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 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. 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.

(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

[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 paper. Interested readers are referred to Kauermann and Carroll (2001) and Kauermann et al. (2009).

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:

$$\hat{\varphi}^{\alpha} - \varphi^{\alpha} = \hat{\varphi}^{\alpha} - \varphi + \alpha \left(\alpha I + T^{*}T\right)^{-1} T^{*}T\varphi \tag{4.5.7}$$

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 (4.5.7) by:

$$\hat{\varphi}^{*\alpha} - \hat{\varphi}^{\alpha} + \alpha_N \left(\alpha_N I + \hat{T}^{*}\hat{T}\right)^{-1} \hat{T}^{*}\hat{T}\hat{\varphi}^{\alpha} \tag{4.5.8}$$

(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:

$$\hat{\varphi}^{M} - \varphi^{M} = \hat{\varphi}^{M} - \varphi + c \sum_{k=0}^{M-1} \left(I - cT^{*}T\right)^{k} T^{*}T\varphi \tag{4.5.9}$$

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.

We therefore compare the distribution $f(\varphi)$ of $\hat{\varphi} - \varphi$ with the distribution $f^*(\varphi)$ of $\hat{\varphi^*} - \hat{\varphi}$, at each point of the vector Q. For each bootstrap density we compute the absolute deviation 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 $\hat{f}^*(\varphi)$.

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 (4.9), (4.10), (4.11), (4.12), (4.13), (4.14), (4.15), (4.16) and (4.17) present the comparison

[10] See also Ferraty et al. (2010), for a similar approach to the validity of bootstrap.

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.

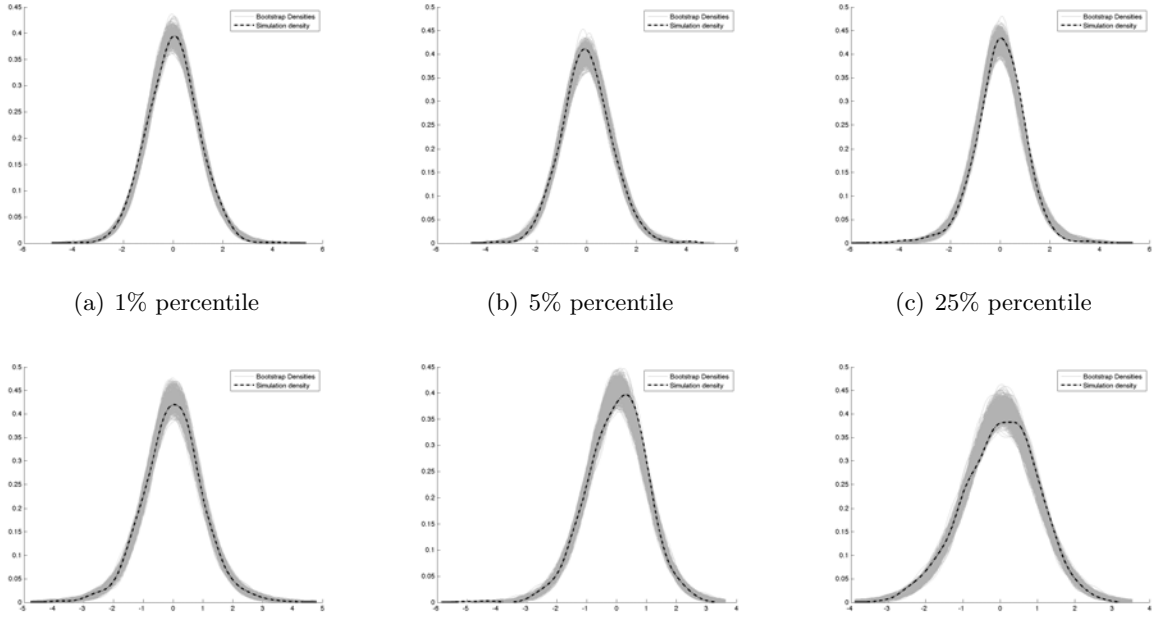
Finally, Table (4.4) reports the median value of the variational distance for each value of the vector Q.[11] The median variational distance is below 0.1 for the majority of the estimators and it therefore confirms that the bootstrap density approximates the true density fairly well. However, its performance 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).

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
Local Constant TH	0.0261	0.0253	0.0451	0.0279	0.0270	0.0381	0.0418
Local Linear TH	0.0710	0.0799	0.0783	0.1979	0.1073	0.0355	0.0664
Spline TH	0.0205	0.0183	0.0730	0.0520	0.0677	0.0541	0.0431
Penalized TH	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 4.4: Median Variational Distance at each point of the vector Q.

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 (4.5) 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 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.

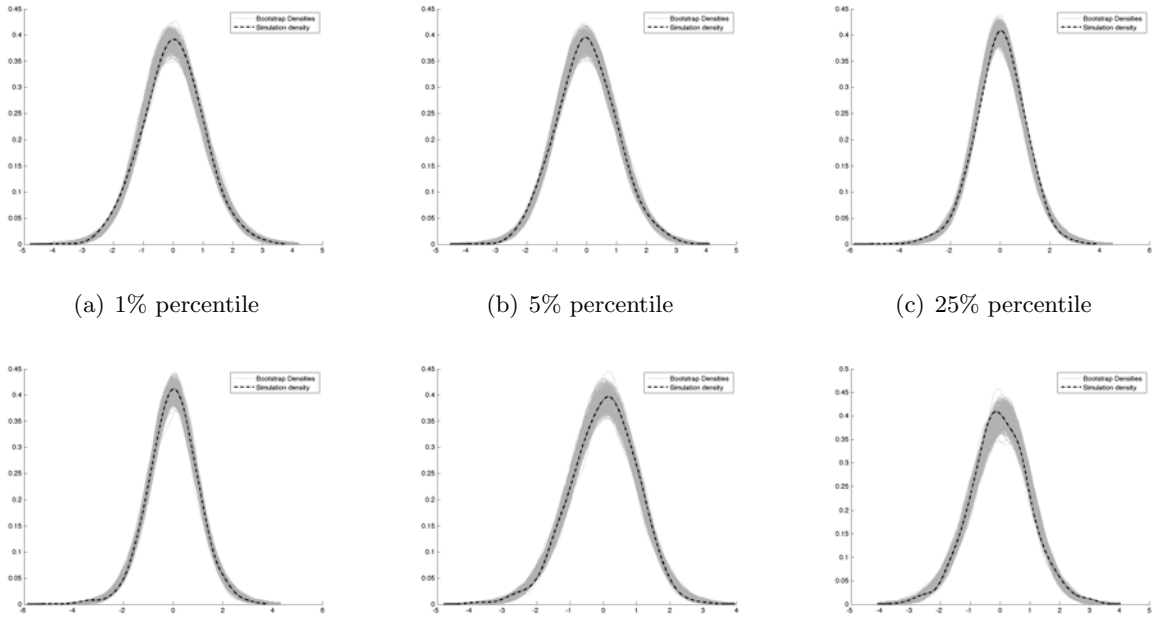
[11] Figures (4.23), (4.24), (4.25), (4.26) and (4.27) in the Appendix report also a box plot comparison of Total Variational Distance.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

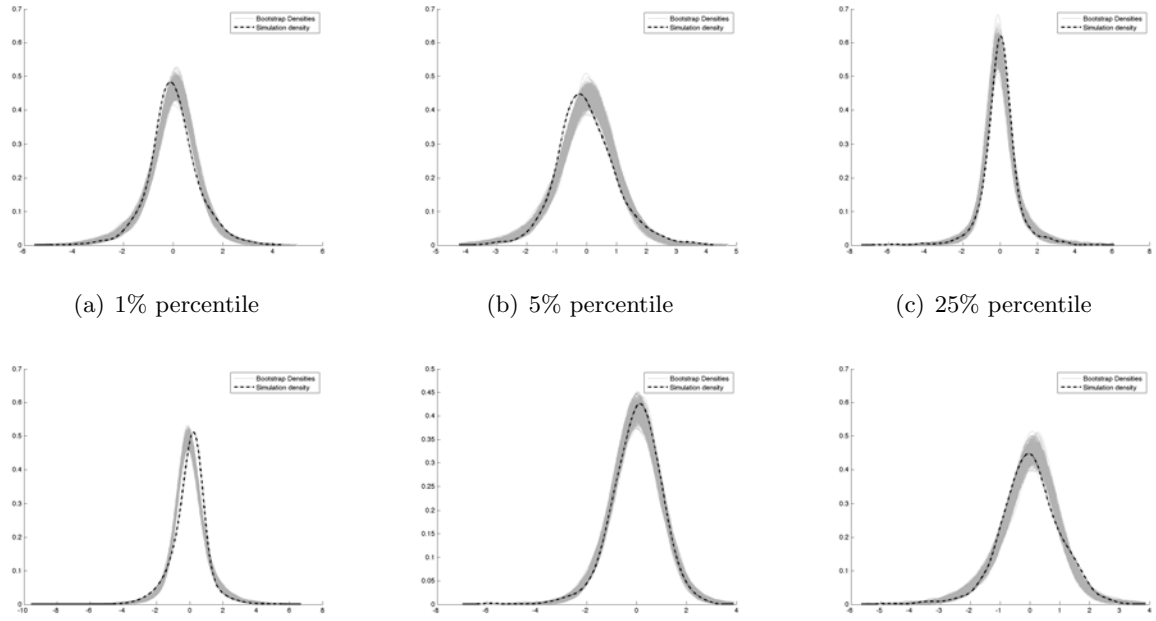
Figure 4.9: Simulation vs Bootstrap Densities for Local Constant Tikhonov.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

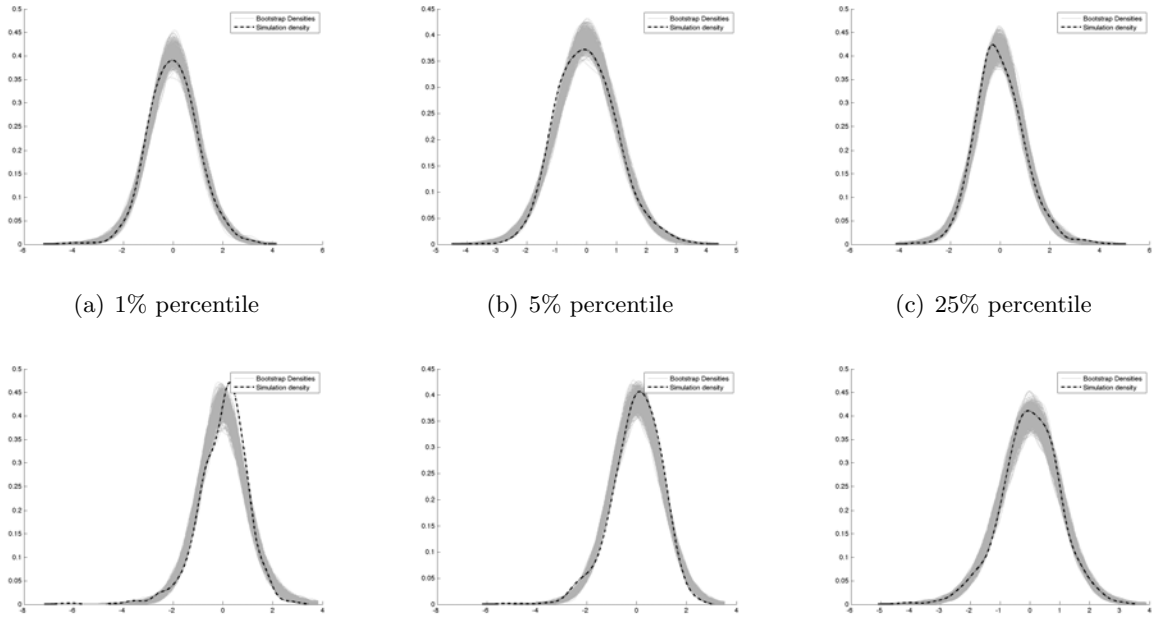
Figure 4.10: Simulation vs Bootstrap Densities for Local Constant Landweber-Fridman.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

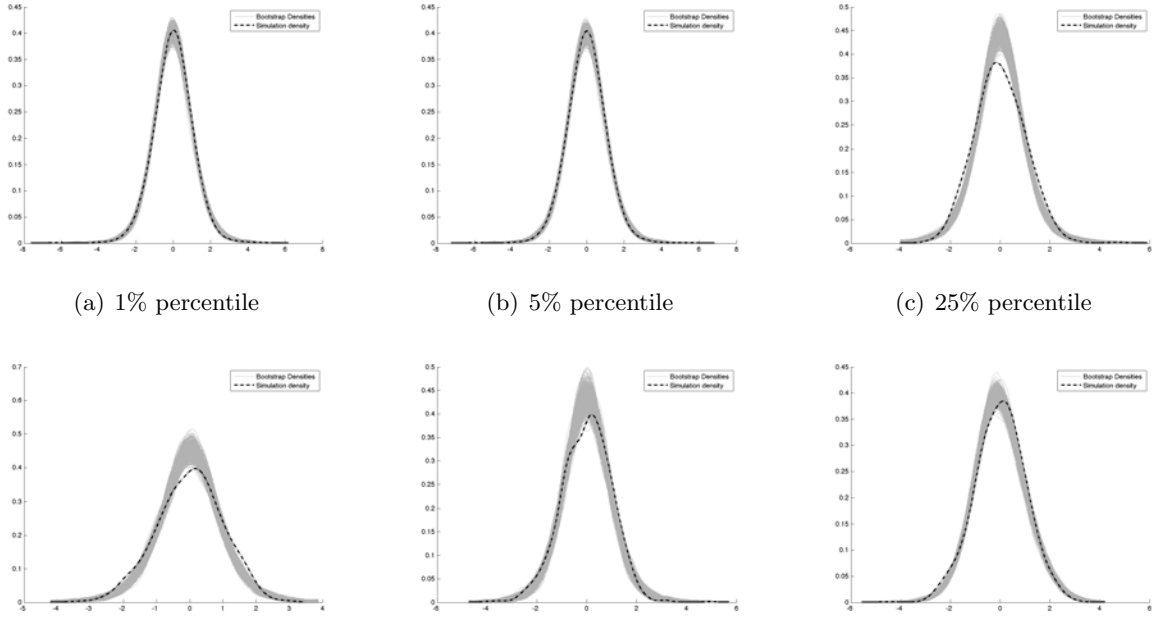
Figure 4.11: Simulation vs Bootstrap Densities for Local Linear Tikhonov.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

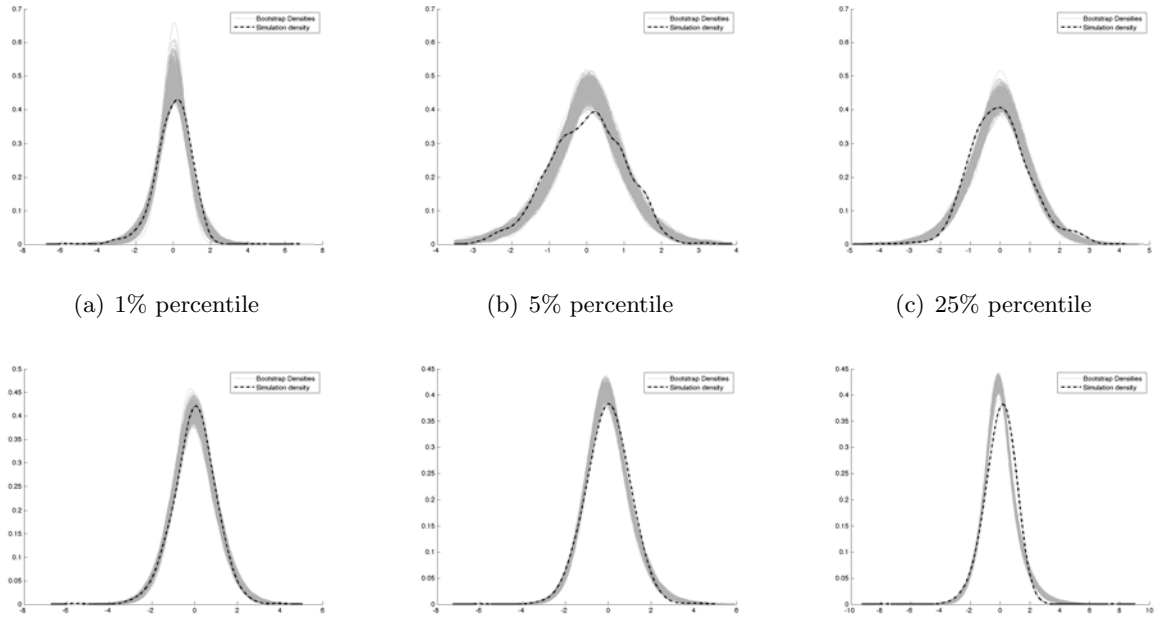
Figure 4.12: Simulation vs Bootstrap Densities for Local Linear Landweber-Fridman.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

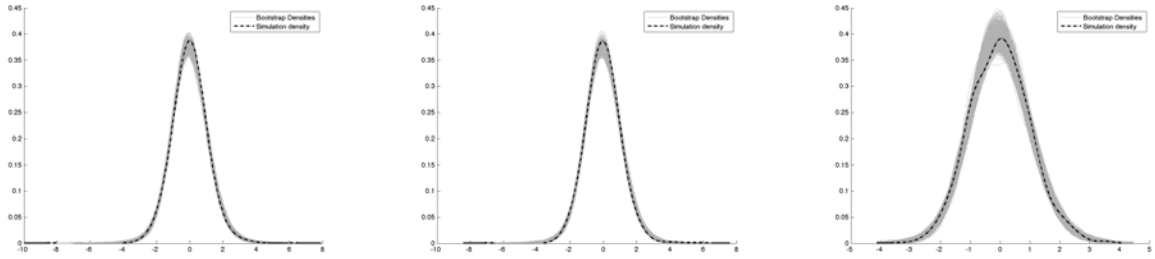
Figure 4.13: Simulation vs Bootstrap Densities for Spline Tikhonov.



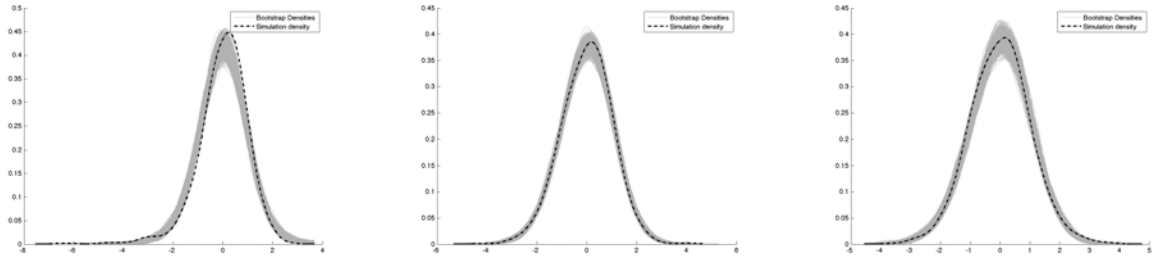
(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

Figure 4.14: Simulation vs Bootstrap Densities for Spline Landweber-Fridman.

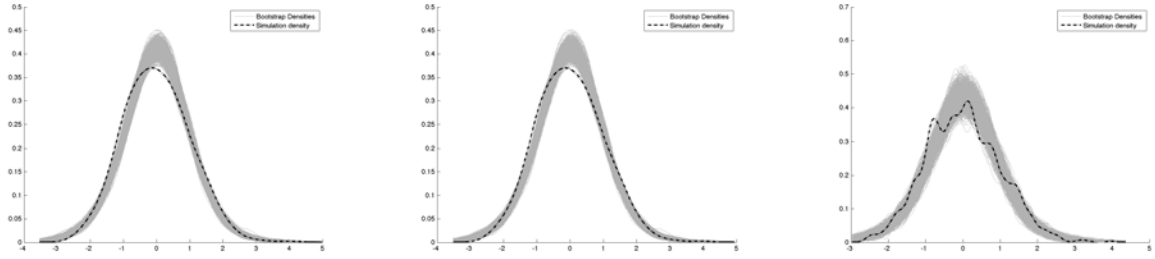


(a) 1% percentile (b) 5% percentile (c) 25% percentile

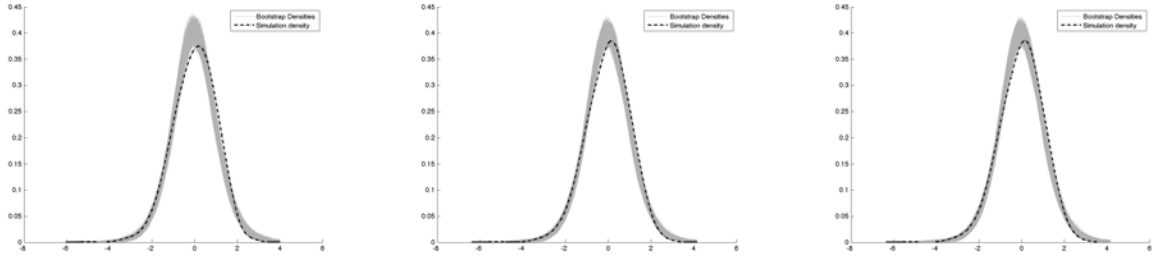


(d) 75% percentile (e) 95% percentile (f) 99% percentile

Figure 4.15: Simulation vs Bootstrap Densities for Local Constant Tikhonov with Penalized first derivative.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

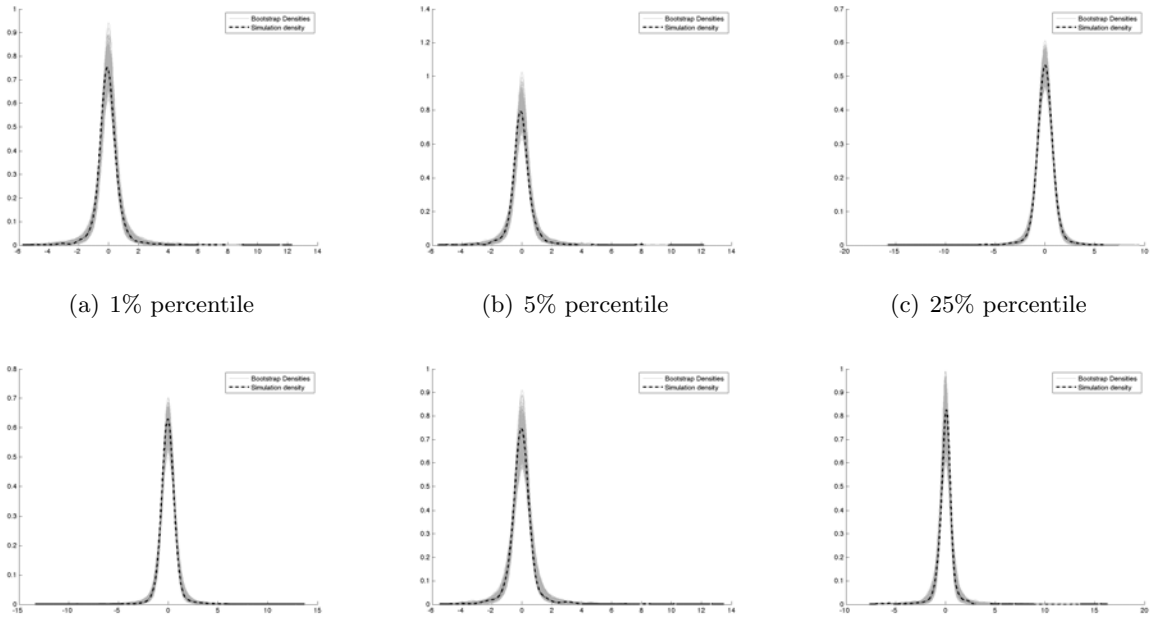


(d) 75% percentile (e) 95% percentile (f) 99% percentile

Figure 4.16: Simulation vs Bootstrap Densities for Local Constant Landweber-Fridman with Penalized first derivative.

		Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
	Local Constant TH	0.8940	0.9030	0.9090	0.8970	0.9050	0.9080	0.8980
	Local Linear TH	0.8920	0.8980	0.9020	0.8810	0.8930	0.9100	0.8970
	Spline TH	0.9000	0.9040	0.8730	0.9110	0.8730	0.8950	0.8920
	Penalized TH	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 TH	0.9560	0.9530	0.9560	0.9470	0.9500	0.9490	0.9500
	Local Linear TH	0.9450	0.9450	0.9440	0.9440	0.9470	0.9550	0.9500
	Spline TH	0.9560	0.9560	0.9610	0.9520	0.9550	0.9540	0.9430
	Penalized TH	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 TH	0.9940	0.9900	0.9840	0.9910	0.9910	0.9930	0.9960
	Local Linear TH	0.9870	0.9870	0.9870	0.9910	0.9910	0.9890	0.9900
	Spline TH	0.9890	0.9890	0.9970	0.9870	0.9970	0.9930	0.9950
	Penalized TH	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 4.5: Pointwise coverage probabilies of wild bootstrap.



(a) 1% percentile (b) 5% percentile (c) 25% percentile

(d) 75% percentile (e) 95% percentile (f) 99% percentile

Figure 4.17: Simulation vs Bootstrap Densities for Splines Galerkin.

4.6 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 regressor. 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 4.6).

	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 4.6: Summary statistics

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

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:

$$Y = \beta_1 Z + \beta_2 Z^2 + \gamma V + U \tag{4.6.1}$$

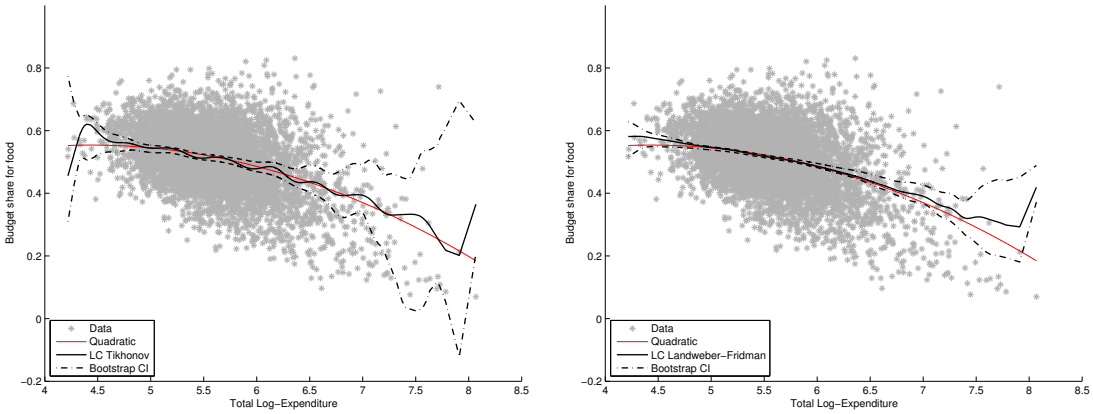
$$Z = \zeta(W) + V \tag{4.6.2}$$

$$\mathbb{E}(U|W,V) = \mathbb{E}(U|V) \tag{4.6.3}$$

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 (4.7). We can see that all coefficients are significant. The one associated with the quadratic component is very small but significantly negative.

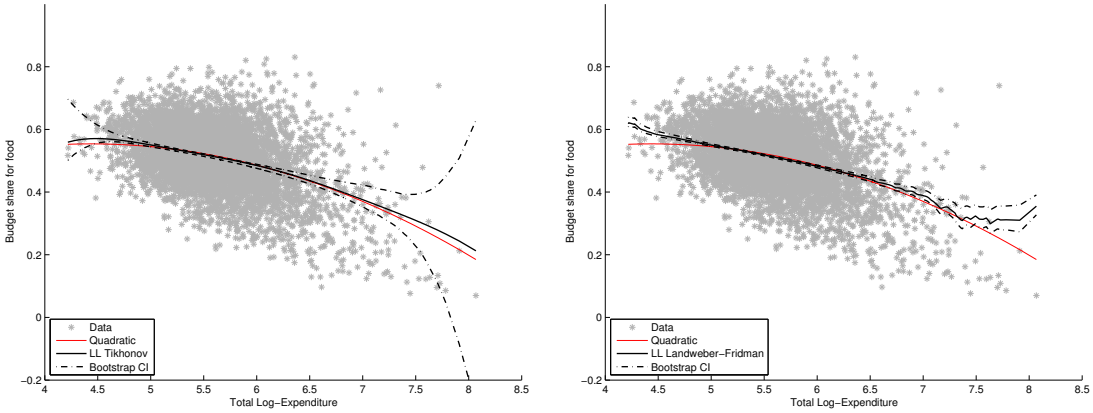
The results of the estimation of the Engel curve for Pakistan data are reported in Figures (4.18), (4.19), (4.20), (4.21) and (4.22). For each kind of nonparametric estimator (local constant, local linear, B-splines and penalized local constant), we present the outcome both using TK and LF regularizations. The final figure (4.22) 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 (4.6.1).

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 the estimated regression function. It is also instructive to observe that these bumps disappear in figure (4.21), 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



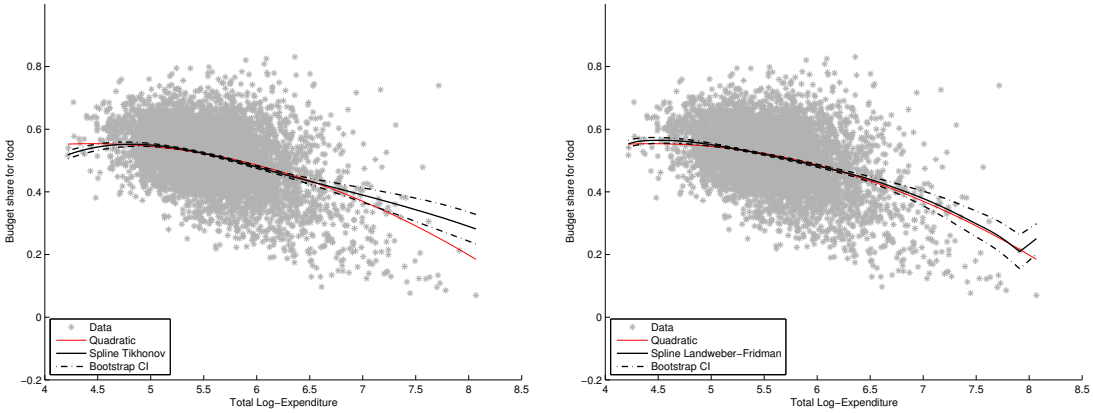
(a) Tikhonov (b) Landweber-Fridman

Figure 4.18: Estimation of the Engel Curve for food (local constant)



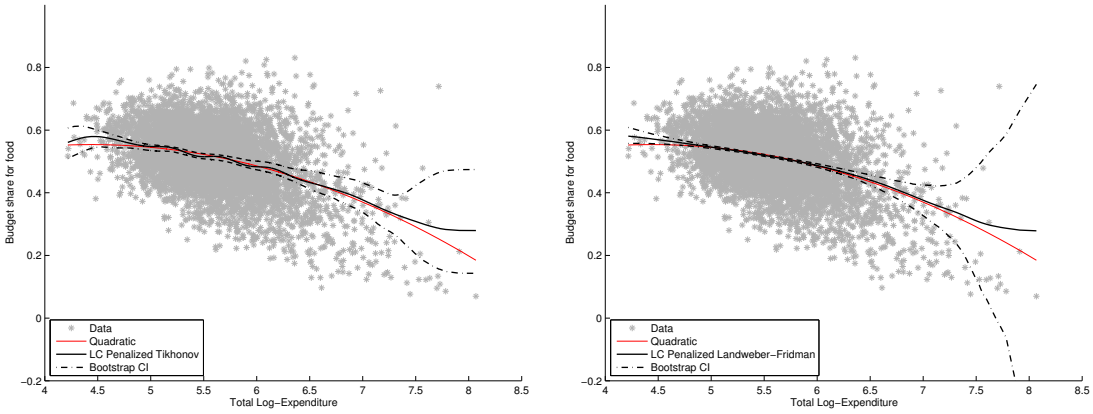
(a) Tikhonov (b) Landweber-Fridman

Figure 4.19: Estimation of the Engel Curve for food (local linear)



(a) Tikhonov (b) Landweber-Fridman

Figure 4.20: Estimation of the Engel Curve for food (splines)



(a) Tikhonov (b) Landweber-Fridman

Figure 4.21: Estimation of the Engel Curve for food (Penalized local constant)

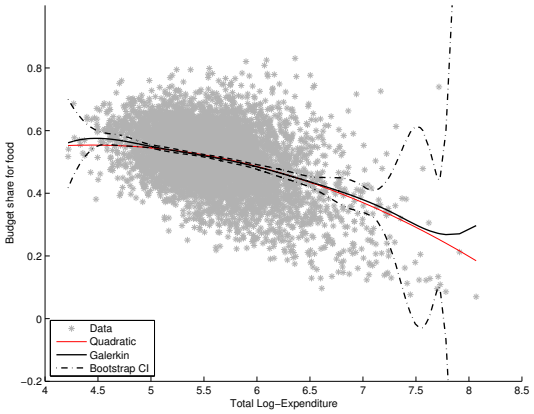


Figure 4.22: Galerkin estimation of the Engel Curve for food

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.

	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 4.7: Results from model (4.6.1). Dependent variable: share of budget for food.

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

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.

4.7 Conclusions

This paper 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.

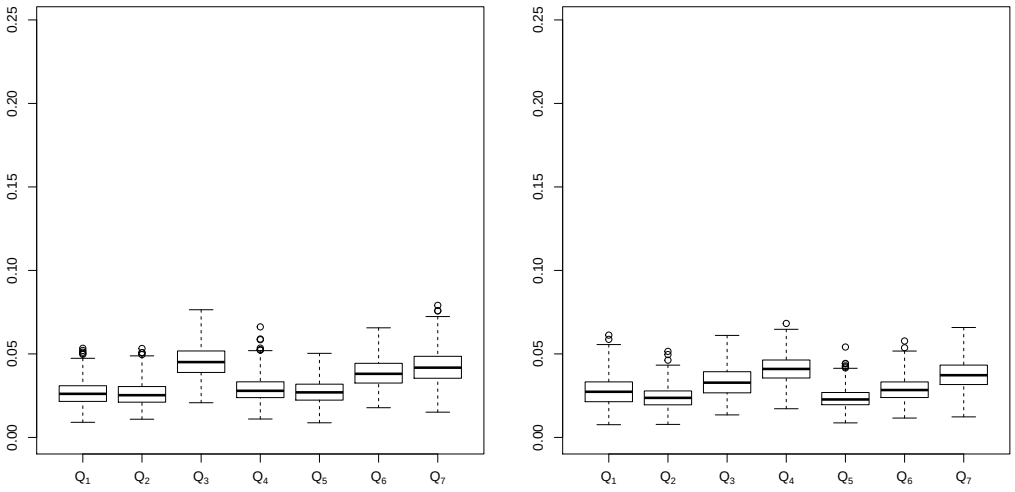
Our intention is to give a unified and simple presentation of the several regularization procedures that can be considered when applied researchers would like to keep the flexibility of nonparametric estimation in presence of endogenous regressors. Our aim is to narrow the gap between the theoretical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

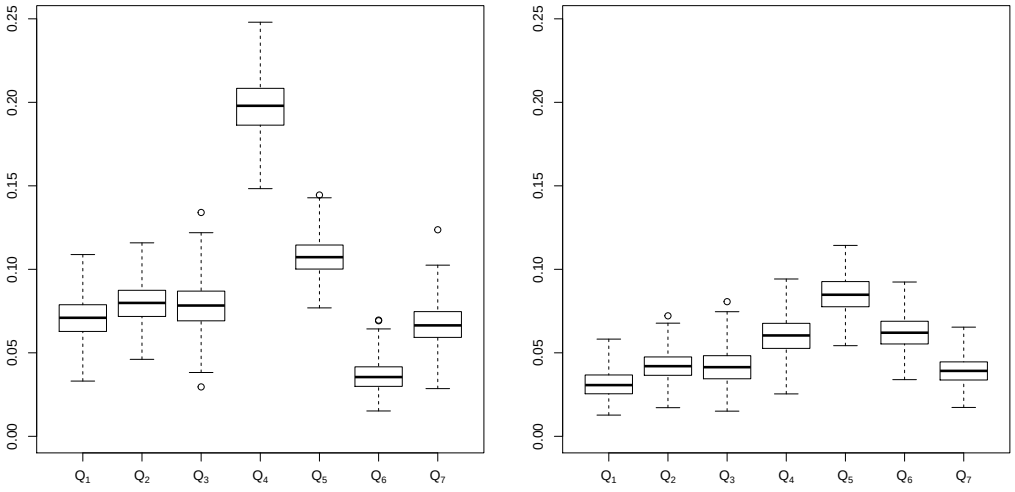
Our analysis deems partial, as we do not explore the properties of the various estimators under several simulation schemes and several degrees of *ill-posedness* of the inverse problem. However, we see this work as a useful first step to make nonparametric instrumental regression readily available to applied economists and econometricians.

4.8 Appendix



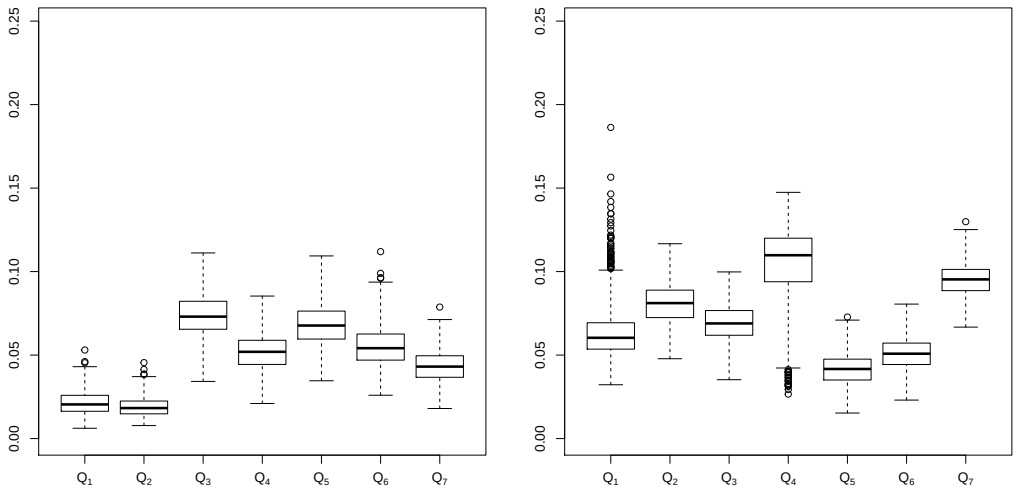
(a) Tikhonov (b) Landweber-Fridman

Figure 4.23: Box plot Total Variational Distance, Local Constant Kernels.



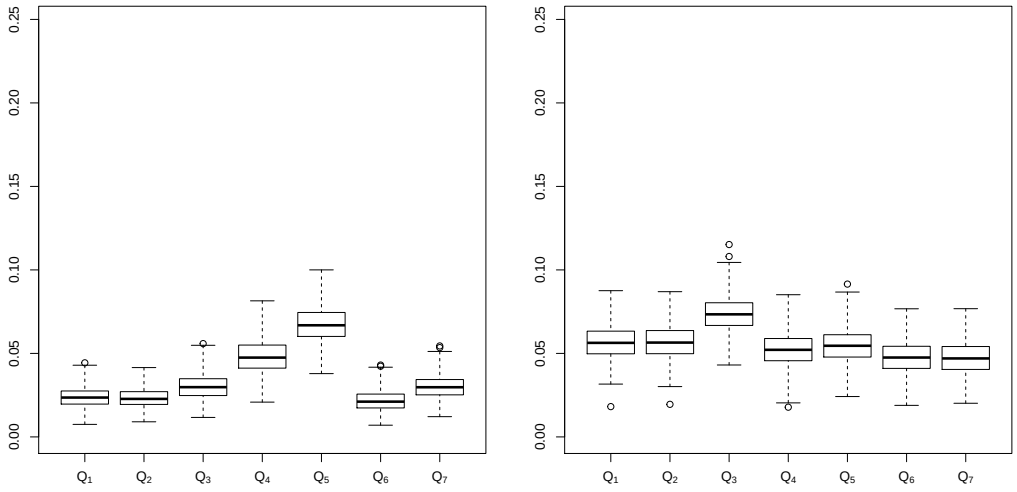
(a) Tikhonov (b) Landweber-Fridman

Figure 4.24: Box plot Total Variational Distance, Local Linear Kernels.



(a) Tikhonov (b) Landweber-Fridman

Figure 4.25: Box plot Total Variational Distance, B-Splines.



(a) Tikhonov (b) Landweber-Fridman

Figure 4.26: Box plot Total Variational Distance, Penalized Local Constant Kernels.

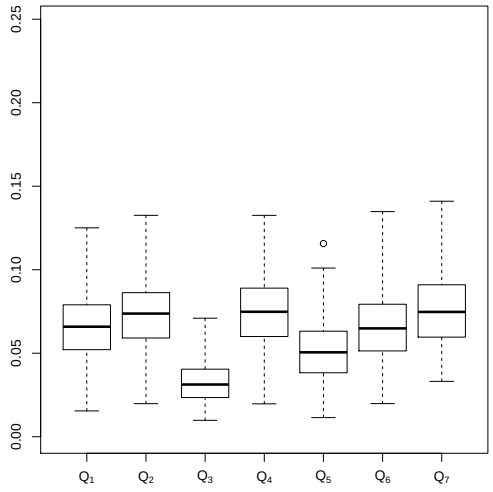


Figure 4.27: Box plot Total Variational Distance, Galerkin.

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 causality and endogeneity in two different type of models and when the object of interest is estimated nonparametrically.

In the standard iid setting, 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.

In continuous time models, we give sufficient condition for the exogeneity of the covariate process and we derive the properties of the estimators of the drift and the diffusion coefficients under Harris recurrence of the joint process.

Of the many issues tackled in this work, we have probably only scratched the surface and future research can proceed in several directions.

The research question tackled in the first chapter is, at the current state, probably the most challenging. Its fortune is intertwined with the one of this first brick that we have started to build, which has to be made more solid in several respects. It remains to be seen if this framework can be useful in the estimation of asset pricing models, especially with respect to the literature on ambiguity on the drift (Chen and Epstein, 2002; Jeong et al., 2009). Moreover, it is also essential to provide a testing procedure for noncausality, as, in practical use, we would like to give hard tools to understand in what context this assumption is satisfied.

The literature on nonparametric instrumental regressions is much better established at the moment, although many aspects could be further developed. The properties and the validity of the wild bootstrap explored in Chapter 4 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 tools could leave room for the simultaneous selection of the bandwidth and the regularization parameter.

Bibliography

Aalen, O. (1980), 'A Model for Nonparametric Regression Analysis of Counting Processes', *Lecture Notes in Statistics* **2**, 1 – 25. 7

Ahn, H., Ichimura, H. and Powell, J. (2004), Simple Estimators for Monotone Index Models, Manuscript, Department of Economics, UC Berkeley. 92

Andrews, D. W. K. (2011), 'Examples of L^2-Complete and Boundedly-Complete Distributions', *Cowles Foundation Discussion Paper* **1801**. 54, 118

Azéma, J., Duflo, M. and Revuz, D. (1969), 'Mesure Invariante des Processus de Markov Récurrents', *Séminaire de Probabilités III Université de Strasbourg* pp. 24–33. 35

Backus, D. K., Foresi, S. and Telmer, C. I. (2001), 'Affine Term Structure Models and the Forward Premium Anomaly', *The Journal of Finance* **56**(1), 279–304. 32

Baillie, R. T. and Bollerslev, T. (1994), 'The Long Memory of the Forward Premium', *Journal of International Money and Finance* **13**(5), 565 – 571. 32

Bandi, F. M. and Moloche, G. (2008), 'On the Functional Estimation Multivariate Diffusion Process', *Working Paper* . 9, 17, 18, 20, 22, 24, 29, 43

Bandi, F. M. and Nguyen, T. N. (2003), 'On the Functional Estimation of Jump-diffusion Models', *Journal of Econometrics* **116**(1-2), 293–328. 8

Bandi, F. M. and Phillips, P. C. B. (2003), 'Fully Nonparametric Estimation of Scalar Diffusion Models', *Econometrica* **71**(1), 241–283. 8, 9, 39, 43

Bandi, F. M. and Phillips, P. C. B. (2010), Nonstationary Continuous-Time Processes, *in* Y. Ait-Sahalia and L. P. Hansen, eds, 'Handbook of Financial Econometrics', Elsevier. 8

Bandi, F. M. and Reno, R. (2009), Nonparametric Stochastic Volatility, Global coe hi-stat discussion paper series, Institute of Economic Research, Hitotsubashi University. 7

Banks, J., Blundell, R. and Lewbel, A. (1997), 'Quadratic Engel Curves and Consumer Demand', *Review of Economics and Statistics* **79**(4), pp. 527–539. 50

Bhalotra, S. and Attfield, C. (1998), 'Intrahousehold Resource Allocation in Rural Pakistan: a Semiparametric Analysis', *Journal of Applied Econometrics* **13**(5), 463–480. 149, 150

Biagini, F., Hu, Y., Øksendal, B. and Zhang, T. (2008), *Stochastic calculus for fractional Brownian motion and applications*, Probability and its applications, Springer. 25

Billingsley, P. (1979), *Probability and Measure*, Wiley series in Probability and mathematical statistics., Wiley. 37

Blanchard, G. and Mathé, P. (2012), 'Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration', *Inverse Problems* **28**(11), 1–24. 71

Blundell, R., Chen, X. and Kristensen, D. (2007), 'Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves', *Econometrica* **75**(6), 1613–1669. 50, 83, 85, 114, 115, 124, 125, 149, 150, 151, 155

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. 114

Blundell, R. and Horowitz, J. (2007), 'A Non-Parametric Test of Exogeneity', *Review of Economic Studies* **74**(4), 1035–1058. 85, 116

Blundell, R. W. and Powell, J. L. (2004), 'Endogeneity in Semiparametric Binary Response Models', *Review of Economic Studies* **71**, 655–679. 91

Borodin, A. N. (1989), 'Brownian Local Time', *Russian Mathematical Surveys* **44**(2), 1–51. 16

Borwein, J. M., Vanderwerff, J. and Wang, X. (2003), 'Local Lipschitz-constant Functions and Maximal Subdifferentials', *Set-Valued Analysis* **11**, 37–67. 19

Breunig, C. and Johannes, J. (2011), 'Adaptive Estimation of Functionals in Nonparametric Instrumental Regressions', *Mimeo* . 52

Brugière, P. (1993), 'Théorème de Limite Centrale pour un Estimateur Non Paramétrique de la Variance d'un Processus de Diffusion Multidimensionnelle', *Annales de l'Institut Henri Poincaré* **29**(3), 357–389. 24

Burda, M. C. (1993), 'The determinants of East-West German migration: Some first results', *European Economic Review* **37**(2-3), 452 – 461. 107

Cardot, H. and Johannes, J. (2010), 'Thresholding projection estimators in functional linear models', *Journal of Multivariate Analysis* **101**(2), 395 – 408. 115, 124

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. 49, 53, 69, 74, 91, 97, 98, 116, 117

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'. 73, 74, 140

Centorrino, S. (2013), On the Choice of the Regularization Parameter in Nonparametric Instrumental Regressions, Technical report, Toulouse School of Economics. 101, 115, 121, 129, 131

Centorrino, S., Fève, F. and Florens, J.-P. (2013a), 'Implementation, Simulations and Bootstrap in Nonparametric Instrumental Variable Estimation', *Mimeo - Toulouse School of Economics* . 73, 79, 81, 85, 109

Centorrino, S. and Florens, J.-P. (2013), 'Nonparametric Instrumental Variable Estimation of Binary Regression Models', *Mimeo - Toulouse School of Economics* . 53

Chan, K. C., Karolyi, G. A., Longstaff, F. A. and Sanders, A. B. (1992), 'An Empirical Comparison of Alternative Models of the Short-Term Interest Rate', *The Journal of Finance* **47**(3), pp. 1209–1227. 28

Chen, X. and Pouzo, D. (2012), 'Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals', *Econometrica* **80**(1), 277–321. 49, 91, 116, 117, 137

Chen, X. and Reiss, M. (2011), 'On Rate Optimality for Ill-Posed Iinverse Problems in Econometrics', *Econometric Theory* **27**(3), 497–521. 56, 93, 139

Chen, Z. and Epstein, L. (2002), 'Ambiguity, Risk, and Asset Returns in Continuous Time', *Econometrica* **70**(4), pp. 1403–1443. 159

Comte, F. and Renault, E. (1996), 'Noncausality in Continuous Time Models', *Econometric Theory* **12**(02), 215–256. 11, 12

Conway, J. (2000), *A Course in Operator Theory*, Graduate Studies in Mathematics, American Mathematical Society. 53

Creedy, J., Lye, J. and Martin, V. L. (1996), 'A Non-Linear Model of the Real US/UK Exchange Rate', *Journal of Applied Econometrics* **11**(6), 669–686. 7

Creedy, J. and Martin, V. L. (1994), 'A Model of The Distribution of Prices', *Oxford Bulletin of Economics and Statistics* **56**(1), 67–76. 7

Darolles, S., Fan, Y., Florens, J. P. and Renault, E. (2011*a*), 'Nonparametric Instrumental Regression', *Econometrica* **79**(5), 1541–1565. 2, 49, 50, 52, 54, 55, 57, 60, 63, 64, 71, 79, 91, 92, 93, 97, 98, 99, 102, 111, 115, 116, 117, 118, 120, 123, 130, 141

Darolles, S., Fan, Y., Florens, J. P. and Renault, E. (2011*b*), 'Supplement to Nonparametric Instrumental Regression', *Econometrica Online Appendix* . 112

D'Haultfoeuille, X. (2011), 'On the Completeness Condition in Nonparametric Instrumental Problems', *Econometric Theory* **27**, 460–471. 55, 118

Dong, Y. (2010), 'Endogenous Regressor Binary Choice Models without Instruments, with an Application to Migration', *Economics Letters* **107**(1), 33 – 35. 104, 107

Dozzi, M. (2003), Occupation Density and Sample Path Properties of N-parameter Processes, *in* V. Capasso, E. Merzbach, B. Ivanoff, M.Dozzi, R. Dalang and T. Mountford, eds, 'Topics in Spatial Stochastic Processes', Springer-Verlag. 19

Engl, H. W., Hanke, M. and Neubauer, A. (2000), *Regularization of Inverse Problems*, Vol. 375 of *Mathematics and Its Applications*, Kluwer Academic Publishers, Dordrecht. 56, 64, 66, 71, 72, 73, 75, 78, 115, 117

Escanciano, J. C., Jacho-Chavez, D. and Lewbel, A. (2011), Identification and Estimation of Semiparametric Two Step Models, Technical report, Boston College. 104, 107

Evans, M. D. D. and Lewis, K. K. (1995), 'Do Long-Term Swings in the Dollar Affect Estimates of the Risk Premia?', *The Review of Financial Studies* **8**(3), pp. 709–742. 32

Fernandes, M. (2006), 'Financial Crashes as Endogenous Jumps: Estimation, Testing and Forecasting', *Journal of Economic Dynamics and Control* **30**(1), 111–141. 7

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. 142, 143

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). 52, 56, 57, 62, 63, 70, 71, 79, 82, 97, 101, 115, 119, 121

Fève, F. and Florens, J.-P. (2013), 'Non Parametric Analysis of Panel Data Models with Endogenous Variables', *Journal of Econometrics* **Forthcoming**. 82, 129, 131

Florens, J. and Fougere, D. (1996), 'Noncausality in Continuous Time', *Econometrica* **64**(5), 1195–1212. 11

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. 72, 73, 77

Florens, J.-P., Johannes, J. and Van Bellegem, S. (2012), 'Instrumental Regressions in Partially Linear Models', *The Econometrics Journal* **15**(2), 304–324. 107, 155, 160

Florens, J.-P. and Racine, J. (2012), 'Nonparametric Instrumental Derivatives', *Mimeo* . 73, 81, 86, 97, 115, 122, 123, 127, 128, 129, 130

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. 93, 102, 130, 139

Florens-Zmirou, D. (1993), 'On Estimating the Diffusion Coefficient from Discrete Observations', *Journal of Applied Probability* **30**(4), 790–804. 20, 43

Geman, D. and Horowitz, J. (1980), 'Occupation Densities', *The Annals of Probability* **8**(1), 1–67. 9, 16

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. 52

Hall, P. and Horowitz, J. L. (2005), 'Nonparametric Methods for Inference in the Presence of Instrumental Variables', *Annals of Statistics* **33**(6), 2904–2929. 49, 50, 78, 79, 91, 116

Hall, P. and Racine, J. S. (2013), 'Infiinite Order Cross-Validated Local Polynomials Regressions', *WP - McMaster University, Department of Economics* **5**. 160

Hansen, B. E. (2008), 'Uniform Convergence Rates for Kernel Estimation with Dependent Data', *Econometric Theory* **24**(03), 726–748. 56, 97, 112

Hansen, L. P. and Sargent, T. J. (1983), 'The Dimensionality of the Aliasing Problem in Models with Rational Spectral Densities', *Econometrica* **51**(2), 377–387. 8

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. 136

Härdle, W., Liang, H. and Gao, J. (2000), *Partially Linear Models*, Contributions to Statistics Series, Heidelberg: Physica-Verlag. 106

Härdle, W. and Mammen, E. (1993), 'Comparing Nonparametric Versus Parametric Regression Fits', *The Annals of Statistics* **21**(4), pp. 1926–1947. 137

Härdle, W. and Marron, J. S. (1991), 'Bootstrap Simultaneous Error Bars for Nonparametric Regression', *The Annals of Statistics* **19**(2), pp. 778–796. 136, 137

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. 50

Hazelton, M. L. (2007), 'Bias Reduction in Kernel Binary Regression', *Computational Statistics and Data Analysis* **51**(9), 4393 – 4402. 100

Heckman, J. J. (1978), 'Dummy Endogenous Variables in a Simultaneous Equation System', *Econometrica* **46**(4), pp. 931–959. 96

Herrero, D. A. (1991), 'Diagonal Entries of a Hilbert Space Operator', *Rocky Mountain Journal of Mathematics* **21**(2), 857–865. 69

Hoderlein, S. and Holzmann, H. (2011), 'Demand Analysis as an Ill-posed Inverse Problem with Semiparametric Specification', *Econometric Theory* **27**, 609–638. 50, 84, 85, 114, 115

Höpfner, R. and Löcherbach, E. (2003), *Limit Theorems for Null Recurrent Markov Processes*, American Mathematical Society. 15, 17, 35

Horowitz, J. L. (1992), 'A Smoothed Maximum Score Estimator for the Binary Response Model', *Econometrica* **60**(3), 505–31. 94

Horowitz, J. L. (2011), 'Applied Nonparametric Instrumental Variables Estimation', *Econometrica* **79**(2), 347–394. 2, 49, 50, 83, 93, 97, 115, 124

Horowitz, J. L. (2012), 'Adaptive Nonparametric Instrumental Variables Estimation: Empirical Choice of the Regularization Parameter', *Mimeo - NorthWestern University* . 52, 83, 115, 125, 131

Horowitz, J. L. and Lee, S. (2012), 'Uniform Confidence Bands for Functions Estimated Nonparametrically with Instrumental Variables', *Journal of Econometrics* **168**(2), 175 – 188. 116, 137, 142

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. 106

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. 62, 131

Iacus, S. (2008), *Simulation and inference for Stochastic Differential Equations: with R examples*, Springer Series in Statistics, Springer. 28

Ichimura, H. (1993), 'Semiparametric Least squares (SLS) and Weighted SLS Estimation of Single-Index Models', *Journal of Econometrics* **58**(1–2), 71 – 120. 94

Jäger, S. and Kostina, E. (2005), 'Parameter Estimation for Forward Kolmogorov Equation with Application to nonlinear Exchange rate Dynamics', *PAMM* **5**(1), 745–746. 7

Jeong, D., Kim, H. and Park, J. Y. (2009), 'Does Ambiguity Matter? Estimating Asset Pricing Models with a Multiple-Priors Recursive Utility', *Mimeo* . 159

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. 97, 115, 122

Karatzas, I. and Shreve, S. (1991), *Brownian Motion and Stochastic Calculus*, Springer-Verlag. 6, 11, 13

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. 140

Kauermann, G., Claeskens, G. and Opsomer, J. D. (2009), 'Bootstrapping for Penalized Spline Regression', *Journal of Computational and Graphical Statistics* **18**(1), 126–146. 140

Kleibergen, F. and Paap, R. (2006), 'Generalized Reduced Rank Tests using the Singular Value Decomposition', *Journal of Econometrics* **133**(1), 97 – 126. 106

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. 92, 94

Kleptsyna, M., Le Breton, A. and Roubaud, M.-C. (2000), 'Parameter Estimation and Optimal Filtering for Fractional Type Stochastic Systems', *Statistical Inference for Stochastic Processes* **3**, 173–182. 25, 26

Krein, S. and Petunin, Y. (1966), 'Scales of Banach Spaces', *Russian Math. Survey* **21**(2), 89–168. 72

Kress, R. (1999), *Linear Integral Equations*, Applied mathematical sciences, Springer-Verlag. 51, 53, 118

Lewbel, A. (1991), 'The Rank of Demand Systems: Theory and Nonparametric Estimation', *Econometrica* **59**(3), pp. 711–730. 50

Li, Q. and Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press. 62, 71, 85, 119, 131

Liese, F. and Vajda, I. (2006), 'On Divergences and Informations in Statistics and Information Theory', *Information Theory, IEEE Transactions on* **52**(10), 4394 –4412. 142

Locherbach, E. and Loukianova, D. (2008), 'On Nummelin Splitting for Continuous Time Harris Recurrent Markov Processes and Application to Kernel Estimation for Multi-dimensional Diffusions', *Stochastic Processes and their Applications* **118**(8), 1301–1321. 9, 17

Lukas, M. A. (1993), 'Asymptotic Optimality of Generalized Cross-Validation for Choosing the Regularization Parameter', *Numerische Mathematik* **66**(1), 41–66. 52

Lukas, M. A. (2006), 'Robust Generalized Cross-Validation for choosing the Regularization Parameter', *Inverse Problems* **22**(5), 1883–1902. 52

Ma, S. and Racine, J. (2013), 'Additive Regression Splines With Irrelevant Categorical and Continuous Regressors', *Statistica Sinica* **23**, 515–541. 131

Manski, C. F. (1985), 'Semiparametric Analysis of Discrete Response : Asymptotic Properties of the Maximum Score Estimator', *Journal of Econometrics* **27**(3), 313–333. 94

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. 126

Mark, N. C. and Moh, Y.-K. (2007), 'Official Interventions and the Forward Premium Anomaly', *Journal of Empirical Finance* **14**(4), 499 – 522. 33

Marteau, C. and Loubes, J.-M. (2012), 'Adaptive Estimation for an Inverse Regression Model with Unknown Operator', *Statistics & Risk Modeling* **29**(3), 215–242. 52

Mathé, P. and Tautenhahn, U. (2011), 'Regularization under General Noise Assumptions', *Inverse Problem* **27**(3), 35–41. 71

Matzkin, R. L. (1991), 'Semiparametric Estimation of Monotone and Concave Utility Functions for Polychotomous Choice Models', *Econometrica* **59**(5), 1315–27. 94

Matzkin, R. L. (1992), 'Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models', *Econometrica* **60**(2), 239–70. 94

McKean, H. P. (1969), *Stochastic Integrals*, Academic Press, Inc. 35

Morozov, V. (1967), 'Choice of a Parameter for the Solution of Functional Equations by the Regularization Method', *Sov. Math. Doklady* **8**, 1000–1003. 71

Neal, R. M. (2003), 'Slice Sampling', *Annals of Statistics* **31**(3), 705–767. 78

Newey, W. K. and Powell, J. L. (2003), 'Instrumental Variable Estimation of Nonparametric Models', *Econometrica* **71**(5), 1565–1578. 49, 91, 93, 116, 118

Norros, I., Valkeila, E. and Virtamo, J. (1999), 'An Elementary Approach to a Girsanov Formula and Other Analytical Results on Fractional Brownian Motions', *Bernoulli* **5**(4), pp. 571–587. 25, 27

Øksendal, B. (2003), *Stochastic Differential Equations: an Introduction with Applications*, Universitext (1979), Springer. 6, 17, 39

Pagan, A. and Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge University Press. 18, 22

Panel Study of Income Dynamics (2003). 104

Park, J. Y. (2005), The Spatial Analysis of Time Series, Discussion papers, Indiana University. 16

Park, J. Y. (2008), 'Martingale Regression and Time Change', *Working Paper* . 7, 11

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge University Press. 1

Phillips, P. (1973), 'The Problem of Identification in Finite Parameter Continuous Time Models', *Journal of Econometrics* **1**(4), 351–362. 8, 28

Phillips, P. C. B. and Ploberger, W. (1996), 'An Asymptotic Theory of Bayesian Inference for Time Series', *Econometrica* **64**(2), 381–412. 40

Phillips, P. and Park, J. (1998), 'Nonstationary Density Estimation and Kernel Autoregression', *Cowles Foundation Discussion Paper* . 9

Protter, P. E. (2003), *Stochastic Integration and Differential Equations, 2nd ed.*, Springer-Verlag. 13, 43, 45

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* 131

Rao, B. (2010), *Statistical Inference for Fractional Diffusion Processes*, Wiley Series in Probability and Statistics, John Wiley & Sons. 25

Revuz, D. (1984), *Markov chains*, North-Holland mathematical library, North-Holland. 45

Revuz, D. and Yor, M. (1999), *Continous Martingale and Brownian Motion*, Springer-Verlag. 9, 14

Rothe, C. (2009), 'Semiparametric estimation of binary response models with endogenous regressors', *Journal of Econometrics* **153**(1), 51 – 64. 91, 92, 102, 107, 108

Ruppert, D. and Wand, M. (1994), 'Multivariate Locally Weighted Least Squares Regression', *The Annals of Statistics* **22**(3), 1346–1370. 18

Santos, A. (2012), 'Inference in nonparametric instrumental variables with partial identification', *Econometrica* **80**(1), 213–275. 116, 137

Schienle, M. (2011), Nonparametric Nonstationary Regression with Many Covariates, Discussion papers, Humboldt University. 9, 29

Signorini, D. F. and Jones, M. C. (2004), 'Kernel Estimators for Univariate Binary Regression', *Journal of the American Statistical Association* **99**(465), 119–126. 100

Sokullu, S. (2010), 'Nonparametric Analysis of Two-Sided Markets'. 115, 137

Stone, C. J. (2005), 'Nonparametric M-regression with free knot Splines', *Journal of Statistical Planning and Inference* **130**(1–2), 183 – 206. 131

Stone, C. J. and Huang, J. Z. (2003), 'Statistical Modeling of Diffusion Processes with Free Knot Splines', *Journal of Statistical Planning and Inference* **116**(2), 451–474. 7

Vogel, C. (2002), *Computational Methods for Inverse Problems*, Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics. 52, 115

Wahba, G. (1977), 'Practical Approximate Solutions to Linear Operator Equations when the Data are Noisy', *SIAM Journal on Numerical Analysis* **14**(4), 651–667. 52

Zivot, E. (2000), 'Cointegration and Forward and Spot Exchange Rate Regressions', *Journal of International Money and Finance* **19**(6), 785 – 812. 32