UNIVERSITÀ
DEGLI STUDI
DI FERRARA
- EX LABORE FRUCTUS -

PH.D. COURSE IN
ECONOMICS AND MANAGEMENT OF INNOVATION AND SUSTAINABILITY -
EMIS

Cycle: XXX
Director: Prof. Stefano Azzali
Coordinator (Ferrara): Prof. Emidia Vagnoni

Essays in nonparametric econometrics with applications to the economics of productivity and innovation

Scientific/Disciplinary Sector (SDS): SECS-P/05 ECONOMETRIA

Candidate:	*Supervisor:*
Dott. Georgios GIOLDASIS	Prof. Antonio MUSOLESI
(signature)	(signature)

Years: 2014/2017

UNIVERSITÀ
DEGLI STUDI
DI FERRARA
- EX LABORE FRUCTUS -

DOTTORATO DI RICERCA IN
ECONOMIA E MANAGEMENT DELL'INNOVAZIONE E DELLA
SOSTENIBILITÀ - EMIS

Ciclo: XXX
Coordinatore: Prof. Stefano Azzali
Coordinatore (locale Ferrara): Prof. Emidia Vagnoni

Saggi in econometria nonparametrica con applicazioni all'economia della produttività e dell'innovazione

Settore Scientifico Disciplinare: SECS-P/05 ECONOMETRIA

Dottorando:
Dott. Georgios GIOLDASIS

Tutore:
Prof. Antonio MUSOLESI

(firma)

(firma)

Anni: 2014/2017

Dedication

To my mother Katerina, to my daughter Katerina.

Acknowledgements

I would like to thank from my heart my supervisor, Antonio Musolesi, for his patience, guidance and support. His constant and uninterrupted help are precious to me and have allowed me to improve considerably my knowledge and scientific understanding.

I would like to express my gratitude to all coauthors, who have contributed during these years to my improvement as a researcher. In alphabetic order: Davide Antonioli and Michel Simioni.

I would like to thank Jeffrey Racine, Jeffrey Wooldridge, Pierre Mohnen and Patrick Sevestre for their insightful suggestions that gave a boost forward to this thesis.

Special thanks to Massimiliano Mazzanti for all his multi-level support during all this period.

I would like to thank my family Katerina, Alekos and Sofia, as well as Giorgos and Rania, for being so supportive to me.

Finally, I would like to thank my wife Natasha for being so patient with me and always beside me throughout all this exciting journey.

Summary

This thesis is concerned with applying frontier methods in econometric theory, to revisit relevant economic questions concerning productivity, technology and innovation. The focal point of this thesis is to try to employ econometric techniques that may allow us to depart from the usual assumptions of linearity or additivity of the model, as is usually found in economic literature. Therefore, a binding common ground of all the chapters of this thesis is relaxing these restrictive assumptions, in economic topics relevant to productivity and innovation, in micro and macro level. In this framework, recent advancements in nonparametric econometric theory allow relaxing such model restrictions.

The first chapter of the thesis is the only one within a parametric framework. Common literature on econometrics of productivity and technical change (TC) assume additivity of TC in the model, therefore introducing Hicks-neutrality of TC. The chapter estimates a production function at firm level that allows departing from the standard hypothesis of Hicks-neutral technical change. Simultaneously, it is coping with the endogeneity of innovation, the latter being considered a measure of TC. Parametric specifications that allow non-Hicks neutral technical change are derived. The chapter also presents testable conditions, for common parametric approximations, under which Hicks neutrality holds. Cobb-Douglas specifications are estimated adopting IV methods for heterogeneous effect of innovation on productivity. The empirical results reject Hicks neutrality towards the presence of a capital-saving TC. Finally, this chapter serves as a link with the nonparametric approaches developed in the following ones.

The second chapter addresses the issue of localization of technical change using the firm level dataset of the previous chapter. The orthodoxy that technical progress increases productivity at all factor proportions is questioned by introducing a nonparametric model. In order to allow the detection of local effects of technical change, a kernel regression approach is adopted. The endogeneity of innovation is tackled by employing an IV estimation procedure based on regularization. The results reveal that technical change cannot be modeled as a shift in the production function and that technical change is localized and not Hicks neutral.

The third chapter revisits the issue of international R&D spillovers by using nonparametric methods and tests the validity of the main results provided in the literature with respect to the possible existence of nonlinearities, threshold effects and non-additive relations. It considers a sieve estimation of a panel data model of technology diffusion among countries, paying attention to the issue of error cross sectional dependence. The adopted semiparametric approach is an extension of the parametric factor model by Pesaran (2006). The comparison between the parametric and the semiparametric approach reveals a better performance of the latter. From

an economic viewpoint, the results show new evidence with respect to the benefits of countries from domestic and foreign R&D.

Finally, the last chapter of the thesis is a review of the nonparametric kernel regression. Following, mainly, the work of Li and Racine (2007), it summarizes the key features of least squares cross validation and the kernel regression using local constant, local linear or local polynomial approaches. Moreover, the nonparametric IV kernel regression is presented, along with the relevant topics of ill-posed inverse problems and regularization methods. This chapter serves also as an informal *"appendix"* of the previous chapters, especially for the ones concerning kernel regression, because it provides useful insights of the underlying methodologies.

Contents

Chapter 1

Estimating a non-neutral production function: a heterogeneous treatment effect approach

joint with Davide Antonioli and Antonio Musolesi

Abstract

This paper addresses the issue of estimating a production function that allows us to depart from the standard hypothesis of Hicks neutrality while also coping with the endogeneity of a dummy innovation variable. We consider specifications that relax Hicks neutrality, and we derive the testable conditions for common parametric approximations under which Hicks neutrality holds. The model is estimated through instrumental variables methods, allowing for a heterogeneous effect of innovation on the production process. The econometric analysis rejects Hicks neutrality and highlights three main features: i) a capital-saving technology of innovative with respect to non-innovative firms, ii) a locally progressive technical change and iii) fully heterogeneous technologies when comparing innovative to non-innovative firms.

Keywords: Biased technical change; Hicks neutrality; Innovation; Productivity; Knowledge production function; CDM model; Instrumental variables; heterogeneous treatment effect

JEL classification: C26; C31; D24; O33

1.1 Introduction

Assessing the effect of technical change (TC) on the production process, both theoretically and empirically and at all levels of aggregation, is one of the major concerns amongst economists because TC is largely recognized as one of the fundamental drivers of economic development. In neoclassical production theory, one of the most commonly used classifications of TC dates to the seminal work of Hicks (1963). In particular, according to Hicks, *"we can classify inventions according as their initial effects are to increase, leave unchanged or diminish the ratio of the marginal product of capital to that of labor"* (Hicks, 1963, p.121). The above types of inventions are referred to as *labour-saving*, *neutral* or *capital-saving* inventions, respectively.

Formally, *Hicks neutrality* (HN) requires that the marginal rate of technical substitution (MRTS) between each pair of inputs be independent of technical change. However, this definition of neutrality has received different interpretations; Hicks' requirement that the firm be in a state of *"internal equilibrium"* (Hicks, 1963, p.113, 234, 236) does not specify explicitly whether the effect of TC should be measured along the firm's expansion path or at the optimal factor proportions, which has resulted in controversy (see, e.g., Antle and Capalbo, 1988).

To clarify this ambiguity, Blackorby et al. (1976) consider three different definitions. The first was suggested by Hicks and is already referred to as HN. The second is *implicit Hicks neutrality* (IHN) and accounts for a ray preserving TC. They also introduce a third definition that they term *extended Hicks neutrality* (EHN). While HN translates equivalently to a weak separability assumption between TC and the inputs (Morimoto, 1974), EHN implies TC that is strongly separable from the inputs in a production function. The three different definitions considered in Blackorby et al. (1976), namely HN, IHN and EHN, are not equivalent in general and coincide only if the production function is input homogeneous.

While HN has received significant attention, several streams of literature question its plausibility. Jones (1965) describes a two-sector economy in which HN is unlikely, although HN may appear in each industry. Further, Steedman (1985) presents a model of interconnected industries in which, under weak alternative sufficient conditions, HN is impossible, ultimately concluding that the compatibility of HN with the other assumptions of a model should be examined. Acemoglu (2015) also mentions that in Atkinson and Stiglitz's (1969) seminal paper, it is implied that when TC is localized to specific factor proportions, then it is biased. Chambers (1988, p. 206) describes other fairly common types of TC that depart from being HN. In particular, if TC is locally progressive or regressive, its effect on productivity results in isoquants of the production function that intercept the old ones. Obviously, such TC does not follow HN, IHN or EHN.

Empirically, at a microeconomic level, since the seminal works by Pakes and Griliches (1984) and Griliches (1998), the question concerning the effect of TC on the production process has often been addressed within the framework of the so-called knowledge production function (KPF), where innovation activity is the main source of technical and knowledge improvements. The KPF is a conceptual framework that suggests a possible causal relationship between unobservable knowledge capital and related observable variables such as innovation inputs (e.g., research and development, R&D), innovation outputs (e.g., patents) and firms' performance. The

KPF provided the basis for econometric analyses connecting different and relevant aspects of innovative activities. In their influential article, Crépon et al. (1998) propose a multiple-equation econometric model – commonly labeled CDM – that has a similar structure to Griliches' original conception and uses appropriate estimation methods for taking into account both the potential endogeneity of some of the explanatory variables and the particular nature of the dependent variables. In recent years, the availability of survey data, such as those obtained from the Community Innovation Surveys, has allowed the use of a direct and binary measure of the innovative output that is then introduced into a production function framework as an endogenous dummy variable that accounts for TC (see, e.g., Mairesse and Mohnen, 2010).

Recently, developments and generalizations of the CDM approach have been accomplished, such as the introduction of dynamics into the model, the assessment of measurement errors or the consideration of a Schumpeterian perspective (Lööf et al., 2016). The extremely vast literature clearly indicates a positive and significant effect of innovation on productivity (Mohnen and Hall, 2013) and sheds light on many other relevant relationships among variables.

However, a crucial maintained assumption in this literature is HN. In standard econometric models, innovation is additively introduced into the production function specification. Additivity is equivalent to a multiplicative decomposability of the production function into a function of innovation and a function of the inputs. Uzawa and Watanabe (1961) prove the equivalence of HN and the decomposability of the production function. Therefore, standard econometric models usually impose a strict condition, specifically HN, when assessing the effect of innovation on productivity.

The present study contributes to the literature on the econometrics of productivity and TC in two ways. First, we estimate a model that allows us to relax the neutrality assumption. Second, we extend the analysis by Blackorby et al. (1976) and provide testable conditions, for common parametric specifications, namely Cobb-Douglas (CD) and translog (TL), under which HN, IHN or EHN hold.

The econometric analysis developed here exploits recent advances in the econometric theory of instrumental variables (IVs) with cross-sectional data and a binary endogenous regressor. In the empirical literature, the standard approach considers a parametric approximation of the production function – typically CD – in which innovation enters additively, and the main focus is on the endogeneity of innovation and its binary nature (Musolesi and Huiban, 2010; Mohnen and Hall, 2013). We depart from this literature by allowing for an heterogeneous effect of the dummy endogenous innovation variable by adopting the approach proposed by Wooldridge (2003, 2010). This allows the technology parameters to differ between innovative and non-innovative firms, finally permitting us to address endogeneity while relaxing the assumption of neutrality.

The remainder of the paper is organized as follows. Section 1.2 describes the theory of TC and presents HN, IHN and EHN, along with their relationships. It also introduces empirical specifications that allow non-neutrality and provides simple conditions under which each of the different definitions of neutrality holds. The data set is described in section 1.3. This section also presents the results and comments. Finally, section 1.4 concludes the paper. In supplementary appendices, we present the relationships among HN, IHN and EHN (Appendix A), and we also provide conditions under which neutrality holds within a TL framework (Appendix B). Appendix C provides robustness checks.

1.2 Hicks neutrality: theory and econometrics

1.2.1 A selective review of theory

Consider a production function $f : \mathbb{R}_+^{k+1} \to \mathbb{R}_+^1$ that is expressed as

$$Y = f\left(\mathbf{X}, I\right). \tag{1.1}$$

$\mathbb{R}_+$ denotes the positive real numbers. It is commonly assumed that f is twice-differentiable in $\mathbf{X}$, strictly quasiconcave in $\mathbf{X}$ and non-decreasing in I. Y is a measure of output such as value added. $\mathbf{X} = \{X_1, X_2, \ldots, X_k\}^T$ is a *k*-dimensional vector of inputs that includes conventional production inputs such as capital and labour. I represents innovation activity, which is regarded in production theory as the main source of TC.

In a production function with multiple inputs, i.e., for $k > 1$, technical change might differently affect the marginal productivity of each input. Therefore, it can be classified as *biased* or *neutral*, according to whether the ratios between marginal products are changed. Hicks (1963) is the first to distinguish inventions according to the above observation. Particularly, he focuses on cases with two input factors, namely capital and labour. Under the requirement that the firm remains in a state of *"internal equilibrium"*, he defines an invention as *labour-saving*, *neutral* or *capital-saving* according to whether its initial effect is to increase, leave unchanged or decrease the ratio of the marginal product of capital to that of labour, respectively. Uzawa and Watanabe (1961) generalize Hicks' classification to cases of more than two factors of production.

Hicks' classification has resulted in controversial interpretations concerning whether the effect described above should be observed along the expansion path of the firm or along the optimal factor proportions. In particular, while Blackorby et al. (1976) state that this effect should be considered along the firm's expansion path, Kennedy and Thirlwall (1972, 1977) consider HN along a ray from the origin and argue that Hicks' definition of neutrality does not imply an expansion *path preserving* innovation.

Nevertheless, Blackorby et al. (1976) attempt to clarify this ambiguity by distinguishing three different definitions of neutrality. The first generalizes Hicks' definition for cases of firms with more than two inputs; neutrality holds if the MRTS between each pair of inputs is independent of I. This definition is denoted HN and translates to the following expression:

$$\frac{\partial f(\mathbf{X}, I)/\partial X_r}{\partial f(\mathbf{X}, I)/\partial X_l} = \phi_{rl}(\mathbf{X}),\ \ \forall r \neq l, \tag{1.2}$$

where $\phi_{rl}, r, l = 1, 2, \ldots, k, r \neq l$ are functions of $\mathbf{X}$. Therefore, HN requires that the MRTS for every $X_r,\ X_l, r \neq l$ component of $\mathbf{X}$ be a function of $\mathbf{X}$ only. Blackorby et al. (1976) refer to innovation as HN if it is expansion path preserving. Further, it is proven (see Morimoto, 1974) that HN holds if and only if the inputs $\mathbf{X}$ are weakly separable from I in f, such that the production function is described by:

$$f(\mathbf{X}, I) = g(h(\mathbf{X}), I), \tag{1.3}$$

where g and h are real functions. The above implies that HN and weak separability of $\mathbf{X}$ from I impose equivalent restrictions on f.

Blackorby et al. (1976) introduce a second definition of neutrality, namely IHN, that requires the MRTS between each pair of inputs to be independent of I, at constant factor proportions. This definition implies that the MRTS is homogeneous of degree zero in the inputs $\mathbf{X}$. Further, they refer to innovation as IHN if it is *ray preserving* and prove that IHN holds if and only if the transformation function $\tilde{f}(\mathbf{X}, Y, I) = \max_{\lambda}\{\lambda > 0 : f(\lambda^{-1}\mathbf{X}, I) \geq Y\}$ can be written in the following form:

$$\tilde{f}(\mathbf{X}, Y, I) = \tilde{g}(\tilde{h}(\mathbf{X}, Y), Y, I), \tag{1.4}$$

where $\tilde{g}$ and $\tilde{h}$ are real functions. $\tilde{f}$ is a distance function that uniquely represents the technology f. According to (1.4), I is IHN if it is separable from $\mathbf{X}$ in $\tilde{f}$.

A third definition of neutrality accounts for cases in which I is strongly separable from $\mathbf{X}$ in the production function f (see Chambers, 1988, p.45). By definition, EHN holds if:

$$\frac{\partial}{\partial X_r}\left[\ln f(\mathbf{X}, I)\right] = \phi_r(\mathbf{X}), \ \forall r = 1, 2, \ldots, k, \tag{1.5}$$

where $\phi_r, \ r = 1, 2, \ldots, k$ are functions of $\mathbf{X}$. Moreover, it is proven that EHN holds if and only if the production function can be multiplicatively decomposed into a function $\bar{h}$ of inputs only and a function $\bar{g}$ of I only, such that:

$$f(\mathbf{X}, I) = \bar{g}(I)\bar{h}(\mathbf{X}). \tag{1.6}$$

Obviously, if innovation is EHN, it is also HN, because eq.(1.6) implies eq.(1.3).

In general, the three definitions of neutrality are not equivalent. Antle and Capalbo (1988, p.38) note that innovation that results in a renumbering of the isoquants is also neutral in terms of the MRTS at points on the expansion path but may not be neutral in terms of optimal factor proportions. Moreover, a priori, neither HN nor IHN is sufficient for EHN. Nevertheless, there are conditions under which HN, IHN and EHN coincide. First, under the assumption that the production function is input homogeneous, the three definitions are equivalent (see Morimoto, 1974; Blackorby et al., 1976). IHN and HN are equivalent if and only if the production function is input homothetic, while the equivalence of IHN and EHN implies the homotheticity of f. A detailed presentation of the relationships among HN, IHN and EHN is provided in Appendix A.

1.2.2 Econometric specification: relaxing and testing Hicks neutrality

The above definitions of neutrality can be assessed using a suitable econometric framework. We assume that the inputs of production are the conventional factors capital (K) and labour (L). The innovation activity of the firm is described by a binary variable $I \in \{0, 1\}$, as in many previous works (see, e.g., Mairesse and Mohnen, 2010; Mohnen and Hall, 2013). We consider a specification that allows us to relax the neutrality assumption. This is achieved by focusing on a model in which innovation not only produces a shift in the production technology – as is usually imposed in econometric analyses – but is also interacted with labour and capital inputs, thus allowing the technology parameters to differ between innovative and non-innovative firms. For simplicity, comparability with previous studies and data congruence, the main analysis is conducted assuming a CD technology, while in Appendix B, we derive testable conditions concerning the TL form and its relationship with HN.

Empirical studies usually consider a CD production function in which innovation enters additively. Given a sample of n observations, the econometric model is described by:

$$\ln Y_i = \ln f^{\texttt{CD}}(K_i, L_i, I_i, \epsilon_i) = \alpha + \alpha_I I_i + \alpha_K \ln K_i + \alpha_L \ln L_i + \epsilon_i, \tag{1.7}$$

where i denotes the i-th observation, and ϵ is the error term. Then, the MRTS[1] between L and K is given by:

$$MRTS_i^{\texttt{CD}} = -\frac{\alpha_L}{\alpha_K}\frac{K_i}{L_i}, \tag{1.8}$$

which, being independent of I, implies that HN holds. Moreover, IHN is also imposed because for constant factor proportions, the MRTS is both independent of I and constant. Finally, I is EHN because (1.5) holds, that is:

$$\frac{\partial}{\partial K_i} E\left[\ln Y_i | K_i, L_i, I_i\right] = \alpha_K \frac{1}{K_i} \quad \& \quad \frac{\partial}{\partial L_i} E\left[\ln Y_i | K_i, L_i, I_i\right] = \alpha_L \frac{1}{L_i}.$$

Alternatively, it suffices to observe that I is strongly separable from the inputs in $f_{\texttt{CD}}$. In summary, in the case of a CD production function with added innovation as described by (1.7), all definitions of neutrality described above are satisfied. This equivalence is also implied by the input homogeneity of $f^{\texttt{CD}}$.

In (1.7), innovation additively enters the production function. To relax HN, we consider a CD production function in which innovation also interacts with the inputs, as given by:

$$\ln Y_i = \ln f_{\texttt{CDh}}(K_i, L_i, I_i, \epsilon_i) = \alpha + \alpha_I I_i + \alpha_K \ln K_i + \alpha_{KI} I_i \ln K_i + \alpha_L \ln L_i + \alpha_{LI} I_i \ln L_i + \epsilon_i. \tag{1.9}$$

Then, the MRTS between L and K becomes:

$$MRTS_i^{\texttt{CDh}} = -\frac{\alpha_L + \alpha_{LI} I_i}{\alpha_K + \alpha_{KI} I_i}\frac{K_i}{L_i}. \tag{1.10}$$

In this case, neither HN nor IHN hold, unless the following condition holds:

$$\alpha_{KI}\alpha_L = \alpha_{LI}\alpha_K. \tag{1.11}$$

[1] Computation of the MRTS involves deriving the marginal products of the inputs. The marginal product of $X_r, r = 1, 2, \ldots, k$ at $(Y_i, \mathbf{X}_i, I_i), i = 1, 2, \ldots, n$ is given by $\partial E(Y_i|\mathbf{X}_i, I_i)/\partial X_{r,i}$, which assumes the exogeneity of all variables, that is, $E(\epsilon_i|\mathbf{X}_i, I_i) = 0, i = 1, 2, \ldots, n$. (see Verbeek, 2008, for a discussion of marginal effects in the linear model) This assumption is considered in this section only to simplify the presentation without losing any relevant information.

Moreover, the definition of EHN does not hold, generally, because:

$$\frac{\partial}{\partial K_i}E\left[\ln Y_i|K_i,L_i,I_i\right]=(\alpha_{_K}+\alpha_{_{KI}}I_i)\frac{1}{K_i} \text{ \& } \frac{\partial}{\partial L_i}E\left[\ln Y_i|K_i,L_i,I_i\right]=(\alpha_{_L}+\alpha_{_{LI}}I_i)\frac{1}{L_i}.$$

Alternatively, it suffices to show that I is not strongly separable from the inputs in $f^{\texttt{CDh}}$, unless the following condition holds:

$$\alpha_{_{LI}}=\alpha_{_{KI}}=0. \tag{1.12}$$

In summary, in the case of a CD production function with interactions as described in (1.9), the definitions of HN, IHN and EHN are not satisfied, unless particular conditions are met.

1.2.3 Estimation methods

The particular structure of models in which a dummy endogenous innovation additively enters the production function (1.7) is typically referred to as the dummy endogenous variable model. This specification, which is standard in the econometric literature focusing on the effect of innovation on productivity (Mohnen and Hall, 2013), can be consistently estimated using, among other methods, the standard IV estimator (IV-2SLS, hereafter; see, e.g., Wooldridge, 2010; Kelejian, 1971; Angrist and Krueger, 2001) and selecting the instruments within the determinants of the innovation function (Musolesi and Huiban, 2010). However, in a model with interactions between the endogenous dummy and the explanatory variables, the adoption of the IV-2SLS is more problematic. The main problem arises because each interaction term IX_r will be also endogenous. Therefore, estimating such a model by standard IV-2SLS would require finding instruments for all the endogenous variables $I, IX_r, r=1,2,\ldots,k$. If $\mathbf{Z}$ is a set of ρ valid instruments for I, then a natural set of instruments for IX_r is $\{Z_jX_r : Z_j\in\mathbf{Z}\}$. This approach results in a total of $(k+1)\rho$ IVs, while it is well known that the estimation by standard IV-2SLS in the presence of many instruments exhibits substantial bias and makes inference inaccurate (see Hansen et al., 2008)

Consequently, we use an alternative IV approach proposed by Wooldridge (2010) (IV-W), which is more efficient than IV-2SLS and has a number of other interesting features. The implementation of IV-2SLS requires the *zero correlation assumption*, i.e., $\mathbf{E}\left(\epsilon\right)=\mathbf{E}(\epsilon\mathbf{X})=\mathbf{E}(\epsilon\mathbf{Z})=0,$ whereas to use IV-W, the error term ϵ should have *zero conditional mean* – $\mathbf{E}\left(\epsilon\mid\mathbf{X},\mathbf{Z}\right)=0$ – which is a stronger exogeneity assumption that ensures that $\mathbf{E}\left(\epsilon\right)=\mathbf{E}(\epsilon\mathbf{X})=\mathbf{E}(\epsilon\mathbf{Z})=0$ but also implies that ε is uncorrelated with any function of $\mathbf{X}$ and $\mathbf{Z}$. Under the *zero conditional mean assumption*, the two-step approach proposed by Wooldridge for a model in which the endogenous dummy enters additively, as in (1.7), consists in estimating $P\left(I=1\mid\mathbf{X},\mathbf{Z}\right)=F\left(\mathbf{X},\mathbf{Z};\gamma\right)$ by (probit) maximum likelihood (ML) and then estimating the structural equation by IV-2SLS using 1, $\mathbf{X}$ and the estimated conditional probability $\hat{P}$ as instruments.

This two-step approach has a number of very interesting features since i) although generated instruments are used, the usual IV-2SLS standard errors and test statistics remain asymptotically valid for the second stage; ii) provided that the homoskedasticity assumption $\mathrm{Var}(\epsilon|\mathbf{X},\mathbf{Z})=\sigma^2$ holds, the IV-2SLS estimator of the second step is asymptotically the most efficient for the class of estimators in which the instruments are functions of $(\mathbf{X},\mathbf{Z})$; and, possibly most important, iii) this estimator possesses an important robustness property since the estimator of the structural equation in the second step is consistent, even if the model for $P(I = 1|\mathbf{X}, \mathbf{Z})$ is not correctly specified, while the requirements that are needed for consistency are much weaker (White, 1982). In other words, the innovation function does not have to be correctly specified to obtain consistent estimates for the parameters in the augmented production function.

To estimate a model with interactions such as (1.9), Wooldridge (2003, 2010) proposes a method (IV-W-H) that is a simple extension of that presented above and provides a solution to the problem of many instruments in the standard IV-2SLS. This is achieved by using the optimal – in terms of efficiency – instruments $P \equiv P(I = 1|\mathbf{X}, \mathbf{Z})$ for I and PX_r for $IX_r, r = 1, 2, \ldots, k$. This approach consists of the following two steps:

a. Estimate $P(I = 1|\mathbf{X}, \mathbf{Z}) = G(\mathbf{X}, \mathbf{Z}; \gamma)$ by ML and obtain the fitted values $\hat{P}$.

b. Estimate the structural equation by IV-2SLS using 1, $\mathbf{X}$, $\hat{P}$ and $\hat{P}\mathbf{X}$ as instruments.

As before, under the *zero conditional mean assumption*, the IV-2SLS of the structural equation is a consistent and asymptotically normal estimator, and again, the binary response model does not need to be correctly specified to achieve consistency. Some additional remarks are in order.

First, note that to estimate (1.9), we use the same parametrization as in Wooldridge (2003, 2010), where the interaction terms are mean centred, i.e., $I\left(\mathbf{X} - \overline{\mathbf{X}}\right)$. Second, note that $\overline{\mathbf{X}}$ is introduced in $\left(\mathbf{X} - \overline{\mathbf{X}}\right)$ as an estimator of $\mathbf{E}(\mathbf{X})$ and this should be accounted for when computing the standard errors. However, according to Wooldridge (2010), this will not have serious consequences, and heteroskedasticity-robust standard errors could still be employed; alternatively, bootstrapped standard errors are a viable alternative. Third, the parameter associated with innovation, $\alpha_{_I}$, measures, under weak assumptions, the average treatment effect (ATE), i.e., $\alpha_{_I} = \alpha_{_I}^{ATE}$ for both the additive (1.7) and the interaction model (1.9). As the interaction terms are mean centred and using the same notation as in Cerulli (2014), we can define the following:

$$\begin{aligned} ATE(\mathbf{X}) &= \mathbf{E}\left(\ln Y \mid \mathbf{X}, \mathbf{Z}, I = 1\right) - \mathbf{E}\left(\ln Y \mid \mathbf{X}, \mathbf{Z}, I = 0\right) \\ &= \alpha_{_I} + \alpha_{_{KI}}(\ln K - \overline{\ln K}) + \alpha_{_{LI}}\left(\ln L - \overline{\ln L}\right). \end{aligned} \tag{1.13}$$

While in the additive model, the innovation effect is constant, specifying (1.9) allows for an heterogeneous effect of innovation across firms, which is a function of the production inputs. This is why, in the heterogeneous case, we will also focus attention on the estimation of the distribution of such an effect. Finally, suppose that the unobservable stochastic part of the model, ϵ, differs between innovative and non-innovative firms, that is, $\epsilon = \epsilon_0 + I(\epsilon_1 - \epsilon_0), \epsilon_1 \neq \epsilon_0$. In such a case, under a fairly weak assumption, the above procedure continues to be consistent. The assumption that is required for consistency is a mean independence assumption $\mathbf{E}[I(\epsilon_1 - \epsilon_0)|\mathbf{X}, \mathbf{Z}] = \mathbf{E}[I(\epsilon_1 - \epsilon_0)]$, which is generally reasonable for continuously distributed responses (see Angrist, 1991; Wooldridge, 2010).

1.3 Data and Results

1.3.1 Data

The data used for the analysis come from the tenth and last Survey on Manufacturing Firms ("Indagine sulle Imprese Manifatturiere") provided by Unicredit-Capitalia, which is complemented with balance sheets sourced from either AIDA (the Italian Balance Sheet Dataset of the Bureau van Dijk) or from the Chambers of Commerce Registry (UNICREDIT, 2008). The same survey, although in different waves, has been widely used in the economics literature on firms' innovation activities (for example Parisi et al., 2006; Hall et al., 2009). Information on the innovation activity of firms was derived from the survey that was conducted in 2007 and posed questions referring to the three-year period 2004-2006, while the variables derived from balance sheets refer to the year 2006. The initial sample comes from a stratified survey: all firms with more than 500 employees are included, while for the firms with fewer than 500 employees, a sample is extracted and stratified according to the information collected from the company registry for the variables size, value added, geographical location and industry. To estimate the econometric model, the main variables we consider are as follows:

1. The natural logarithm of value added (lnY): the measure refers to 2006 and is reported on the balance sheet.

2. The natural logarithm of the capital stock (lnK): the measure refers to 2006, is calculated by summing the value of fixed assets, and is estimated through a perpetual inventory method considering the usual rate of depreciation of 0.05, including investments. Both measures of fixed assets and investments are available from 1998 to 2006, and both are deflated with the respective aggregate price index (derived from ISTAT, the Italian National Statistical Office).

3. The natural logarithm of labour (lnL): number of employees in 2006 reported on the balance sheet.

4. Innovation, I: an innovation dummy taking value 1 if the firm affirmed having introduced at least one product or one process innovation in the previous three years (2004-2006), 0 otherwise.

5. Sectors: the manufacturing firms are classified by sector according to the two-digit ATECO2002 classification, which derives from the NACE Rev.1.1 Eurostat classification.

The aim of our preliminary data analysis is to identify the outliers in the sample, and the econometric sample is obtained by adopting the cleaning procedure detailed below.

First, we drop observations with missing or inconsistent values, resulting in a sample of 3237 firms. Second, outliers are detected using the boxplot rule (Tukey's method) on the variables under investigation. In so doing, another 232 observations (7% of the total) that exceed the boxplot's outer fences are dropped, resulting in a dataset of 3005 firms and a significant reduction in the range of the variables. We also search for outliers with respect to productivity.

Consequently, we define total factor productivity as $TFP = Y/(K^{0.3}L^{0.7})$ and then again apply the boxplot rule to detect another 81 outlying observations. Therefore, the final dataset consists of 2924 firms.

Descriptive statistics of the main inputs and output are presented in Table 1, while the $\mathbf{Z}$ variables are described in Table 4 of Appendix C. Note that the percentage of innovating firms in the final dataset employed in this work is 64%. This value is the same as in Hall et al. (2009), who construct a panel data set starting from different waves of the same survey used in this work. This value is also very close to the percentage of innovating firms obtained using the Italian CIS survey (Hall et al., 2008).

1.3.2 Main results

First, we focus on the choice of the set of the $\mathbf{Z}$ variables. Importantly, as explained in subsection 1.2.3, Wooldridge's approach for both the additive and non-additive specifications (IV-W and IV-W-H) is consistent even if the model for $P(I = 1|\mathbf{X}, \mathbf{Z})$ is not correctly specified. This is an important robustness property.

In Table 2, we report the main results, which are obtained by using Wooldridge's approach and selecting the $\mathbf{Z}$ variables using a backward selection procedure, with a threshold of 0.10 for the p-value. The presentation of this single set of results, leaving the remaining to Appendix C, is chosen because of the extreme stability of the results across the different specifications. Appendix C presents many additional results. These results are obtained using alternative definitions of the vector $\mathbf{Z}$ and also adopting the standard IV-2SLS method, where the instruments are selected to be strong and valid.

Examining the results reported in the first two columns of Table 2, which are obtained estimating the additive specification, reveals their consistency with previous empirical work. While the baseline OLS approach does not provide evidence supporting the positive role of innovation in the production process, with $\widehat{\alpha_I} = 0.0189$ and being non-significant, when we use the IV-W method, we find a positive and significant Hicks-neutral effect of innovation, with $\widehat{\alpha_I} = 0.285$, which is in line with the existing literature (Hall, 2011), where this parameter ranges from approximately 0.2 to approximately 0.3. Moreover, the estimated coefficients of labour and capital are also in line with previous work, with $\widehat{\alpha_K} = 0.179$ and $\widehat{\alpha_L} = 0.741$, and a resulting elasticity of scale equal to 0.92.

We then turn to the estimation of the non-additive specification using the IV-W-H method, which is the main interest of this paper. The estimated average effect of innovation (0.272) is very close to that estimated using the additive model IV-W (0.285). Note that the results presented in Table 2 are all obtained using heteroskedasticity-robust standard errors. As stressed in subsection 1.2.3, when adopting the IV-W-H approach, bootstrapping the standard errors can be a viable solution to account for the fact that mean-centred variables are used for the interaction terms. We also apply a bootstrap with 1000 replications, which does not affect the results. The second important result that emerges is that the estimated parameters associated with the interaction terms, $\widehat{\alpha_{KI}}$ and $\widehat{\alpha_{LI}}$, appear to be statistically significant. This is a crucial result from this paper indicating that innovation does not have a neutral effect on production output. In fact, the production technology of innovative firms differs significantly from that of non-innovative firms, and this difference is produced by two elements: 1) a significant estimated parameter allowing for a shift, $\widehat{\alpha_I}$; 2) a significant change in the slope parameters, which is measured by $\widehat{\alpha_{KI}}$ and $\widehat{\alpha_{LI}}$, thus also affecting the shape of the production function.

Moreover, the additive specification can be tested against the non-additive one using a modified F test (Wooldridge, 1995). Unlike the standard F test, this statistic uses the sum of squared residuals from the second stage of the IV-2SLS regressions in the second step of the IV-W and IV-W-H approaches. The test rejects the additive specification in favour of the non-additive specification at the 5% significance level. In other words, the test rejects the condition (1.12) under which EHN is true. Further, we examine the condition under which HN and IHN hold in the non-additive specification using a Wald-type test; the test rejects, at the 5% level, the null hypothesis that (1.11) holds. Therefore, all definitions of Hicks neutrality are firmly rejected.

By examining the estimated parameters, we can obtain further insights into how technology differs between innovative and non-innovative firms. The estimated elasticity of labour for innovative firms equals 0.94 and is much higher than that for non-innovative firms, which is estimated at 0.43. The opposite holds for capital, with elasticity values equal to 0.33 for non-innovative and to 0.08 for innovative firms. As explained below, we find evidence of a capital-saving innovation. This supports the idea that innovation has an heterogeneous effect on the production process, which depends substantially on the production input with which it interacts. The consequences of the different interaction effects are visible in the shape of the estimated production function (fig.1.1).

The estimated $ATE(\mathbf{X})$ is equal to $0.272+0.510\left(\ln L - \overline{\ln L}\right) - 0.248(\ln K - \overline{\ln K})$, where $\widehat{\alpha_I} = 0.272$ is the estimated ATE, and we can focus our attention on the estimation of the underlying density function using the kernel approach. We use a second-order Gaussian kernel and cross-validation to choose the smoothing parameter. The estimated density is plotted in fig.1.2. As noted above, in our case, the ATE corresponds to the innovation coefficient. The estimated ATE returns an increase in value added of approximately 27% if innovation is introduced. Note that the ATE is the mean value of the $ATE(\mathbf{X})$. The estimated $ATE(\mathbf{X})$ is positive for most of the domain of the inputs: it ranges from -.677 to 1.453 and is positive for approximately 82% of the observations. Innovative firms are less productive than non-innovative firms only for very low values of labour associated with relatively high values of capital. On the contrary, the highest $ATE(\mathbf{X})$ appears for high values of labour associated with low values of capital. These results are directly related to the notion of locally progressive TC, which according to Chambers (1988), is fairly common in practice and has relevant implications. To the best of our knowledge, this is the first paper providing empirical evidence of such a situation.

Next, we discuss the elasticity of scale (see Basu and Fernald, 1997, for a through discussion on estimated returns to scale). While non-innovative firms are characterized by decreasing returns to scale, with an estimated elasticity of scale $\widehat{\alpha_K} + \widehat{\alpha_L} = 0.75$, innovative firms exhibit slightly increasing returns to scale, with an estimated elasticity of scale $\widehat{\alpha_K} + \widehat{\alpha_{KI}} + \widehat{\alpha_L} + \widehat{\alpha_{LI}} = 1.02$. Using Wald tests, the hypothesis of constant returns to scale is not rejected at the 10% significance level for innovative firms, while it is rejected at the 0.1% level for non-innovative firms. Instead, when we estimated a model assuming a common technology (the additive one), the elasticity of scale equals 0.92. In this case, the hypothesis of constant returns to scale is rejected at the 0.1% level. This indicates that when estimating a production function with added innovation, we face a kind of heterogeneity bias since the value 0.92 is obtained mingling two heterogeneous technologies: that of innovative firms characterized by constant returns to scale and that of non-innovative firms with decreasing returns to scale.

A final object that is of great relevance is the MRTS, the analysis of which provides us with information on the nature of technological progress. To obtain insights into such an object, we first use contour plots (in figure 1.3) to represent the estimated production function. Such isolines have a direct economic interpretation as the estimated isoquants. Given the shapes of the isoquants in figure 1.3, we find that the MRTS – the slope of the isoquants – is higher for innovative than for non-innovative firms. The substitution opportunities are reduced for innovative firms. The values of the relative MRTS tell us that to compensate for a 1% change in labour, an innovative firm should change capital by approximately 12%, while a non-innovative firm needs a change in capital of only approximately 1.3%.

To obtain complementary information, we also calculate the MRTS using (1.10) and then focus on the estimation of its density function (figure 1.4). We specifically estimate the conditional density of the MRTS, conditional to innovation. With innovation being a discrete variable, we adopt the approach of (Hall et al., 2004), which uses generalized product kernels to deal with mixed data and cross-validation to choose the smoothing parameters. Interestingly, the smoothing parameter associated with innovation goes to zero. This not only suggests that innovation is relevant – meaning that the two densities are not the same –but also indicates that the generalized estimator collapses to the standard frequency estimator. This result goes further in the direction of fully heterogeneous technology.

1.4 Conclusion

In standard econometric models, innovation additively enters the production function, imposing a strict condition of Hicks neutrality. We depart from this restrictive framework by considering a production function that allows a heterogeneous effect of innovation on the production process and relaxes the Hicks neutrality assumption. We derive conditions, for common parametric specifications, under which neutrality holds and that are easily testable through common Wald-type tests. Further, taking into consideration the endogenous character of innovation, we estimate the model by adopting an instrumental variables approach that addresses the problem of many instruments when estimating a model with interactions. The econometric analysis rejects Hicks neutrality and indicates that innovation produces a non-neutral effect on the production process, which is obtained because of the joint presence of a shift in the production technology and a change in the slope of the isoquants. The latter indicates that innovative firms are *capital saving* compared with non-innovative firms. Moreover, as a consequence of the joint effect described above, a *locally progressive* technical change is also observed, because, while for most of the domain of the inputs, innovation has a positive effect, for a small part of it, innovative firms are less productive than non-innovative firms. Overall, our results indicate fully heterogeneous production technologies when comparing innovative to non-innovative firms. These findings have interesting policy implications, as they highlight the complex fashion in which innovation affects the production process. To the best of our knowledge, this is the first study supporting such evidence.

Further studies may consider extending the analysis to other countries or to specific sectors. Methodological extensions may be achieved by considering a panel data framework to account for the time dimension or by adopting nonparametric methods to highlight the potential presence of localized technical change.

Table 1: Descriptive statistics

	lnVA			lnK			lnL		
	All firms	I=0	I=1	All firms	I=0	I=1	All firms	I=0	I=1
minimum	2.52	2.52	2.75	3.65	4.11	3.65	2.30	2.30	2.30
median	7.43	7.25	7.53	7.71	7.50	7.81	3.50	3.30	3.58
mean	7.46	7.32	7.53	7.60	7.46	7.68	3.52	3.40	3.59
maximum	9.32	9.29	9.33	9.78	9.77	9.78	5.50	5.48	5.50
st. deviation	0.81	0.82	0.80	1.11	1.13	1.09	0.73	0.73	0.72

Table 2 - Main results: lnVA as dependent variable

	OLS	*IV-W*	*IV-W-H*
lnL	.750***	.741***	.428***
	(.0166)	(.0218)	(.1044)
lnK	.185***	.179***	.327***
	(.0108)	(.0135)	(.0639)
$Innovation$	.0189	0.285***	0.272***
	(.0159)	(.0774)	(.0806)
$IlnL$			0.510***
			(.1545)
$IlnK$			-0.248**
			(.0970)
intercept	3.228***	2.870***	3.120***
	(.0578)	(.1203)	(.2618)
N	2924	2239	2239
R^2	.758	.727	.704
adj. R^2	.755	.724	.701
Endogeneity test		12.177***	23.341***
Montiel-Pflueger F		168.204	
[critical value]		[37.418]	
ATE	.019	.285	.272
Elasticity of labour	.75	.74	
Elasticity of labour $(I = 0)$			.43
Elasticity of labour $(I = 1)$			.94
Elasticity of capital	.19	.18	
Elasticity of capital $(I = 0)$			.33
Elasticity of capital $(I = 1)$			.08
Elasticity of scale	.94	.92	
Elasticity of scale $(I = 0)$			.75
Elasticity of scale $(I = 1)$			1.02
Relative MRTS	-4.12	-4.14	
Relative MRTS $(I = 0)$			-1.31
Relative MRTS $(I = 1)$			-11.78

Montiel-Pflueger test for weak instruments, null hypothesis that instruments are weak. $\tau = 5\%$, confidence level $\alpha = 5\%$, test not applicable in regressions with one endogenous variable.
Sectors not presented in the table.
Wooldridge's (1995) robust score test for overidentifying restrictions not applicable.
Endogeneity test according to Wooldridge's (1995) score test.
Robust standard errors, Huber/White/sandwich estimator
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, robust standard errors in parentheses

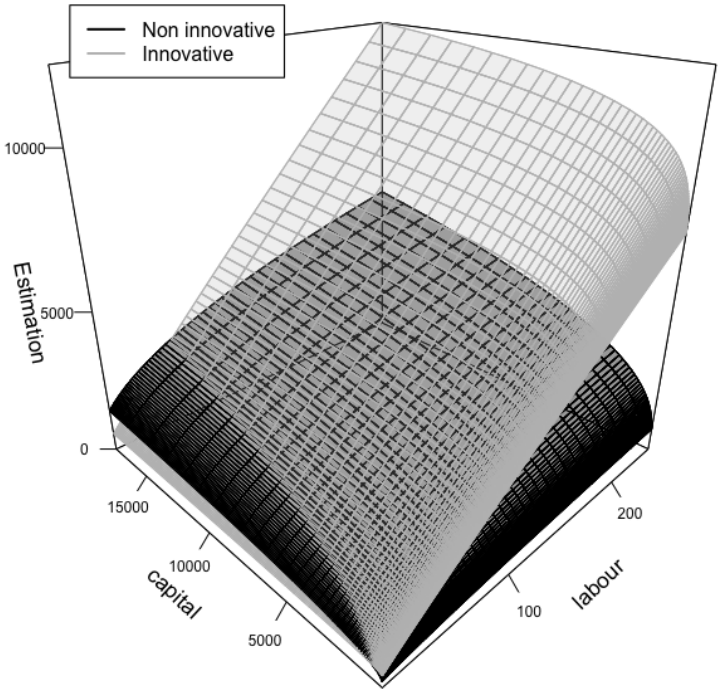


Figure 1.1: Estimated production function (IV-W-H method)

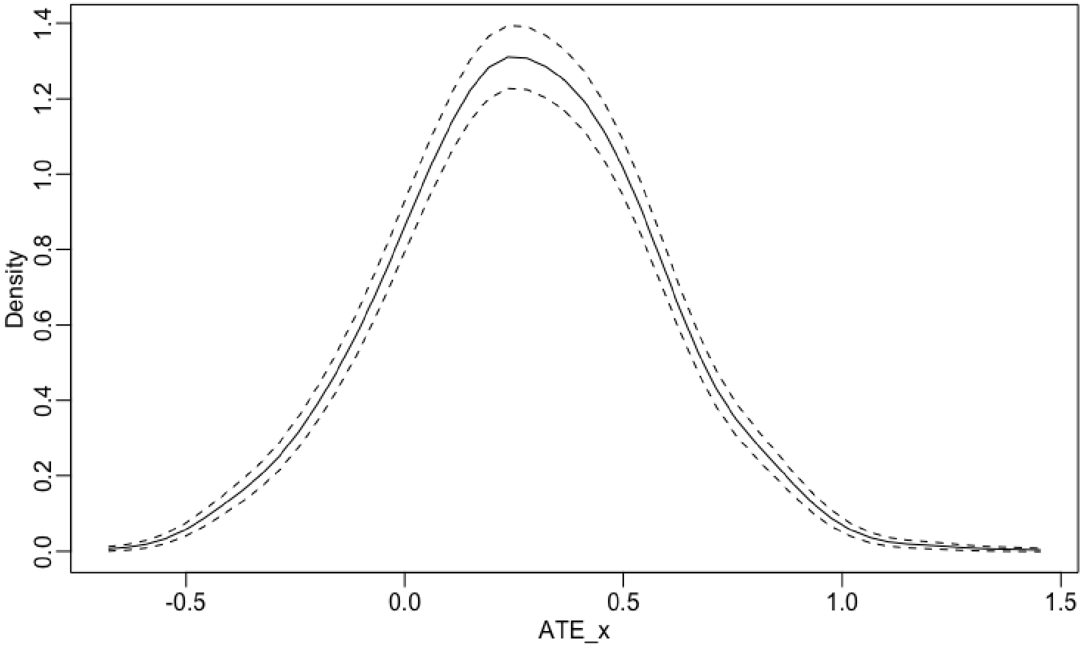


Figure 1.2: Distribution of the ATE(**X**), estimated using gaussian kernels. The dashed lines represent 95% bootstrapped confidence bands. The bandwidth (0.07) is selected by least squares cross validation. The mean value of ATE (**X**) corresponds to ATE and is equal to 0.272.

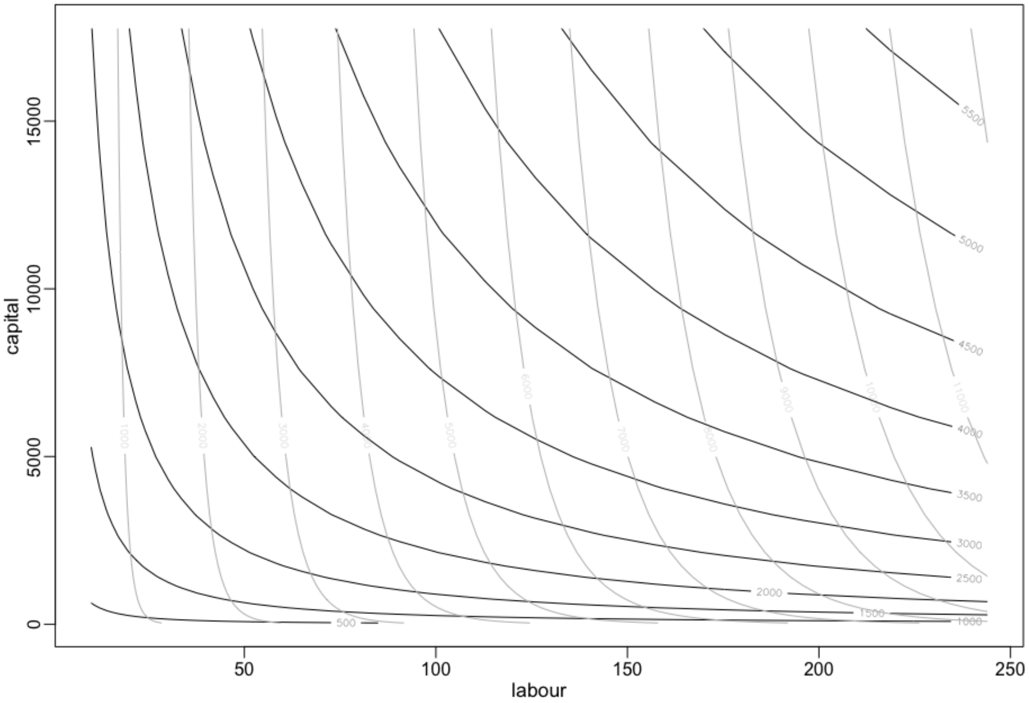


Figure 1.3: Contours of the production function estimated with IV-W-H. The grey lines correspond to the contours of the innovative firms, while the black lines correspond to those of non-innovative firms.

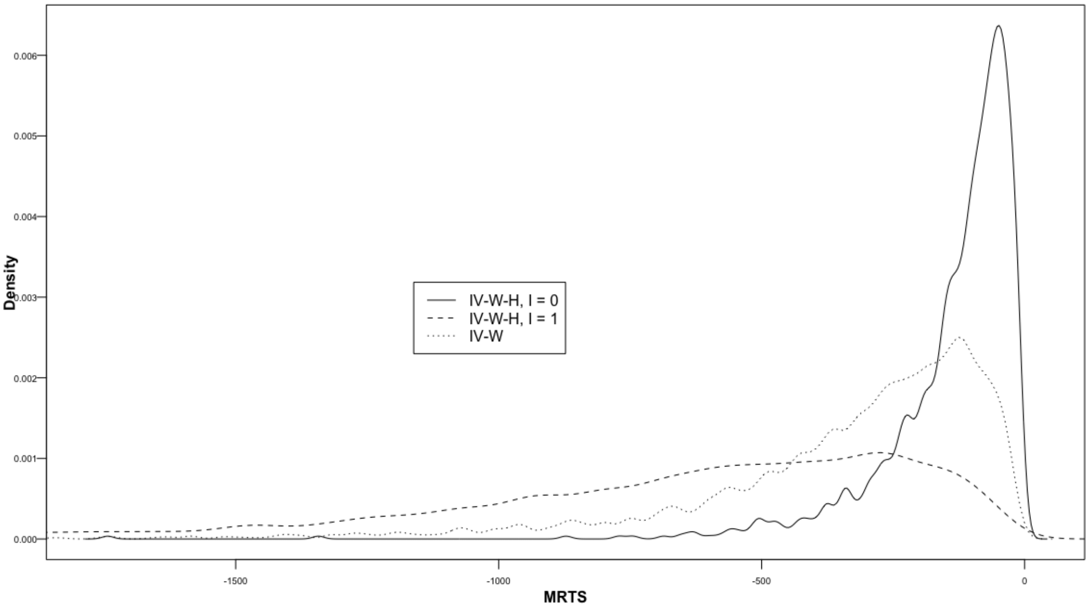


Figure 1.4: Density distribution of the MRTS. The dotted line corresponds to the MRTS estimated by the IV-W approach. The solid line and the dashed line correspond to the MRTS of non-innovative and innovative firms in the IV-W-H case.

Supplementary Material

Appendix A: Relationships among HN, IHN and EHN

The definitions of HN, IHN and EHN are described in subsection 1.2.1. In this appendix we provide additional insights into the work of Blackorby et al. (1976, sec.4). Specifically, we provide an illustrative, simplified presentation of the conditions under which the above definitions are equivalent, and we highlight their relationships using the following figure.

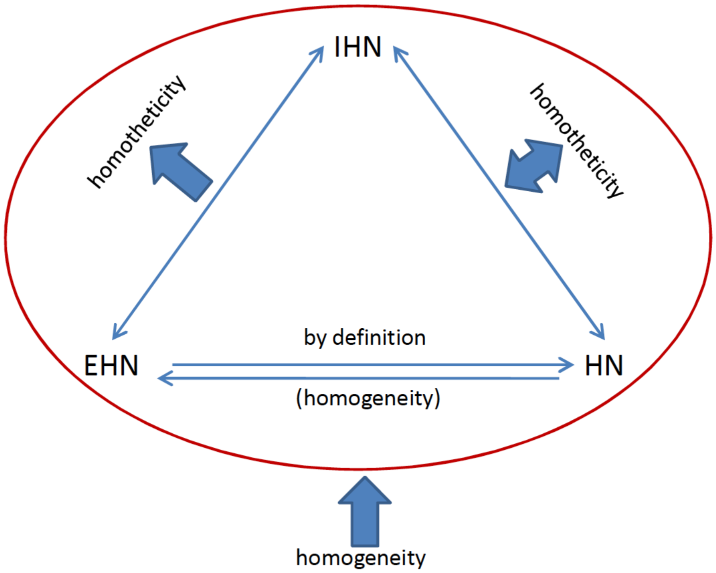


Figure 1.5: Illustration of the relationships among the different definitions of Hicks neutrality; double arrows symbol implies equivalence.

The relationship between EHN and HN

Consider the production function $Y = f(\mathbf{X}, I)$, where $\mathbf{X}$ is a vector of inputs and I is a variable that measures innovation. If I is EHN, then I is strongly separable from $\mathbf{X}$. Therefore, it is also weakly separable from $\mathbf{X}$. For definitions of weak and strong separability, we refer to Chambers (1988, p. 42-46). Morimoto (1974) proves that weak separability of I from $\mathbf{X}$ is equivalent to I being HN. Therefore, if I is EHN, it is also HN. In general, the converse does not hold, that is, I being HN does not necessarily imply that it is also EHN. Uzawa and Watanabe (1961) prove that in the case of an input-homogeneous production function f, I is HN if and only if f is multiplicatively decomposable into a function of I and a function of $\mathbf{X}$. Therefore, if the production function is input homogeneous, then HN implies EHN.

The relationship between HN and IHN

Moreover, HN and IHN are generally not equivalent terms. Blackorby et al. (1976, fig.I) describe technical progress that is HN but not IHN, and vice versa. They prove that HN and IHN are equivalent if and only if the production function is input homothetic. An obvious consequence of this theorem is that in a homothetic production function, if I is not HN (IHN), then it is also not IHN (HN). Moreover, if I is HN (IHN) but not IHN (HN), then f is not input homothetic.

The relationship between IHN and EHN

In general, IHN and EHN are also not equivalent. Blackorby et al. (1976) prove that if I is IHN and EHN, then the production function is input homothetic. By negation, it can be shown that if f is homothetic and I is EHN (IHN), then I should also be IHN (EHN). For example, in (1.14), we describe an empirical specification of a TL function with an added I. In this case, I is EHN but, generally, not IHN, unless certain homotheticity conditions hold.

Finally, note that according to the above, if the production function is input homogeneous, then HN, IHN and EHN are equivalent. Indeed, baring that homogeneity implies homotheticity, if f is homogeneous and I follows one of these definitions of neutrality, then I should also follow the other two (see also fig.1.5).

Appendix B: Hicks neutrality and the translog specification

In this study, we have also estimated by IV-2SLS, IV-W and IV-W-H the TL specifications that are presented in (1.14) and (1.17) below. Nevertheless, the results of the F-type tests (see Wooldridge, 1995) that compare the CD and the TL models do not reject the CD specification and the estimations of the TL model provide poor results in terms of magnitude and significance level for most of the coefficients.[2] For these reasons, the main interest in this paper is on the CD specification. However, as the TL specification is the most popular Diewert-flexible form and a widely used direct generalization of the CD specification, we extend the analysis in section 1.2.2 to account for the TL case. Particularly, in this appendix, we provide insights into HN within a TL framework and provide easily testable conditions for the presence of HN in TL specifications. Therefore, the following analysis may be useful in empirical studies that adopt a TL framework while relaxing HN.

Under the assumption that innovation enters additively, the econometric specification of a TL production function with conventional inputs K and L is:

$$\begin{aligned}\ln Y_i &= \ln f^{\text{TL}}\left(K_i, L_i, I_i, \epsilon_i\right) = \\ &= \alpha + \alpha_{_I} I_i + \alpha_{_K} \ln K_i + \alpha_{_L} \ln L_i + \alpha_{_{KL}} \ln K_i \ln L_i + \alpha_{_{K2}} \frac{1}{2} \ln K_i^2 + \alpha_{_{L2}} \frac{1}{2} \ln L_i^2 + \epsilon_i,\end{aligned} \tag{1.14}$$

and the MRTS between L and K is described by:

$$MRTS_i^{\text{TL}} = -\frac{\alpha_{_L} + \alpha_{_{KL}} \ln K_i + \alpha_{_{L2}} \ln L_i}{\alpha_{_K} + \alpha_{_{KL}} \ln L_i + \alpha_{_{K2}} \ln K_i} \frac{K_i}{L_i}, \tag{1.15}$$

according to which I is HN but not IHN, because the MRTS is not constant along a ray from the origin, unless $\alpha_{_{KL}} + \alpha_{_{L2}} = 0$ and $\alpha_{_{KL}} + \alpha_{_{K2}} = 0$. Moreover, EHN holds because f^{TL} is

[2] The results of the estimations of the TL specifications are available upon request.

multiplicatively decomposable into a function of I and a function of the inputs. Alternatively, it can be shown that:

$$\frac{\partial}{\partial K_i} E\left[\ln Y_i | K_i, L_i, I_i\right] = (\alpha_{K} + \alpha_{KL} \ln L_i + \alpha_{K2} \ln K_i)\frac{1}{K_i} \tag{1.16a}$$

$$\frac{\partial}{\partial L_i} E\left[\ln Y_i | K_i, L_i, I_i\right] = (\alpha_{L} + \alpha_{KL} \ln K_i + \alpha_{L2} \ln L_i)\frac{1}{L_i}. \tag{1.16b}$$

Note that while in the case of a CD production function with added innovation, all the above definitions of neutrality are imposed, the TL specification with added innovation imposes HN and EHN but not IHN.

Further, a TL function with interaction terms between innovation and the explanatory variables is described by:

$$\begin{aligned}\ln Y_i = \alpha + \alpha_{I} I_i + \alpha_{K} \ln K_i + \alpha_{KI} I_i \ln K_i + \alpha_{L} \ln L_i + \alpha_{LI} I_i \ln L_i + \alpha_{KL} \ln K_i \ln L_i +\\ + \alpha_{KLI} I_i \ln K_i \ln L_i + \alpha_{K2}\frac{1}{2}\ln K_i^2 + \alpha_{K2I}\frac{1}{2} I_i \ln K_i^2 + \alpha_{L2}\frac{1}{2}\ln L_i^2 + \alpha_{L2I}\frac{1}{2} I_i \ln L_i^2 + \epsilon_i.\end{aligned} \tag{1.17}$$

In this case, the MRTS between L and K is given by the following equation:

$$MRTS_i^{\mathtt{TLh}} = -\frac{\alpha_{L} + \alpha_{LI} I_i + (\alpha_{KL} + \alpha_{KLI} I_i)\ln K_i + (\alpha_{L2} + \alpha_{L2I} I_i)\ln L_i}{\alpha_{K} + \alpha_{KI} I_i + (\alpha_{KL} + \alpha_{KLI} I_i)\ln L_i + (\alpha_{K2} + \alpha_{K2I} I_i)\ln K_i}\frac{K_i}{L_i}. \tag{1.18}$$

Generally, the definitions of HN and IHN are not satisfied because the MRTS in (1.18) is dependent on I. By contradiction, it is proven that the definition of EHN does not hold either, unless $\alpha_{LI} = \alpha_{KI} = \alpha_{KLI} = \alpha_{K2I} = \alpha_{L2I} = 0$. Finally, assuming non-zero coefficients, HN holds only if at least one of the following conditions is satisfied:

$$\frac{\alpha_{K}}{\alpha_{L}} = \frac{\alpha_{KL}}{\alpha_{L2}} = \frac{\alpha_{K2}}{\alpha_{KL}} = \frac{\alpha_{KI}}{\alpha_{LI}} = \frac{\alpha_{K2I}}{\alpha_{KLI}} = \frac{\alpha_{KLI}}{\alpha_{L2I}}, \text{ or} \tag{1.19a}$$

$$\frac{\alpha_{LI}}{\alpha_{L}} = \frac{\alpha_{KI}}{\alpha_{K}} = \frac{\alpha_{KLI}}{\alpha_{KL}} = \frac{\alpha_{K2I}}{\alpha_{K2}} = \frac{\alpha_{L2I}}{\alpha_{L2}}. \tag{1.19b}$$

Under condition (1.19a), I in (1.17) is also IHN because (1.8) and (1.18) coincide. Under condition (1.19b), (1.18) is identical to (1.15). In this case, I is IHN if, in addition to (1.19b), it also holds that $\alpha_{KL} + \alpha_{L2} = 0$ and $\alpha_{KL} + \alpha_{K2} = 0$.

In summary, in the case of a TL production function with interactions, the definitions of HN, IHN and EHN are not satisfied, unless particular conditions are met. These conditions are summarized in Table 3.

Table 3 - Conditions for neutrality

Type	TL additive	TL non-additive
HN	–	$\frac{\alpha_K}{\alpha_L} = \frac{\alpha_{KL}}{\alpha_{L2}} = \frac{\alpha_{K2}}{\alpha_{KL}} = \frac{\alpha_{KI}}{\alpha_{LI}} = \frac{\alpha_{K2I}}{\alpha_{KLI}} = \frac{\alpha_{KLI}}{\alpha_{L2I}}$ or $\frac{\alpha_{LI}}{\alpha_L} = \frac{\alpha_{KI}}{\alpha_K} = \frac{\alpha_{KLI}}{\alpha_{KL}} = \frac{\alpha_{K2I}}{\alpha_{K2}} = \frac{\alpha_{L2I}}{\alpha_{L2}}$
IHN	$\alpha_{KL} + \alpha_{L2} = 0$ and $\alpha_{KL} + \alpha_{K2} = 0$	$\frac{\alpha_K}{\alpha_L} = \frac{\alpha_{KL}}{\alpha_{L2}} = \frac{\alpha_{K2}}{\alpha_{KL}} = \frac{\alpha_{KI}}{\alpha_{LI}} = \frac{\alpha_{K2I}}{\alpha_{KLI}} = \frac{\alpha_{KLI}}{\alpha_{L2I}}$ or $\frac{\alpha_{LI}}{\alpha_L} = \frac{\alpha_{KI}}{\alpha_K} = \frac{\alpha_{KLI}}{\alpha_{KL}} = \frac{\alpha_{K2I}}{\alpha_{K2}} = \frac{\alpha_{L2I}}{\alpha_{L2}}$, $\alpha_{KL} + \alpha_{L2} = 0,$ $\alpha_{KL} + \alpha_{K2} = 0$
EHN	–	$\alpha_{LI} = \alpha_{KI} = \alpha_{KLI} = \alpha_{K2I} = \alpha_{L2I} = 0$

Appendix C: Robustness checks

In this appendix, we provide robustness checks of the main results presented in subsection 1.3.2. A natural approach is first to estimate the additive model by IV-2SLS and focus attention on the choice of the IVs. We apply the IV selection method described below and arrive at different choices of sets $\mathbf{Z}$ of strong and valid instruments. Then, we apply Wooldridge's approach to perform IV-W and IV-W-H estimations using the $\mathbf{Z}$ sets selected previously. As presented below, the results are very robust in all estimations.

We first select an initial set of potential IVs according to the literature on KPF (see, e.g., Hall et al., 2009; Musolesi and Huiban, 2010). This set is given in the table below.

Table 4 - Initial set of IVs

Category	Variables	Description	Type
Firm Characteristics	NW, NE, C, S, Age, Group, Consortium	Firm's characteristics including geographical location; age and group or consortium membership	binary (0,1)
Human capital	RD_Personnnel	Percentage of employees in Research and Development activities	numeric
Objectives of investment	BetterProd, MoreProd, NewProd, Env, CostRed, Advert, SellNet, SellAss	Objectives including ameliorating the product, produce more, introduce a new one, reduce the environmental impact, reduce costs, to increase the selling network or to ameliorate it, respectively	binary (0,1)
Market penetration	MarkPenEU15, MarkPenEU2004, MarkPenRussia, MarkPenOtherEU, MarkPenAfrica, MarkPenAsia, MarkPenCina, MarkPenUSMex, MarkPenSouthAm, MarkPenOce	Market penetration in different world regions, including EU member states, Africa, Asia, China, U.S., Canada, Mexico, South America and Oceania	binary (0,1)
Commercial agreements	CommAgrEU15, CommAgrEU2004, CommAgrRussia, CommAgrOtherEU, CommAgrAfrica, CommAgrAsia, CommAgrCina, CommAgrUSMex, CommAgrSouthAm, CommAgrOce	Commercial agreements in world regions, as mentioned above	binary (0,1)
Patent acquisition	PatBuyEU15, PatBuyEU2004, PatBuyRussia, PatBuyOtherEU, PatBuyAfrica, PatBuyAsia, PatBuyCina, PatBuyUSMex, PatBuySouthAm, PatBuyOce	Location of the aforementioned world regions where the firm acquired patents	binary (0,1)
Production overseas	ProdAbroadEU15, ProdAbroadEU2004, ProdAbroadRussia, ProdAbroadOtherEU, ProdAbroadAfrica, ProdAbroadAsia, ProdAbroadCina, ProdAbroadUSMex, ProdAbroadSouthAm, ProdAbroadOce	Production located in the aforementioned world regions	binary (0,1)
Competitiveness	LowCompet, HighCompet, SmallProdScale	Perceived level of competitiveness and scale of production compared to competitors	binary (0,1)
Financial specs	ListedComp, FinanIncent	Listed company or receiving financial incentives	binary (0,1)

As highlighted in Bound et al. (1995), a major pitfall that results in inconsistency and large finite sample bias exists when selecting instruments that are weakly correlated with the endogenous variable. To avoid the presence of weak instruments and ensure the validity of the IVs, we follow a two-step procedure.

First, we regress innovation on $\mathbf{X}$ and the above set and adopt a backward selection algorithm to choose an initial set of potential IVs that could be strongly correlated with innovation. The sets corresponding to a 10% and a 5% threshold are presented in the table below.

Table 5 - IV sets

set	instruments
10% set	FinanIncent, BetterProd, MarkPenEU15, C, EUCompet, NewProd, MarkPenEU2004, ProdAbroadEU15, Age, CommAgrAfrica, RD_Personnnel, MoreProd, HighCompet
5% set	FinanIncent, BetterProd, MarkPenEU15, C, EUCompet, NewProd, HighCompet, Age
IV1	BetterProd, MarkPenEU15, EUCompet
IV2	FinanIncent, BetterProd, MarkPenEU15, C, EUCompet
IV3	FinanIncent, BetterProd, MarkPenEU15, EUCompet, NewProd

Generally, the bias of the IV-2SLS estimator increases as the correlation between the IVs and the endogenous variable decreases and as the number of instruments increases. For this reason, in a second step, we estimate by IV-2SLS the additive specification using all the possible combinations of IVs from the 10% set. Since heteroskedasticity-robust standard errors are considered, for each specification, we apply the robust score tests by Wooldridge (1995) to test endogeneity and overidentifying restrictions. We also use the Montiel-Pflueger test to detect the presence of weak instruments (Montiel Olea and Pflueger, 2013). Unlike traditional tests, the Montiel-Pflueger test also accounts for heteroskedastic, serially correlated errors.

The above post-estimation tests indicate IV sets of strong and valid instruments. According to the Montiel-Pflueger test, 42 combinations provide strong IVs. We also find that for 54 sets, the robust score test cannot reject the validity of the IVs. We finally select 13 sets of valid and strong instruments for which the exogeneity test is not rejected. These combinations provide very robust results; both the estimated coefficients and the significance levels are stable across choices of sets. For reasons of brevity, in Table 6, we present the IV-2SLS, IV-W and IV-W-H estimations using these three sets[3], while Table 7 shows the averages of the IV-2SLS, IV-W and IV-W-H estimations for the CD specifications using the above 13 sets.

The first three columns of Table 6 present the IV-2SLS results of the additive specification. The estimated effect of innovation on productivity is between 0.24 and 0.31 and is significant at 0.01 level. In columns 4 to 6, we present the results of the IV-W estimations. The estimated effects of labour, capital and innovation are similar to those from the IV-2SLS estimations, in terms of both the magnitude of the coefficients and the confidence levels. The estimated innovation parameter is between 0.23 and 0.28 and significant at the 0.01 level. Finally, in the last three columns, we also present the results of the IV-W-H estimation. The estimated innovation parameter is significant at the 0.05 level, ranges between 0.22 and 0.31 and is similar to the IV-2SLS and IV-W estimates. An exhaustive presentation of the IV-W-H results and the comparison to the other approaches is given in subsection 1.3.2.

[3] The remaining sets and the results they provide are available upon request.

Table 6 - Cobb Douglas estimations

	IV-2SLS			*IV-W*			*IV-W-H*		
	IV1	IV2	IV3	IV1	IV2	IV3	IV1	IV2	IV3
lnL	.735***	.737***	.737***	.737***	.738***	.737***	.457***	.471***	.482***
	(.0197)	(.0200)	(.0200)	(.0196)	(.0200)	(.0200)	(.1002)	(.0939)	(.0900)
lnK	.179***	.181***	.181***	.179***	.181***	.181***	.294***	.294***	.300***
	(.0125)	(.0126)	(.0126)	(.0123)	(.0125)	(.0125)	(.0646)	(.0601)	(.0595)
$Innovation$	.312***	.247***	.242***	.283**	.233***	.229**	.308***	.221**	.219**
	(.1097)	(.0862)	(.0890)	(.1102)	(.0862)	(.0903)	(.117)	(.0903)	(.0935)
$IlnL$							.451***	.438***	.420***
							(.1536)	(.1430)	(.1372)
$IlnK$							-.193*	-.189**	-.198**
							(.1020)	(.0933)	(.0923)
intercept	2.850***	2.943***	2.949***	2.885***	2.960***	2.964***	3.256***	3.270***	3.189***
	(.1481)	(.1223)	(.1246)	(.1489)	(.1230)	(.1265)	(.2969)	(.2535)	(.2586)
N	2635	2474	2474	2635	2474	2474	2635	2474	2474
R^2	.725	.736	.737	.731	.738	.738	.710	0.723	0.724
adj. R^2	.723	.733	.734	.728	.735	.736	.707	.720	.721
Robust score test	4.521	5.573	6.492						
Endogeneity test	7.685***	6.878***	6.004**	6.094**	5.974**	5.182**	16.044***	16.187***	15.904***
Montiel-Pflueger F	24.814	21.055	20.787	74.160	109.329	104.643			
[critical value]	[15.760]	[20.981]	[20.751]	[37.418]	[37.418]	[37.418]			
	First stage regression			Probit					
FinanIncent		.110***	.109***		.333***	.331***			
		(.0224)	(.0224)		(.0698)	(.0699)			
MarkPenEU15	.135***	.126***	.123***	.431***	.422***	.414***			
	(.0268)	(.0275)	(.0275)	(.0943)	(.0994)	(.0994)			
BetterProd	.120***	.106***	.115***	.348***	.310***	.333***			
	(.0196)	(.0203)	(.0207)	(.0585)	(.0606)	(.0612)			
EUCompet	.058***	.064***	.062**	.176**	.191***	.182**			
	(.0233)	(.0239)	(.0240)	(.0699)	(.0729)	(.0727)			
C		.075***			.206***				
		(.0259)			(.0739)				
NewProd			.099**			.280**			
			(.0405)			(.1224)			
1st-stage partial R^2	.0241	.0369	.0358	.0248	.0382	.0362			
1st-stage robust F	25.47***	22.73***	22.61***	74.16***	109.33***	104.64***			
ATE	.312	.247	.242	.283	.233	.229	.308	.221	.219
Elasticity of labour	.74	.74	.74	.74	.74	.74			
Elasticity of labour $(I = 0)$							.46	.47	.48
Elasticity of labour $(I = 1)$							.91	.91	.90
Elasticity of capital	.18	.18	.18	.18	.18	.18			
Elasticity of capital $(I = 0)$							.29	.29	.30
Elasticity of capital $(I = 1)$							.10	.11	.10
Elasticity of scale	.91	.92	.92	.92	.92	.92			
Elasticity of scale $(I = 0)$							.75	.77	.78
Elasticity of scale $(I = 1)$							1.00	1.01	1.00
Relative MRTS	-4.11	-4.07	-4.07	-4.12	-4.08	-4.07			
Relative MRTS $(I = 0)$							-1.55	-1.60	-1.61
Relative MRTS $(I = 1)$							-8.99	-8.66	-8.84

Montiel-Pflueger test for weak instruments, null hypothesis that the instruments are weak, $\tau = 5\%$, confidence level $\alpha = 5\%$
Sectors not presented in the table. Only the instruments are presented in the first-stage regression of the IV-2SLS and in the probit of the IV-W.
Results of the probit regression in the IV-W-H estimation are same as in the IV-W case.
Wooldridge's (1995) robust score test for overidentifying restrictions. Endogeneity test according to Wooldridge's (1995) score test.
Robust standard errors, Huber/White/sandwich estimator
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, robust standard errors in parentheses

Regarding the IV-W and IV-W-H estimations, the results presented in Table 6 are similar, in both values and significance levels, to the respective results in Table 2. The robustness of the results is clearly shown in Table 7, where we present the averages of the estimated parameters that are obtained by the estimations that use the 13 sets mentioned above. The standard deviations of the parameters are also presented in parentheses. The results show that the estimated coefficients are stable across the different IV sets, for all estimation approaches.

In summary, the above sensitivity analysis shows that the results on the estimated ATE are very stable across the different estimation methods: IV-2SLS, IV-W and IV-W-H. Moreover, the results are very robust across different sets of $\mathbf{Z}$ and, further, strongly support the results presented in the main text.

Table 7 - Cobb Douglas estimation averages

	IV-2SLS	*IV-W*	*IV-W-H*
lnL	.737	.738	.440
	(.002)	(.002)	(.054)
lnK	.179	.180	.310
	(.002)	(.002)	(.027)
$Innovation$	.285	.266	.256
	(.056)	(.055)	(.053)
$IlnL$			.489
			(.090)
$IlnK$			-.218
			(.048)
intercept	3.176	3.180	3.233
	(.021)	(.020)	(.073)
ATE	.285	.266	.256
Elasticity of labour	.74	.74	
Elasticity of labour $(I = 0)$			.44
Elasticity of labour $(I = 1)$			.93
Elasticity of capital	.18	.18	
Elasticity of capital $(I = 0)$			.31
Elasticity of capital $(I = 1)$			.09
Elasticity of scale	.92	.92	
Elasticity of scale $(I = 0)$			.75
Elasticity of scale $(I = 1)$			1.02
Relative MRTS	-4.12	-4.11	
Relative MRTS $(I = 0)$			-1.44
Relative MRTS $(I = 1)$			-10.85

Average values of the estimated parameters, ATEs, elasticities and rel. MRTS. Standard deviations of the estimated parameters in parentheses.
Use of 13 IV sets of strong and valid instruments, for which the exogeneity test is rejected at 0.05.
Sectors not presented in the table.

Chapter 2

Innovation and productivity: new insights from nonparametric instrumental regression

joint with Antonio Musolesi

Abstract

We exploit recent advances on generalized kernel instrumental regression to revisit the relationship between innovation and productivity. This allows to highlight the possible presence of a localized effect of an endogenous innovation variable and, thanks to smoothing discrete variables, also to account for fully heterogeneous technologies across sectors. Such issues are extremely relevant from both a theoretical and a policy oriented perspective but cannot be addressed by adopting common parametric approaches. We also address the issue of the predictive performances of this nonparametric estimator when compared to some parametric alternatives. The results i) indicate that the proposed nonparametric estimator performs better than parametric ones and ii) reveal some relevant patterns that can only be detected using the nonparametric estimator.

Keywords: Nonparametric instrumental regression; localized and biased technical change; Innovation; Productivity;

JEL classification: C14; D24; O33;

2.1 Introduction

Though Solow (1957) describes technical change (TC) *"as a shorthand expression for any kind of shift in the production function"*, in the production theory, TC is typically considered to raise output at all factor proportions (e.g., see Stewart, 1978). Its effect is usually represented by a general shift of the production function outwards, or equivalently, by a uniform shrinking of the unit isoquant towards the origin (Lapan and Bardhan, 1973). Therefore, it is assumed that TC affects productivity globally, that is, in the entire domain of the production function, rather than being localized in specific techniques (David, 1975).

A specific type of global shift refers to the notion of Hicks neutrality (Hicks, 1932), which is often imposed both in theoretical models and in empirical applications. In its original formulation, Hicks neutrality (HN) requires that the marginal rate of technical substitution of each pair of inputs is independent of TC. Blackorby et al. (1976) extend the work by Hicks by introducing two other non-equivalent definitions, i.e. implicit and extended Hicks neutrality. Acemoglu (2015) highlights that in the language of modern growth theory, technological progress is neutral - in the simplest form, Hicks neutral - creating the same proportional gain in output regardless of factor proportions.

This orthodoxy that technical progress would increase productivity at all input levels was first questioned by Atkinson and Stiglitz (1969). They suggest a formalization of the way by which TC can be localized (see also Acemoglu, 2008, ch.18) and describe two contexts associated with *learning by doing* and *research activity*, where increases in technical knowledge result in localized technical change (LTC). In the same vein, Salter (1966) mentions that although, potentially, there exists a large range of techniques of varying investment and labour intensity, only the immediately profitable ones are actually developed, resulting in the localization of the TC to the specific ones. In general, irrespective of its source, if TC does not refer to all techniques, it will not result in a global shift of the production function, but it will be rather localized to specific factor proportions. Although Atkinson and Stiglitz's (1969) view draw attention to several theoretical works (e.g., see Lapan and Bardhan, 1973; Stewart, 1978) and despite the large recognition of their original idea, *"the orthodoxy that Atkinson and Stiglitz were criticizing is still fairly influential"* (Acemoglu, 2015, p.445).

At an empirical level, the analysis of the effect of TC on the production process is often considered within the framework of the so-called knowledge production function (KPF) that was proposed in the seminal works by Pakes and Griliches (1984) and Griliches (1998). Crépon et al. (1998) introduce a multiple-equation econometric model, similar to the structure of the KPF, where innovation activity is the main source of technical and knowledge improvements and use appropriate estimation methods to consider the potential endogeneity of some of the explanatory variables and the particular nature of the dependent variables. The availability of survey data such as the ones in Community Innovation Surveys has allowed the use of a direct and binary measure of the innovative output. Innovation enters a production function framework as an endogenous dummy variable that accounts for TC (see, e.g., Mairesse and Mohnen, 2010). Noteworthy, in the above literature, innovation typically enters additively into the production function. This structure imposes a strict shape restriction and specifically implies

HN (Antonioli et al., 2018).

The present study contributes to the literature on econometrics of productivity and innovation by providing an empirical analysis of a fully nonparametric production function model. In contrast to parametric models, which usually employ assumptions that entail a much greater level of specificity, such as neutrality and specific functional form, the adopted model only assumes that the regression curve belongs to some infinite dimensional collection of functions, such as the differentiable ones.

We exploit recent advances on generalized kernel instrumental regression (see, e.g. Horowitz, 2011; Darolles et al., 2011). This allows to highlight the possible presence of a localized effect of an endogenous innovation variable and, thanks to smoothing categorical variables (Racine and Li, 2004; Hall et al., 2007), also to account for fully heterogeneous technologies across sectors. While it is well known that the production technology greatly differs across sectors, due to the existence of small sized sectors, common econometric specifications introduce the sectors' variable into the model additively, therefore assuming that the effect of sectors to productivity is described by just a global shift of the production function. Since the kernel estimator can be written as a convex combination of a frequency and a pooled estimator (Kiefer and Racine, 2009, 2017), with the smoothing parameter determining the balance between these two extremities, our estimation approach provides an insightful measure of the level of heterogeneity across sectors and between innovative and non-innovative firms.

We also address the issue of the predictive performances of this nonparametric estimator when compared to some parametric alternatives, adopting the recent ”revealed performance” approach proposed by Racine and Parmeter (2014). While nonparametric models have been shown to significantly improve the predictive ability of parametric models in some cases, this result is not assured ex ante. For instance, Racine and Parmeter (2014) provide empirical evidence showing that overspecified parametric and nonparametric estimations may not be accurate. The curse of dimensionality problem of nonparametric specifications and the bias-efficiency trade-off, which generally arises when comparing parametric and nonparametric models, are some of the main reasons of this uncertainty. Therefore, in spite of the a priori appeal of nonparametric modeling and because of the great uncertainty surrounding the true DGP, it could be of interest to compare parametric and nonparametric models in the present framework.

The remainder of the paper is organized as follows. Section 2.2 introduces an empirical specification which allows localized innovation to be detected. It also describes the nonparametric methods we use to estimate the above specification. Sections 2.3 and 2.4 present the data that are used and the results of the analysis. Section 2.5 concludes the paper and presents our further steps in this work. Appendix A provides information on the data used and the process of data cleaning that is followed, while it also provides some descriptive statistics. Finally, appendix B describes the procedure that we follow to select instruments.

2.2 Econometric specification and estimation method

2.2.1 Kernel regression with mixed data and shrinkage estimators

For a production function $f : \mathbb{R}_+^k \times \mathbb{D} \to \mathbb{R}_+$, we assume the following econometric model:

$$Y = f\left(\mathbf{Z}\right) + u. \tag{2.1}$$

Y is a scalar, square integrable dependent variable measuring the produced output and $\mathbf{Z} = (\mathbf{X}, S, I)$ is a vector of explanatory variables that includes the $k \times 1$ vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ of continuous inputs, a categorical variable S for sectors and the dummy innovation variable I. $\mathbb{D} = \mathbb{D}_{\mathsf{S}} \times \mathbb{D}_{\mathsf{I}}$ is the range of the discrete variables $\mathbf{D} = (S, I)$. f is the unknown function of interest and u is the error term.

The generalized product kernel approach (see Racine and Li, 2004; Darolles et al., 2011) is adopted to estimate f. We denote by l_{S} and l_{I} the univariate kernels of the discrete variables S and I respectively, and by $w_m, m = 1, 2, \ldots, k$ the kernels of the continuous variables $\mathbf{X}$. Further, we denote by $\boldsymbol{\gamma} = (\boldsymbol{\lambda}, \mathbf{h})$ the vector of smoothing parameters of $\mathbf{Z}$, where $\boldsymbol{\lambda} = (\lambda_{\mathsf{S}}, \lambda_{\mathsf{I}})$ is the vector of bandwidths for $\mathbf{D}$ and $\mathbf{h} = (h_1, h_2, \ldots, h_k)$ is the vector of bandwidths of $\mathbf{X}$. Let $\mathbf{z} = (\mathbf{x}, \mathbf{d})$ and $\mathbf{d} = (\mathsf{s}, \mathsf{i})$. For $\mathbf{z}, \mathbf{Z} \in \mathbb{R}_+^k \times \mathbb{D}$, the generalized product kernel is defined by:

$$K(\mathbf{Z}, \mathbf{z}, \boldsymbol{\gamma}) = L(\mathbf{D}, \mathbf{d}, \boldsymbol{\lambda}) W(\mathbf{X}, \mathbf{x}, \mathbf{h}), \tag{2.2}$$

where $L(\mathbf{D}, \mathbf{d}, \boldsymbol{\lambda}) = l_{\mathsf{S}}(S, \mathsf{s}, \lambda_{\mathsf{S}}) l_{\mathsf{I}}(I, \mathsf{i}, \lambda_{\mathsf{I}})$ is the product kernel of the discrete variables and $W(\mathbf{X}, \mathbf{x}, \mathbf{h}) = \prod\limits_{m=1}^{k} h_m^{-1} w_m(X_m, x_m, h_m)$ is the product kernel for the continuous variables. In this work we employ second order gaussian kernels for the continuous variables and Li-Racine kernels for the discrete variables. In particular, for a variable $D \in \mathbf{D} = (S, I)$, the discrete kernel at point d is defined by:

$$l_{\mathsf{D}}(D, d, \lambda_{\mathsf{D}}) = \begin{cases} 1 & \text{if } D = d \\ \lambda_{\mathsf{D}} & \text{if } D \neq d \end{cases}$$

Hall et al. (2007) show that the Least Squares Cross Validation (LSCV) method selects the bandwidths so that the smoothing parameters associated with irrelevant discrete explanatory variables converge in probability to their upper extreme value.[1] To the other extreme, if the smoothing parameter of a discrete variable is zero, the respective kernel becomes an indicator function and the estimator becomes a frequency estimator. These properties are closely related to the concept of *shrinkage* that is described in Kiefer and Racine (2009, 2017), who provide a Bayesian interpretation of the smoothing parameter. In particular, they associate the smoothing parameter of a discrete kernel in a kernel estimator with the prior variance in a Bayes estimate, highlighting that the smoothing parameter is larger for groups that are more homogeneous, where the prior variance of the Bayes model is small, and smaller for less homogeneous groups, where the prior variance is larger.

Note as well that the kernel estimator can be written as a convex combination of a frequency

[1] This argument is also true for continuous variables, but only in the case of the local constant regression, where a bandwidth going to infinity produces a constant fit.

and a pooled estimator, in which the weight is determined by the smoothing parameter. Indeed, for a regression of Y on a categorical variable G, Kiefer and Racine (2009, 2017) show that the kernel estimator $\hat{y}_g$ of the expectation of Y conditional on a group $G = g$ is described by:

$$\hat{y}_g = (1 - \Phi)\tilde{y}_g + \Phi\bar{y}_{.}, \tag{2.3}$$

G where $\tilde{y}_g$ is the conditional mean frequency estimator and $\bar{y}_{.}$ is the unconditional mean frequency estimator. Φ is a quantity in $[0, 1]$ that depends on the group and the smoothing parameter. If the later is zero, Φ is zero, so that the kernel and the frequency estimators coincide. For a value of the smoothing parameter at the upper extreme, Φ equals one and the groups are pooled together. Selecting the bandwidth, LSCV can determine the appropriate level of heterogeneity between these two limits; the pooled and the frequency estimators are both extreme cases, while *"the truth probably lies somewhere in between. The parameters are not exactly the same, but there is some similarity between them"* (Maddala et al., 1997, p.91).

Indeed, it can be expected that there is some degree of similarity in terms of production technology across technologically close sectors. This remark is also implied in sector taxonomies such as the Industry Classification Benchmark (ICB) or the work in Pavitt (1984), who identifies common patterns at the sectoral level and categorizes the manufacturing industries into groups according to their innovation strategies.[2] Adopting the generalized kernel method, the closeness between sectors is data driven and it can be tuned, for instance, by LSCV instead of using a specific taxonomy.

In summary, the generalized kernel method is robust to functional form specifications and allows to model situations involving complex dependence among categorical and continuous data, like the one that this work aims to highlight.[3] To the best of our knowledge this is the first work applying generalized product kernel to study the relation between innovation and productivity. This is crucial to highlight the possible presence of LTC and also to account for fully heterogeneous technologies across sectors, since the standard frequency approach is unfeasible in practice with small sized sectors and, for such a reason, researchers typically are obliged to assume that there is just a parallel shift in technology across sectors.

2.2.2 Nonparametric regression with endogenous innovation

The endogeneity of innovation is tackled applying instrumental variables (IV) methods which have been recently proposed and allow estimation and inference in nonparametric models with endogenous explanatory variables (see, e.g. Horowitz, 2011; Darolles et al., 2011). In general, if at least one of the explanatory variables in $\mathbf{Z}$ is endogenous, then f in (2.1) cannot be estimated by $E(Y|\mathbf{Z})$. One approach to estimate f is to use a vector $\mathbf{W}$ of IVs so that $E(u|\mathbf{W}) = 0$. f is estimated as the solution to the following inverse problem:

$$E(Y|\mathbf{W}) = E(f(\mathbf{Z})|\mathbf{W}). \tag{2.4}$$

[2] A recent work in sectoral classification based on innovation activity is found in Castellacci (2008).

[3] see also He and Opsomer (2015) and Li et al. (2016).

Assume the following linear, adjoint operators

$$T : \mathbb{L}^2_Z \to \mathbb{L}^2_W,\ \ Th = E(h(\mathbf{Z})|\mathbf{W}) \text{ and} \tag{2.5a}$$

$$T^* : \mathbb{L}^2_W \to \mathbb{L}^2_Z,\ \ T^*k = E(k(\mathbf{W})|\mathbf{Z}), \tag{2.5b}$$

where $\mathbb{L}^2_Z$ and $\mathbb{L}^2_W$ are the sets of all square integrable functions of $\mathbf{Z}$ and $\mathbf{W}$ respectively. Moreover, denote by r the conditional expectation of Y given $\mathbf{W}$, that is $r(\mathbf{W}) = E(Y|\mathbf{W})$. Assuming that $f \in \mathbb{L}^2_Z$, (2.4) is equivalent to the expression

$$r = Tf. \tag{2.6}$$

The identification of f is related to whether there is a unique solution to (2.6). Newey and Powell (2003) and Newey (2013) highlight that f is identified if and only if $\kappa(\mathbf{Z}) = 0$ almost everywhere is the only function that satisfies $E(\kappa(\mathbf{Z})|\mathbf{W}) = 0$. Some remarks are in order. First, the identification of f and the completeness of the conditional expectation are equivalent (see D'Haultfoeuille, 2011). Moreover, the completeness is equivalent to the injectivity of T because it implies that the nullspace of $E(\cdot|\mathbf{W})$ is zero. Canay et al. (2013) show that for nonparametric models and under commonly imposed restrictions, the null hypothesis that the completeness condition does not hold is not testable. Therefore, the usual maintained identification assumption is that T is one-to-one (see Darolles et al., 2011; Horowitz, 2011). Then, T has an inverse T^{-1} and the solution f in (2.6) is given by:

$$f = T^{-1}r. \tag{2.7}$$

Under the conditions of Picard theorem (see, e.g. Kress, 1999b), the solution described in eq.(2.7) is given by:

$$f = \sum_{j=1}^{\infty} \frac{\langle r, \psi_j \rangle}{\lambda_j} \phi_j \tag{2.8}$$

where $(\lambda_j, \phi_j, \psi_j), j = 1, 2, \ldots$ are determined by the Singular Value Decomposition (SVD) of T. $\lambda_j, j = 1, 2, \ldots$ are the singular values of T, arranged in decreasing order so that $\lambda_j \to 0$ as $j \to \infty$, while $\phi_j, j = 1, 2, \ldots$ and $\psi_j, j = 1, 2, \ldots$ are orthonormal sequences in $\mathbb{L}^2_Z$ and $\mathbb{L}^2_W$ respectively, such that $T\phi_j = \lambda_j \psi_j$ and $T^*\psi_j = \lambda_j \phi_j$, $\forall j$. The problem is ill-posed because the mapping from r to f in (2.7) is not continuous. Indeed, consider an estimation $\hat{r} = r + \delta\psi_j$ that enters (2.6), the error level $\delta > 0$ being arbitrarily small. Darolles et al. (2011) mention that this approximation of r introduces an error which leads to a result $f + \frac{\delta}{\lambda_j}\phi_j$ infinitely far from the exact solution f. The large change in f due to an arbitrarily small change in r is also explained in Horowitz (2011).

Because of the ill-posedness of the problem, a consistent estimator of f cannot result from plugging in (2.7) consistent estimators of r (Newey, 2013). A solution to this issue is to create an estimator where ill-posedness does not affect consistency; a sequence of bounded operators $R_\alpha : \mathbb{L}^2_W \to \mathbb{L}^2_Z,\ \alpha > 0$ can be found to approximate the unbounded inverse operator T^{-1}. Regularization is the approximation of an ill-posed problem by a family of well-posed problems such that their solutions converge to the solution of the ill-posed problem. The regularization scheme approximates the solution f in (2.6) by $\hat{f}_\alpha = R_\alpha \hat{r}$. Denote by δ the error level imposed by the estimation of r, so that $||\hat{r} - r|| \leq \delta$, $||\cdot||$ being the usual norm in $\mathbb{L}^2$. The total approximation error is given by the inequality $||\hat{f}_\alpha - f|| \leq \delta ||R_\alpha|| + ||R_\alpha T f - f||$. As described in Kress (1999b, ch.15), the first term of the right hand side is the sample error, which increases as $\alpha \to 0$. The second term is the error introduced by the approximation of T^{-1} by R_α and decreases as $\alpha \to 0$. Therefore, the regularization parameter α minimizes the overall error, balancing between accuracy and stability.

Instead of directly regularizing the problem in (2.6), a common practice (see, e.g. Carrasco et al., 2007; Horowitz, 2011; Centorrino et al., 2017) is to reshape the equation as follows. Assuming that the completeness condition holds and that both operators T and T^* are compact, we project (2.6) onto the space of $\mathbf{Z}$, resulting in:

$$T^* r = T^* T f \tag{2.9}$$

Denote by $\hat{T}^*$, $\hat{T}$ and $\hat{r}$ the estimators of T^*, T and r respectively, obtained by standard non-parametric estimators such as kernels and splines (Darolles et al., 2011; Horowitz, 2011). Then, the sample counterpart of (2.9) is:

$$\hat{T}^* \hat{r} = \hat{T}^* \hat{T} f. \tag{2.10}$$

Darolles et al. (2011) mention that f is identifiable if and only if T^*T is one-to-one. Because $(T^*T)^{-1}$ is noncontinuous, the problem remains ill posed. The regularization procedure is applied in (2.10) rather than in (2.6).

In literature, a plethora of different ways to regularize exists, including Tikhonov methods and sieve estimations. In the present study the Landweber - Fridman approach (Landweber, 1951; Fridman, 1956) is followed. This method avoids the inversion of the $\hat{T}^*\hat{T}$ matrix in (2.10) by using an iterative process (see Carrasco et al., 2007; Johannes et al., 2013) which leads to the recursive solution:

$$\hat{f}_{\nu+1} = \hat{f}_\nu + c\hat{T}^* \left(\hat{r} - \hat{T}\hat{f}_\nu \right), \ \forall \nu = 0, 1, \ldots \tag{2.11}$$

where c ensures the convergence of the iterative process and is selected so that $c||T^*T|| < 1$. The solution in (2.11) can be rewritten as $\hat{f}_{1/\alpha} = c \sum\limits_{\nu=0}^{1/\alpha - 1} \left(I - c\hat{T}^*\hat{T} \right)^\nu \hat{T}^* \hat{r}$. $1/\alpha$ is the total number of iterations needed and α represents the regularization parameter. The number of iterations can be specified minimizing a leave-one-out cross validation criterion (Centorrino, 2015). Following (2.11), after deriving $\hat{f}_0 = c\hat{T}^*\hat{r}$ from a first estimation of r, T and T^*, $\hat{f}_1$ is computed. This step is repeated until a stopping rule, such as the following minimization criterion presented in Centorrino et al. (2017).

$$SSR(\nu) = \nu \sum_{i=1}^{n} \left[(\hat{T}\hat{f}_\nu) - \hat{r} \right]^2, \ \nu = 1, 2, \ldots \tag{2.12}$$

The iteration procedure stops when the above criterion, which describes a locally convex function, starts to increase. In this study we employ a similar stopping rule; the procedure ends when the quantity $||(\hat{r} - \hat{T}\hat{f}_\nu)/\hat{r}||^2$ stops falling.

2.3 Data

Data used for the analysis are sourced by the tenth Survey on Manufacturing Firms (Indagine sulle Imprese Manifatturiere), provided by Unicredit-Capitalia, which are complemented with balance sheets sourced by either AIDA (Italian Balance Sheet Dataset of the Bureau van Dijk) or by the Chambers of Commerce Registry (UNICREDIT, 2008). Noteworthy, these data are also used in Antonioli et al. (2018). The same survey, although in different waves, has been largely used in the economics literature investigating firms innovation activities (see, e.g. Parisi et al., 2006; Hall et al., 2009). Therefore, our present work is fully comparable to previous relevant studies. The initial sample includes all firms with more than 500 employees, while for smaller firms it is stratified by firm size, value added, geographical location and industry. We follow a cleaning procedure, detailed in appendix A, which results in a final sample of 2.750 small and medium-sized firms (SMEs). The main variables of the sample and descriptive statistics are also described in detail in appendix A. Information on the innovation activity of firms is derived from the survey that was conducted in 2007 and refer to the three-year period 2004-2006. Value added and labour are reported in the balance sheets and refer to the year 2006. Capital stock refers to 2006 and is derived from the Italian National Statistical Office (ISTAT). A categorical variable accounts for sectors and follows the ATECO 2002 classification.

2.4 Results

2.4.1 IV selection

As far as IV selection is concerned, we first select an initial set $\mathbf{S}$ of potential IVs according to the literature on KPF (see, e.g. Musolesi and Huiban, 2010).[4] In order to ensure the presence of strong and valid instruments, we follow a two-step procedure. To the best of our knowledge, IV selection approaches are not applicable within a nonparametric framework, and for that reason we adopt a baseline Cobb-Douglas (CD) specification where innovation and sectors are additively introduced. In the first step we regress innovation on $\mathbf{Z}$ and $\mathbf{S}$, and then choose, using a backward selection algorithm, a set of potential instruments that are strongly correlated with innovation. In the second step we estimate the CD model using the two-stage least squares (IV-2SLS) estimator for all the possible combinations of variables from the above selected set, and then we perform tests to investigate the validity and strongness of the selected variables. We finally select 3 sets (IV1, IV2 and IV3) of strong and valid instruments. Noteworthy, the results in the following subsections are robust across these groups. We illustrate figures only for one set (IV1) for which the nonparametric specification exhibits the lowest ASPE. All the details of the IV selection process can be found in Appendix B.

[4] For reasons of brevity, the set of potential IVs is presented in the appendix.

2.4.2 Model comparison

In this section we present the results of the model comparison between the nonparametric specification in (2.1) with alternative parametric counterparts. The generalized product kernels approach (Darolles et al., 2011; Horowitz, 2011) has been used to estimate a local constant kernel regression. The smoothing parameters of the kernels are selected via the LSCV method. We also consider a CD and a translog (TL) production function augmented with endogenous dummy innovation (Mohnen and Hall, 2013; Musolesi and Huiban, 2010). Estimating a production function where innovation enters additively is the standard practice in empirical studies on productivity. We also consider a CD and a TL production function with heterogeneous effect of innovation on the production process (CDh and TLh, respectively), which allow different slope parameters of the input variables capital (K) and labour (L) between innovative and non-innovative firms (see Antonioli et al., 2018). The above models are estimated using standard IV-2SLS methods. It is worth to note that in the cases of CDh and TLh, the interaction terms with innovation are also endogenous, which leads to the requirement for extra variables to instrument these endogenous terms. Therefore, the number of instruments increases, while it is well known that the IV-2SLS estimation in the presence of many instruments may result in substantial bias and inaccurate inference (see, e.g. Hansen et al., 2008).

We perform a pseudo Monte Carlo experiment, as introduced in Racine and Parmeter (2014), to assess the predictive ability of the models. After randomly shuffling and splitting the n observations at a first 90% into n_1 training points and at 10% into n_2 evaluation points, each model is fitted according to the training sample. Then, we compute the average out-of-sample squared prediction error (ASPE) for each specification σ using the evaluation points. ASPE is given by:

$$ASPE = n_2^{-1} \sum_{i=n_1+1}^{n} (y_i - \hat{\sigma}\left(\mathbf{z}_i\right))^2,$$

where y_i and $\hat{\sigma}\left(\mathbf{z}_i\right)$ is the true value of Y and the respective fitted value according to the training observations, at point i. The above steps are repeated a large number of times $\rho = 1000$, which results in a $\rho \times 1$ vector of prediction errors for each model. This procedure is followed using different IV sets for the estimations. For reasons of space and because of the robustness of our findings, while tables 1 and 2 refer to the results for three IV sets (IV1, IV2 and IV3), we illustrate figures only for one set (IV1).

Table 1 presents the median of the ASPE for each specification across different IV sets. A first and very crucial result that emerges is that, in all IV sets, the median that corresponds to the nonparametric model is the smallest among the different specifications. For instance, in the estimations using the IV1 set, the median ASPE of the NP model relative to the ones of LC, LCh, TL and TLh is 0.916, 0.892, 0.922 and 0.776, respectively. A second result, which holds for all the IV sets, is that the median ASPE of the TLh specification is the largest; the model complexity of TLh might be responsible for its poor predictive ability. For example, the median ASPE of the TLh model relative to the other models is 1.178, 1.148, 1.186 and 1.286. A third finding is that the median ASPEs of the CD and TL estimations are very close for all IV sets, their difference being at the level of 10^{-4}. The predictive ability of CD and TL might be similar.

Finally, a fourth result is that the CDh results are ambiguous; its median ASPE is less than the median ASPEs of CD and TL in estimations using the IV2 set and higher in estimations using the IV1 and IV3 sets. The above findings are also illustrated in figure 2.1, which presents the box-and-whisker plots of the distributions of ASPE for the different models.

Figure 2.2 shows the empirical distribution functions (ECDFs) of the ASPEs for each specification. It is clear that the ASPE of the NP model is stochastically dominated by the ASPE of any of the parametric models. This result shows that the nonparametric specification outperforms the parametric ones in terms of predictive ability. Further, the ASPE of the TLh specification stochastically dominates the ASPE of any other specification, indicating that the TLh model underperforms. The above findings are very robust across different IV sets. On the contrary, the results of the comparisons among the parametric specifications are less clear. Results that are not presented in the main text show that while the ECDF of the ASPE for the CDh specification is stochastically dominated by the ones of the CD and the TL in estimations using the IV2 set, it dominates them in estimations with IV1 and IV3 sets. Therefore, the comparisons between the CDh and the CD and between the CDh and the TL specifications do not provide robust results. Noteworthy, the ECDFs of the CD and the TL models almost coincide, which shows that their predictive performance is similar.

We finally compare the different specifications using the test for stochastic dominance by Dunn (1964). The test performs a pairwise comparison that is based on the null hypothesis that the probability of observing a larger randomly selected value of the first group than of the other is one half. The results for all pairwise comparisons for the different IV sets are presented in Table 2. In all cases apart from the CD - TL comparison, the tests reject the null hypothesis. The tests are towards the findings presented in figure 2.2 and indicate that the differences in ECDFs are statistically significant in all cases but the CD - TL pair, for which the test shows that the differences of the ECDFs of the ASPE between the CD and the TL specifications are not significant.

In summary, we provide clear evidence that a) the NP model always outperforms all the others, b) the TLh specification underperforms all the others, and c) the choice across the other three specifications CD, TL and CDh depends crucially on the IV set that is used.

2.4.3 Estimation results

In this section we present the main estimation results of the nonparametric specification in (2.1). For reasons of brevity, in figures 2.3 and 2.4 we show the results with respect to only 4 of the 22 manufacturing sectors under consideration, leaving the results for the remaining ones to be available upon request. In particular, we present the results for firms that belong to sectors 17, 24, 29 and 35, which correspond to the manufacture of *textiles*, *chemical products*, *machinery* and *transport equipment*, respectively. The numbering of the sectors is derived from the NACE Rev.1.1 Eurostat classification. A first result is that innovation has a positive effect on productivity, but not for all parts of the capital-labour domain. Non-innovative firms could be more productive than innovative firms for specific production inputs. This finding is illustrated more clearly in figure 2.4, which presents the contour plots of the difference in the estimation of f between innovative and non-innovative firms, for each sector s, that is $\hat{f}(\mathbf{X}, S = s, I = 1) - \hat{f}(\mathbf{X}, S = s, I = 0)$. The contours are grey colored for positive values of the difference and black colored for negative values. Clearly, the effect of innovation on the production process is not positive for all input sets. This finding is towards the localization of innovation and in contrast to empirical studies that assume an increase in productivity at all factor proportions.

Further, the results in figures 2.3 and 2.4 provide insights into the role of sectors to the relation between innovation and productivity. This role is shown clearly comparing the production output that corresponds to different sectors, fixing the innovation variable. For instance, figure 2.5 shows such a comparison associated with sectors 17 and 24. The figure shows that the effect of sectors on the production process is heterogeneous. This finding is also implied by the value of the smoothing parameter of the sectors' variable, which is 0.081, 0.074 and 0.073 for the kernel estimations that use the IV1, IV2 and IV3 sets, respectively. These values are very small and indicate heterogeneous technologies across sectors.

2.5 Conclusion

This paper provides an empirical analysis of a fully nonparametric production function model adopting recently introduced IV methods that allow estimation and inference in a nonparametric framework while handling endogeneity of explanatory variables (Darolles et al., 2011; Horowitz, 2011). The comparison with commonly used parametric counterparts shows that the nonparametric specification performs better in terms of predictive ability. In contrast, the TL model with heterogeneous effect of innovation on productivity underperforms, while the CD model with heterogeneous effect does not show robust results, as its performance varies depending on the IV set; better than CD and TL for IV2 but worse than CD and TL for IV1 and IV3. Turning to the estimations, the results question the orthodoxy that innovation has a positive effect on productivity at all factor proportions. Our findings show that there are combinations of capital and labour where non-innovative firms are more productive than innovative firms. In general, our findings lean towards the original idea of Atkinson and Stiglitz (1969) about the localization of innovation. The results also question commonly used shape restrictions such as Hicks neutrality and additivity of innovation. Further, they shed light to the role of the sectors' variables to the relation between innovation and productivity. Our findings indicate the presence of heterogeneous technologies across sectors and highlight the non-additive, more complex role of sectors to productivity.

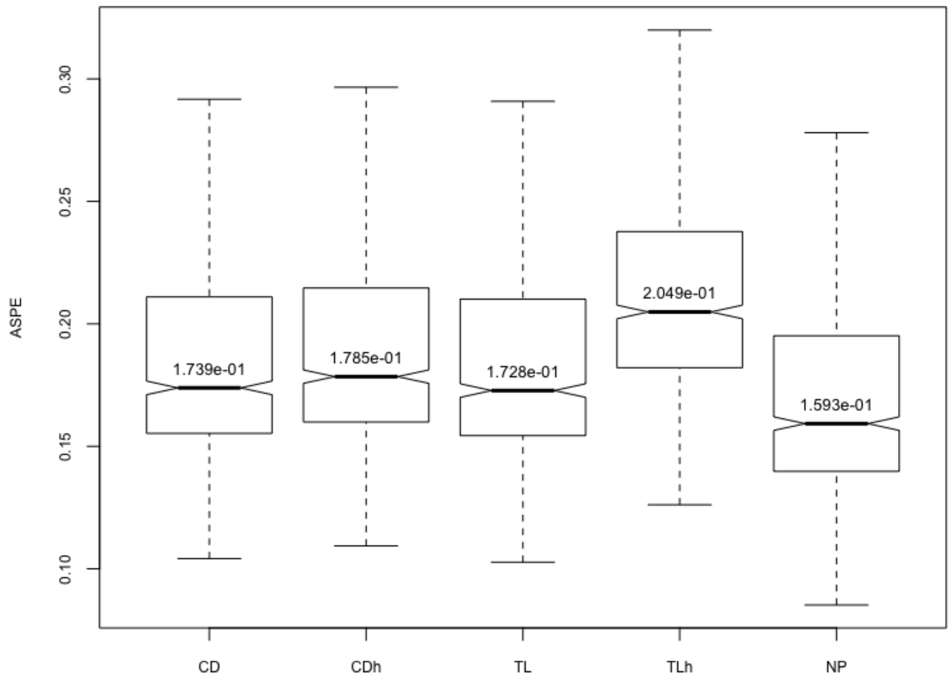


Figure 2.1: Out-of-sample average square prediction error (ASPE) box plots for different models, in the case of the IV1 set.

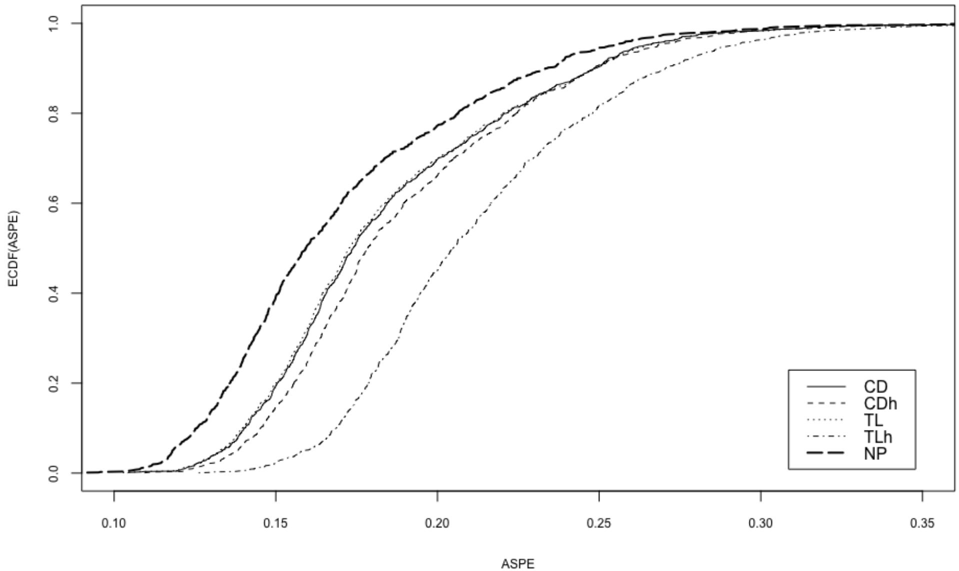


Figure 2.2: Empirical Cumulative Distribution Functions (ECDFs) of the ASPE for different models, in the case of the IV1 set.

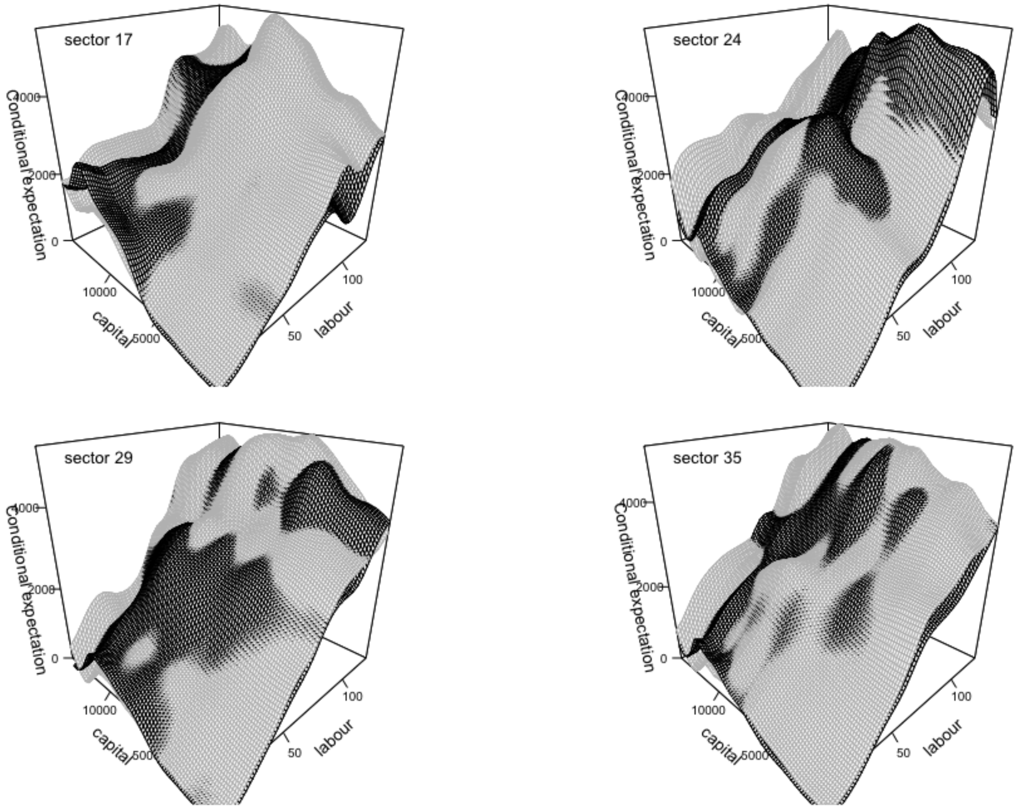


Figure 2.3: Estimated production function for different sectors, for the IV1 set. The function for innovative firms is colored grey and for the non-innovative firms is colored black.

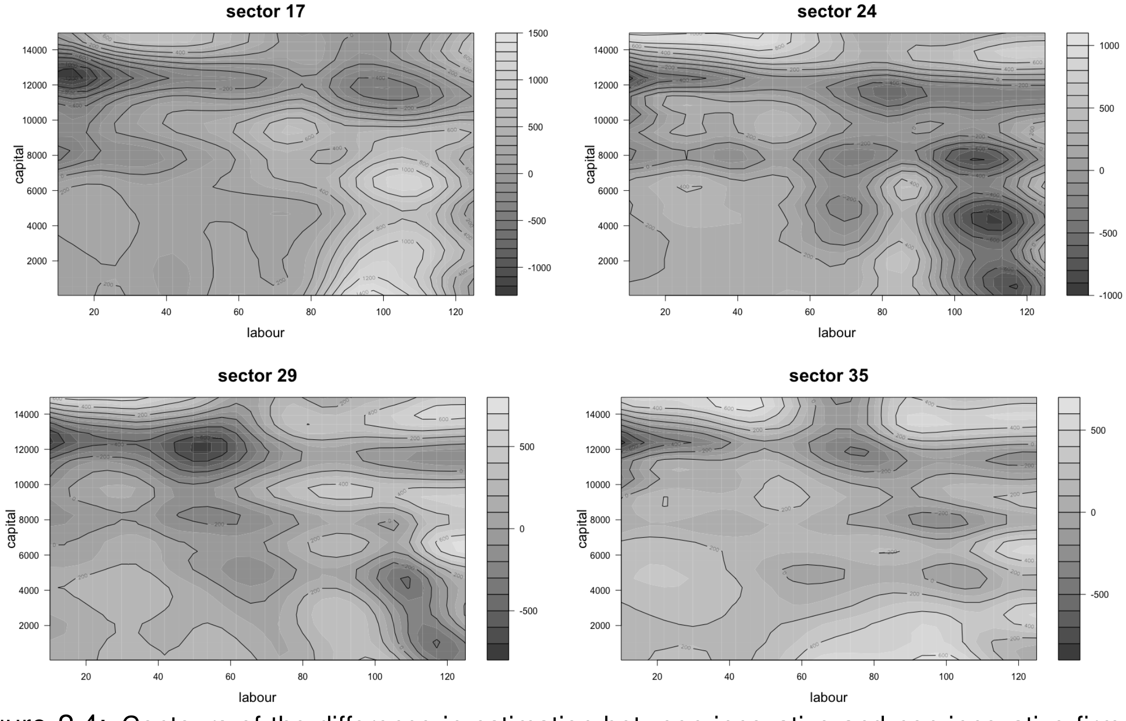


Figure 2.4: Contours of the difference in estimation between innovative and non-innovative firms, for each sector $s = 17, 24, 29$ or 35. The estimations correspond to the IV1 set.

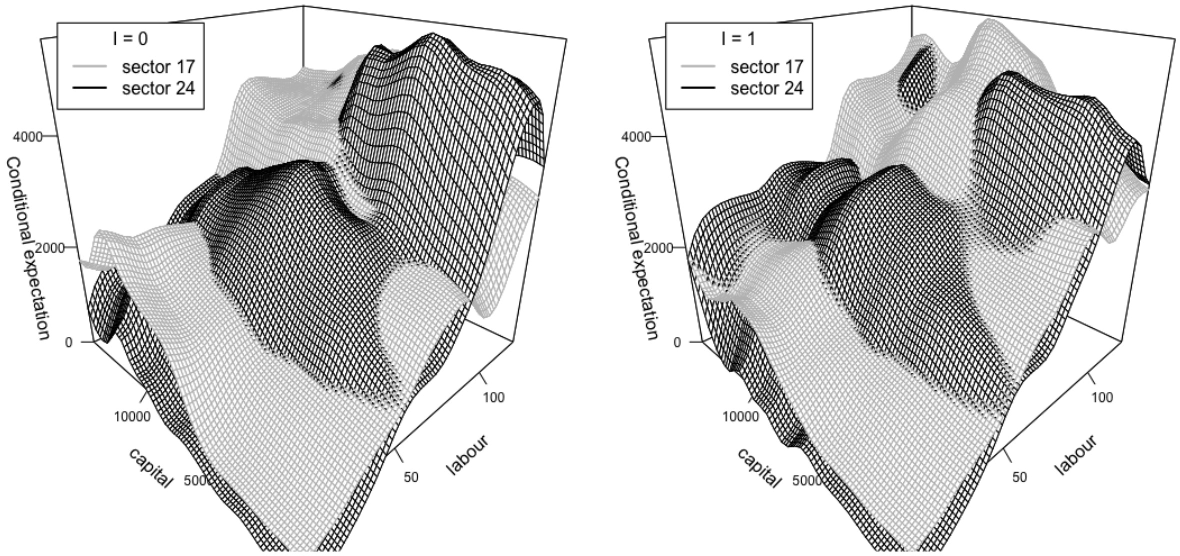


Figure 2.5: Comparison between two sectors, for non-innovative and innovative firms, associated with the IV1 set.

Table 1 - Median ASPE

	Median of ASPE					Values relative to NP			
IV set	CD	CDh	TL	TLh	NP	CD	CDh	TL	TLh
IV1	.1739	.1785	.1728	.2049	.1593	1.092	1.121	1.085	1.286
IV2	.1736	.1681	.1729	.1834	.1605	1.082	1.047	1.077	1.143
IV3	.1707	.1762	.1706	.1846	.1620	1.054	1.088	1.053	1.140

Median ASPEs (left) and relative values of median ASPEs with respect to NP (right).

Table 2 - Stochastic dominance tests

models	**IV sets**	CDh	TL	TLh	NP
	IV1	-2.652^{***}	$.434$	-15.296^{***}	8.420^{***}
CD	IV2	3.420^{***}	$.413$	-4.681^{***}	8.217^{***}
	IV3	-2.727^{***}	$.017$	-6.598^{***}	4.666^{***}
	IV1		3.085^{***}	-12.644^{***}	11.072^{***}
CDh	IV2		-3.008^{***}	-8.101^{***}	4.797^{***}
	IV3		2.744^{***}	-3.871^{***}	7.392^{***}
	IV1			-15.729^{***}	7.986^{***}
TL	IV2			-5.094^{***}	7.804^{***}
	IV3			-6.615^{***}	4.648^{***}
	IV1				23.716^{***}
TLh	IV2				12.898^{***}
	IV3				11.263^{***}

Dunn's (1964) test for stochastic dominance. Null hypothesis: 50% probability of observing a larger randomly selected value from the first distribution than from the second.

Classification of p-value: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Appendix A: Data, variables and sample

Data used for the analysis are sourced by the tenth Survey on Manufacturing Firms (Indagine sulle Imprese Manifatturiere), provided by Unicredit-Capitalia, which are complemented with balance sheets sourced by either AIDA (Italian Balance Sheet Dataset of the Bureau van Dijk) or by the Chambers of Commerce Registry (UNICREDIT, 2008). The same survey, although in different waves, has been largely used in the economics literature investigating firms innovation activities (see, e.g. Parisi et al., 2006; Hall et al., 2009). The initial sample consists of 3.770 observations. It includes all firms with more than 500 employees, while for smaller firms it is stratified by firm size, value added, geographical location and industry. The main variables of the data set that we use in our study are:

1. Value added (K €) Y: referred to year 2006 and reported in the balance sheets.
2. Capital stock (K €) K: referred to year 2006, calculated by summing the investments with the value of fixed assets, as estimated through a perpetual inventory method considering the usual rate of depreciation of 0.05. Both measures of investments and fixed assets are available from 1998 to 2006, and both are deflated with the respective aggregate price index (derived from ISTAT).
3. Labour L: number of employees in 2006, reported in the balance sheets.
4. Innovation I: innovation dummy taking value 1 if the firm affirmed to have introduced at least one product or one process innovation during the period 2004-2006, and 0 otherwise.
5. Sectors S: the manufacturing firms are classified by sectors according to the two digit ATECO2002 classification, which derives from NACE Rev.1.1 Eurostat classification.

Our preliminary data analysis is focused mainly on two purposes; the first one is to identify and remove the outliers of the sample using an appropriate detection method, while the second is to tackle sparsity of data in the two-dimensional (capital - labour) domain of the production function, so that the final data set be appropriate to recover a fully nonparametric production function. For these reasons we follow the cleaning procedure detailed below. Initially, 13 observations with missing or inconsistent values are dropped. Then, outliers and sparse data are detected studying the univariate distributions of the variables under investigation. Our focus is primarily on the 3.239 small and medium sized firms (SMEs), that is enterprises between 10 and 250 employees. Restricting the analysis to SMEs reduces significantly the range of labour and facilitates the identification of sparse data. We observe that, even focusing solely on SMEs, the distribution of labour is still extremely right skewed. There is a drastic reduction of the frequency of the observations for firms with more than 120-130 employees, which is also confirmed by the cluster analysis that is described below. After this threshold the observations are few and sparse; SMEs with more than 125 employees account to less than 8% of the total. Therefore, we restrict the analysis to the subgroup of SMEs with labour $L \leq 125$. This set consists of 2.985 observations and it may be a reasonable compromise between studying a relevant fraction of the small and medium-sized firms that were initially sampled, while obtaining a dense data set for nonparametric estimation.

With regard to value added and capital, a visual inspection of the empirical distribution does not allow to identify the data that have to be deleted. In order to detect outliers, the boxplot rule is applied to both variables, from which another 150 observations (5% of the total) which exceed the boxplot's outer fences are dropped. This results in a data set of 2.822 firms and a significant reduction to the range of the variables. The same process is followed with respect to productivity. In particular, we define Total Factor Productivity as $TFP = Y/(K^{0.5}L^{0.5})$, and afterwards we apply the boxplot rule again to detect another 72 outlying observations. 2D and 3D scatterplots do not reveal any other outliers apart from the already omitted ones. The final data set consists of 2.750 firms.

Generally, Tukey's method applies in both symmetric and skewed distributions. Nevertheless, its outlier detection capacity may be affected if the data are significantly skewed. For this reason we also use adjusted boxplot for skewed distributions, as proposed by Hubert and Vandervieren (2008). The method takes into consideration skewness through the medcouple quantity, which was first introduced by Brys et al. (2003). Although the total number of outliers detected is quite the same (226 for adjusted boxplot, 222 for Tukey's rule), the resulting range for both value added and capital is much larger in the case of the adjusted boxplot, therefore creating a comparatively sparse dataset. For this reason Tukey's rule is preferable.

In order to check the sparsity of data in the capital - labour plane, we perform cluster analysis to the initial sample of 3.239 SMEs. Data clustering gives an insight into data, reveals their structure and provides class identification (Jain, 2010). Moreover, clustering can also be an effective way to identify outliers (Zhao, 2012). By grouping data into clusters, observations which are not assigned to any cluster can be considered as outliers. In this framework, *dbscan* is a well known density based cluster analysis method which has these characteristics. It is proposed by Ester et al. (1996) and is based on the definition of a neighborhood around each observation and a specification of a threshold density. The neighborhood of a sample point p is specified by a radius r, while the threshold density of a neighborhood is determined by a number of points k inside. Both r and k are pre-fixed by the analyst. Core points of the sample are the observations whose neighborhood has at least the threshold density. Based on the above definitions, a point p is in a cluster C if it belongs to the neighborhood of a core point of this cluster. Therefore, the method groups points of densely populated areas into clusters (Zhao, 2012). Unlike other clustering algorithms such as *k-means* (Hartigan and Wong, 1979) or *k-medoid* (Kaufman and Rousseeuw, 2009), *dbscan* does not impose any restrictions to the shape of the clusters. Moreover, it does not require a pre-identification of the number of clusters. The method is suitable for removing sparse observations because it identifies them as unclustered.

Nevertheless, the specification of clusters and unclustered points with *dbscan* is sensitive to the specification of r and k (Kandylas et al., 2010). Different values of these parameters may lead to different clustering. For this reason we adopt the following graphical heuristic. After standardizing the variables K and L so that the definition of neighborhood is meaningful, k is fixed to 4 as suggested by Ester et al. (1996) for 2-dimensional spaces, such as the (K, L)

plane. To specify r, we compute the distance d of every point to its first 4 nearest ones. The values are sorted in ascending order in a graph. The point in the graph beyond which the slope becomes much steeper[5] provides an appropriate radius to detect unclustered observations. In our case r is between 0.10 and 0.13 (see fig.2.7).

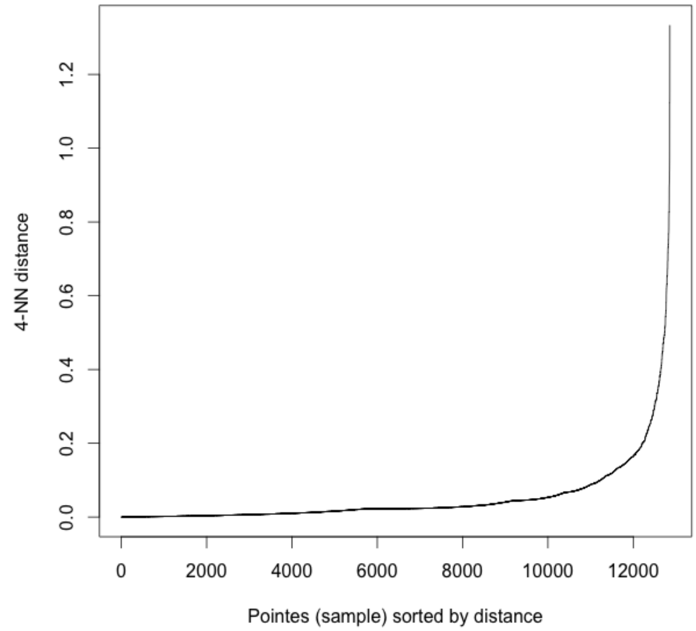


Figure 2.6: Total number of neighborhood points with respect to different prefixed radii r. The threshold density is determined by $k = 4$. The graph refers to the sample of all SMEs, after removing observations with inconsistent values.

The resulting clustering shows that in regions of the plane where labour is greater than 120-130 there are practically no clustered data. This appears to be extremely consistent with the data cleaning we performed using univariate analyses (note that in the final dataset the maximum value for capital is 14.952). Therefore, we do not drop other observations to the sample previously defined.

[5] This point is referred as "knee" of the graph (Ester et al., 1996).

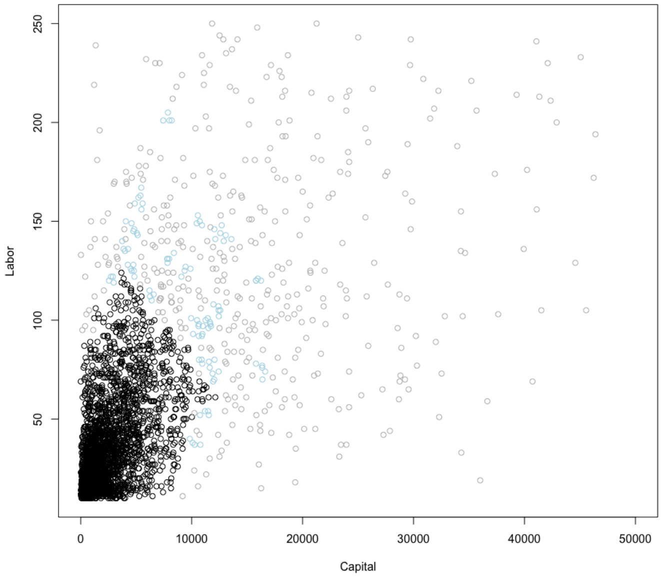


Figure 2.7: Data clustering with *dbscan* for $r = 0.1$ and $k = 4$. Black depicts the largest cluster, blue the remaining clusters and grey the unclustered data. The (K, L) plane is scarse for values of labour above $120 - 130$ employees.

Descriptive statistics are presented in table 3 below. Note that the percentage of innovating firms in the final dataset employed in this work is 64%. This value is the same as in Hall et al. (2009), who built a panel data set starting from different waves of the same survey used in this work. This value is also very close to the percentage of innovating firms obtained using the Italian CIS survey (Hall et al., 2008).

Table 3 - Descriptive statistics

	Value added			Capital			Labour		
VALUE	all firms	non innov.	innov.	all firms	non innov.	innov.	all firms	non innov.	innov.
minimum	12	12	16	39	61	39	10	10	10
1st quantile	922	846	979	900	749	990	17	16	19
median	1576	1353	1758	2025	1677	2230	31	26	34
mean	2070	1876	2181	2934	2670	3086	38.5	34.8	40.7
3rd quantile	2770	2485	2956	4048	3571	4216	53	46	56
maximum	9163	9163	8822	14952	14521	14952	125	125	125
st. deviation	1555	1531	1558	2796	2708	2834	26.1	24.9	26.5
skewness	1.49	1.84	1.32	1.63	1.76	1.57	1.13	1.35	1.03

Appendix B: IV selection

For the estimation of (2.1) we adopt a nonparametric instrumental regression approach (see Racine and Li, 2004; Darolles et al., 2011). In this appendix we present the procedure that we follow for the selection of the instruments. As a first step, we select an initial set $\mathbf{S}$ of potential IVs following the literature on KPF (see, e.g. Hall et al., 2009; Musolesi and Huiban, 2010). This group is described in table 4 below.

Table 4 - Initial set of IVs

Category	**Variables**	**Description**	**Type**
Firm Characteristics	NW, NE, C, S, Age, Group, Consortium	Firm's characteristics including geographical location; age and group or consortium membership	binary (0,1)
Human capital	RD_Personnnel	Percentage of employees in Research and Development activities	numeric
Objectives of investment	BetterProd, MoreProd, NewProd, Env, CostRed, Advert, SellNet, SellAss	Objectives including ameliorating the product, produce more, introduce a new one, reduce the environmental impact, reduce costs, to increase the selling network or to ameliorate it, respectively	binary (0,1)
Market penetration	MarkPenEU15, MarkPenEU2004, MarkPenRussia, MarkPenOtherEU, MarkPenAfrica, MarkPenAsia, MarkPenCina, MarkPenUSMex, MarkPenSouthAm, MarkPenOce	Market penetration in different world regions, including EU member states, Africa, Asia, China, U.S., Canada, Mexico, South America and Oceania	binary (0,1)
Commercial agreements	CommAgrEU15, CommAgrEU2004, CommAgrRussia, CommAgrOtherEU, CommAgrAfrica, CommAgrAsia, CommAgrCina, CommAgrUSMex, CommAgrSouthAm, CommAgrOce	Commercial agreements in world regions, as mentioned above	binary (0,1)
Patent acquisition	PatBuyEU15, PatBuyEU2004, PatBuyRussia, PatBuyOtherEU, PatBuyAfrica, PatBuyAsia, PatBuyCina, PatBuyUSMex, PatBuySouthAm, PatBuyOce	Location of the aforementioned world regions where the firm acquired patents	binary (0,1)
Production overseas	ProdAbroadEU15, ProdAbroadEU2004, ProdAbroadRussia, ProdAbroadOtherEU, ProdAbroadAfrica, ProdAbroadAsia, ProdAbroadCina, ProdAbroadUSMex, ProdAbroadSouthAm, ProdAbroadOce	Production located in the aforementioned world regions	binary (0,1)
Competitiveness	LowCompet, HighCompet, SmallProdScale	Perceived level of competitiveness and scale of production compared to competitors	binary (0,1)
Financial specs	ListedComp, FinanIncent	Listed company or receiving financial incentives	binary (0,1)

Bound et al. (1995) and Wooldridge (2010, ch.5) highlight the potential problems of inconsistency and finite sample bias that result from the use of weak instruments in IV methods. To avoid this pitfall, we follow a two-step procedure that ensures the strongness and validity of the instruments. To the best of our knowledge, IV selection approaches are not applicable within a nonparametric framework. For that reason we adopt a baseline Cobb-Douglas (CD) specification where innovation and sectors are additively introduced.

In the first step we regress innovation on $\mathbf{Z}$ and $\mathbf{S}$, and then choose, using a backward selection algorithm, a set of potential instruments that are strongly correlated with innovation. The sets corresponding to a 10% and a 5% threshold are described in table 5 below.

Table 5 - IV sets

set	instruments
10% set	CommAgrAfrica, ProdAbroadEU15, C, Age, Group, EUCompet, RD_Personnnel, BetterProd, MoreProd, NewProd, FinanIncent, CostRed, Advert, SellNet, SellAss, MarkPenEU15, HighCompet
5% set	CommAgrAfrica, ProdAbroadEU15, C, Age, HighCompet, EUCompet, MarkPenEU15, BetterProd, FinanIncent, NewProd

In the second step we estimate the CD model using the two-stage least squares (IV-2SLS) estimator for all the possible combinations of IVs from the 5% set and then we perform robust score tests by Wooldridge (1995) to test overidentifying restrictions and endogeneity. We also apply the Montiel-Pflueger test (Montiel Olea and Pflueger, 2013) to detect weak instruments. Based on the above post-estimation tests, we find that 62 sets provide strong instruments, while for 85 combinations the validity hypothesis is not rejected. We finally select 14 sets of strong and valid instruments. The results that are presented in section 3.3 show a robust picture across these sets.

For the sake of brevity, we present three of these groups, leaving the results of the remaining IV sets to be available upon request. These sets are $IV1 = (BetterProd, MarkPenEU15, EUCompet)$, $IV2 = (BetterProd, MarkPenEU15, EUCompet, Age)$ and $IV3 = (BetterProd, MarkPenEU15, FinanIncent, NewProd)$. All variables except Age are binary. Age is the number of years of the firm in 2004. $BetterProd$ takes value 1 if the firm's objective is to ameliorate its products and 0 otherwise. $NewProd$ is 1 if the firm's objective is to create new products and 0 otherwise. $MarkPenEU15$ takes value 1 if the firm's market penetration refers to EU15 countries; 0 otherwise. $EUCompet$ is 1 if firm's competitors are from EU countries; 0 otherwise. Finally, $FinanIncent$ is 1 if the firm received financial incentives and 0 otherwise.

Chapter 3

Nonparametric estimation of international R&D spillovers

joint with Antonio Musolesi and Michel Simioni

Abstract

We revisit the issue of international technology diffusion within the framework of large panels with strong cross-sectional dependence by adopting a method which extends the Common Correlated Effects (CCE) approach to nonparametric specifications. Our results indicate that the adoption of a nonparametric approach provides significant benefits in terms of predictive ability. This work also refines previous results by showing threshold effects, nonlinearities and interactions, which are obscured in parametric specifications and which have relevant policy implications.

Keywords: large panels; cross-sectional dependence; factor models; nonparametric regression; spline functions; international technology diffusion.

JEL classification: C23; C5; F0; O3.

3.1 Introduction

With the development of endogenous growth theory since the nineties, there has been an increasing interest in international R&D spillovers. A pioneering empirical work by Coe and Helpman (1995), recently revisited by Coe et al. (2009) – henceforth CH and CHH, respectively – relates total factor productivity (TFP) to both domestic and foreign R&D and, assuming that technology spills over across countries through the channel of trade flows, constructs foreign R&D capital stock as the import-share-weighted average of the domestic R&D capital stocks of the trading partners. Subsequent studies consider other factors as channels of international spillovers, such as foreign direct investment, bilateral technological proximity, patent citations between countries, language skills or geographic proximity.

Recent studies extend the literature on international R&D spillovers by accounting for relevant methodological issues such as cross-sectional dependence and non-stationarity (Coe et al., 2009; Lee, 2006; Ertur and Musolesi, 2017) within a parametric framework.

This paper aims at revisiting the issue of international R&D spillovers using nonparametric methods. This could be relevant from both an economic and a methodological perspective. First, from an economic and poliy oriented perspective, it may allow to test the validity of the main results provided in the literature, especially with respect to the possible existence of nonlinearities, threshold effects, non-additive relations, etc., as nonparametric approaches have been shown to provide new and useful insights in topics very closely related to the present one (Ma et al., 2015; Maasoumi et al., 2007). Second, nonparametric approaches, which are recently developing also in the context of panel data (Rodriguez-Poo and Soberon, 2017; Parmeter and Racine, 2018), have been shown to significantly improve the predictive ability of parametric models in many cases (Racine and Parmeter, 2014; Ma et al., 2015; Delgado et al., 2014), even if this is not assured ex ante because of the curse of dimensionality problem of nonparametric specifications and the bias-efficiency trade-off, which generally arises when comparing parametric and nonparametric models. Therefore, it could be of interest to compare parametric and nonparametric models in the present framework.

The econometric analysis is conducted using annual country-level data for 24 OECD countries from 1971 to 2004. This dataset is also used, among others, in Coe et al. (2009) and in Ertur and Musolesi (2017) and this allows for a comparability with previous studies. The analysis is based on the nonparametric approach by Su and Jin (2012), which allows for a multifactor error structure and extends the approach by Pesaran (2006). Such an approach combines the flexibility of sieves with the ability of factor models to allow for cross-sectional dependence and to account for endogeneity due to unobservables, whereby the explanatory variables are allowed to be correlated with the unobserved factors. Following Su and Jin (2012), the nonparametric component is estimated using sieves, and particularly splines. Specifically, we adopt a regression splines framework, which provides computationally attractive low rank smoothers. We also employ penalized regression splines, as they combine the features of regression splines and smoothing splines, and have proven to be useful empirically in many aspects (Ruppert et al., 2003) while their asymptotic properties have been studied in recent years. The choice of the knots is avoided by using knot-free bases for smooths (Wood, 2003). Finally, as far as model selection is concerned, we compare alternative specifications by focusing on their predictive ability and adopt the approach recently proposed by Racine and Parmeter (2014), which is based on a pseudo Monte Carlo experiment and takes its roots on cross validation.

The paper is organized as follows. In section 3.2 we describe the model specifications that we employ as well as the adopted estimation approach. The comparison among the different model specifications and the results of the estimations, including relevant policy implications, are presented in section 3.3. Finally, section 3.4 concludes.

3.2 Model specification and estimation method

3.2.1 The classical parametric approach

The standard parametric specification *à la* CH/CHH can be expressed as:

$$\log f_{it} = \alpha_i + \theta \log S^d_{it} + \gamma \log S^f_{it} + \delta \log H_{it} + e_{it}, \tag{3.1}$$

where e_{it} is the error term, f_{it} is the TFP of country $i = 1, ..., N$ at time $t = 1, ..., T$; α_i are individual fixed effects, S^d_{it} and S^f_{it} are domestic and foreign R&D capital stocks, respectively; H_{it} is a measure of human capital. Foreign capital stock S^f_{it} is defined as the weighted arithmetic mean of S^d_{jt} for $j \neq i$, that is $S^f_{it} = \sum_{j \neq i} \omega_{ij} S^d_{jt}$, where ω_{ij} represents the weighting scheme. We adopt the same definition proposed by Lichtenberg and van Pottelsberghe de la Potterie (1998), which has been previously adopted in many other papers (Coe et al., 2009; Lee, 2006; Ertur and Musolesi, 2017), incorporating information on bilateral imports.

All the existing literature adopts parametric specifications that are variants of (3.1). Most of the previous studies follow some of the advances in panel time series econometrics over the last two decades. In particular, given the large T dimension of our panel, the likely existence of nonstationarity and cross-sectional dependence (Lee, 2006; Kao et al., 1999; Ertur and Musolesi, 2017) has been investigated.[1] Recently, Ertur and Musolesi (2017) highlight the presence of strong cross-sectional dependence in the data. Further, they use unit roots tests decomposing the panel into deterministic, common and idiosyncratic components (see, e.g. Bai and Ng, 2004) to identify the source of possible nonstationarity. They finally find that the series under investigation are nonstationary and that this property relies on the existence of nonstationary unobserved common factors rather than on idiosyncratic components. Under this scenario, Kapetanios et al. (2011), provide both analytical results and a simulation study according to which the cross-sectional averages augmentation by Pesaran (2006) remains valid.

In the following, for ease of exposition, we employ the notation:

$$y_{it} = \alpha_i + \beta' \mathbf{x}_{it} + e_{it}, \tag{3.2}$$

where $y_{it} = \log f_{it}$, $\mathbf{x}_{it} = [\log S^d_{it}, \log S^f_{it}, \log H_{it}]'$ and $\beta = [\theta, \gamma, \delta]'$.

[1] Another issue, which is out of the scope of this study, questions the homogeneity of the parameters implicit in the use of a pooled estimator in favor of heterogeneous regressions.

3.2.2 A nonparametric model with a multifactor error structure

We adopt the method proposed by Su and Jin (2012), who consider a panel data model that extends the multifactor linear specification proposed by Pesaran (2006). Specifically, Su and Jin (2012) consider the following model, which allows for a nonparametric relation between the dependent variable and the regressors, while the common factors enter the model parametrically:

$$y_{it} = \alpha_i^{'}\mathbf{d}_t + g\left(\mathbf{x}_{it}\right) + e_{it}, \tag{3.3}$$

where $\mathbf{d}_t$ is an $l\times 1$ vector of observed common effects, α_i is the associated vector of parameters and $\mathbf{x}_{it}$ is defined above. The *"one-way"* fixed effect specification is obtained by simply setting $\mathbf{d}_t = 1$. g is an unknown function to be estimated. For identification purposes, the condition $E(g\left(\mathbf{x}_{it}\right)) = 0$ is imposed. The errors e_{it} have a multifactor structure that is described by:

$$e_{it} = \gamma_i^{'}\mathbf{f}_t + \varepsilon_{it}, \tag{3.4}$$

where $\mathbf{f}_t$ is an $m \times 1$ vector of unobserved common factors with country-specific factor loadings γ_i. Combining (3.4) and (3.3), we obtain the following:

$$y_{it} = \alpha_i^{'}\mathbf{d}_t + g\left(\mathbf{x}_{it}\right) + \gamma_i^{'}\mathbf{f}_t + \varepsilon_{it}. \tag{3.5}$$

The idiosyncratic errors ε_{it} are assumed to be independently distributed over $(\mathbf{d}_t, \mathbf{x}_{it})$, whereas the unobserved factors $\mathbf{f}_t$ can be correlated with the observed variables $(\mathbf{d}_t, \mathbf{x}_{it})$. This correlation is allowed by modeling the explanatory variables as linear functions of the observed common factors $\mathbf{d}_t$ and the unobserved common factors $\mathbf{f}_t$:

$$\mathbf{x}_{it} = \mathbf{A}_i'\mathbf{d}_t + \mathbf{\Gamma}_i'\mathbf{f}_t + \mathbf{v}_{it}, \tag{3.6}$$

where $\mathbf{A}_i$ and $\mathbf{\Gamma}_i$ are $l\times 3$ and $m\times 3$ factor loading matrices, and $\mathbf{v}_{it} = (v_{i1t}, v_{i2t}, v_{i3t})'$. Following Pesaran (2006), Su and Jin (2012) proxy the unobservable factors $\mathbf{f}_t$ in (3.5) by the cross-sectional averages $\overline{\mathbf{z}}_t = N^{-1}\sum_{j=1}^{N}\mathbf{z}_{jt}$, where $\mathbf{z}_{it} = [y_{it}, \mathbf{x}_{it}']'$. They estimate the nonparametric part of the model using sieves. It is worth noting that the most common examples of sieve regression are polynomial series expansions and splines.

3.2.3 Alternative specifications

Consider (3.5) for $\mathbf{d}_t = 1$, that is $y_{it} = \alpha_i + g\left(\mathbf{x}_{it}\right) + \gamma_i^{'}\mathbf{f}_t + \varepsilon_{it}$. We are interested in three different specifications. As a benchmark, the parametric specification is obtained for $g\left(\mathbf{x}_{it}\right) = \beta^{'}\mathbf{x}_{it}$. The estimation is performed applying the common correlated effects pooled (CCEP) approach by Pesaran (2006). Then, we consider two specifications where $\mathbf{x}_{it}$ enter the model nonparametrically. The first specification assumes an additive structure of g, as follows:

$$\log f_{it} = \alpha_i + \phi(\log S_{it}^d) + \xi(\log S_{it}^f) + \psi(\log H_{it}) + \gamma_i^{'}\mathbf{f}_t + \varepsilon_{it}, \tag{3.7}$$

where ϕ, ξ and ψ are unknown univariate smooth functions of interest. The second specification assumes a non-additive structure of g, particularly:

$$\log f_{it} = \alpha_i + g(\log S_{it}^d, \log S_{it}^f, \log H_{it}) + \gamma_i^{'} \mathbf{f}_t + \varepsilon_{it}. \tag{3.8}$$

3.2.4 Spline modeling

Su and Jin (2012) estimate the nonparametric component of the model using sieves, and particularly splines, as they typically provide better approximations (see, *e.g.*, Hansen, 2014). Following Su and Jin (2012), we adopt a regression splines (RS) framework. We also employ penalized regression splines (PRS), as they combine the features of both regression splines, which use less knots than data points but do not penalize roughness, and smoothing splines, which control the smoothness of the fit through a penalty term but use all data points as knots. PRS have proven to be useful empirically in many aspects (see, e.g. Ruppert et al., 2003) and, in recent years, their asymptotic properties have been studied and then connected to those of regression splines, to those of smoothing splines and to the Nadaraya - Watson kernel estimators (Claeskens et al., 2009; Li and Ruppert, 2008).

Specifically, for both RS and PRS, we use thin plate regression splines (TPRS), which are a low rank eigen-approximation to thin plate splines. Thin plate splines are somehow ideal smoothers (see Wood, 2017) but are not computationally attractive because they require the estimation of as many parameters as the number of data points. TPRS avoid the problem of knot placement that usually complicates modeling with RS or PRS and more generally have some optimality properties, as they provide optimal low rank approximations to thin-plate splines, while they also are computationally efficient (see Wood, 2003). Since our explanatory variables have different units, in the case of the non-additive specification (3.8), we avoid isotropy by considering a tensor product basis, which is constructed by assigning TPRS as the basis for the marginal smooth of each covariate and then creating their Kronecker product. The tensor product smooths are invariant to the linear rescaling of covariates, and for this reason, they are appropriate when the arguments of a smooth have different units (Wood, 2006). Finally note that in the PRS framework, the smoothing parameter is selected by the restricted maximum likelihood (REML) estimation, which, relative to other approaches, is less likely to develop multiple minima or to undersmooth at finite sample sizes (see, e.g. Reiss and Todd Ogden, 2009).[2]

3.3 Results

3.3.1 Model comparison

To compare the aforementioned specifications, we perform a pseudo Monte Carlo experiment. In particular, along the lines depicted by Racine and Parmeter (2014), Ma et al. (2015) and Delgado et al. (2014), using similar macro panel data variables related to economic growth, the observations are randomly shuffled at 90% into training points and at 10% into evaluation

[2] The nonparametric specifications are estimated by the R package mgcv.

points. Each model is fitted according to the training sample. Then, the average out-of-sample squared prediction error (ASPE) is computed using the evaluation sample. The above steps are repeated a large number of times $B = 1000$, so that a $B \times 1$ vector of prediction errors is created for each model.[3]

The method is linked to cross validation (CV), in the original formulation of which a regression model fitted on a randomly selected first half of the data was used to predict the second half. The division into equal halves is not necessary. For instance, a common variant is the leave-one-out CV, which fits the model to the data excluding one observation each time and then predicts the remaining point. The average of the prediction errors is the CV measure of the error. As highlighted in Racine and Parmeter (2014), the method can provide significant power improvements over existing single-split techniques.

Figure 3.1 presents the box-and-whisker plots of the ASPE distributions for the different specifications. A first relevant result is that the median that corresponds to the parametric model is the largest among the different specifications, while the non-additive penalized model has the smallest median. In particular, the median ASPEs of the non-additive penalized model relative to the other models – the parametric, the additive unpenalized, the additive penalized and the non-additive unpenalized – is 0.6023, 0.9284, 0.9409 and 0.8278, respectively. A second interesting result is that the penalized regression modeling has a smaller median ASPE than its unpenalized counterpart for both additive and non-additive specifications. However, although when imposing an additive structure, the two approaches provide quite similar performances, the gain in terms of predictive ability from using PRS over RS is extremely pronounced when estimating the non-additive specification, which typically suffers more from the curse of dimensionality problem. Also, it is worth noting that within the RS framework, the additive specification provides a better performance than the non-additive one.

Next, figure 3.2 shows the empirical distribution functions of the ASPEs for each model. Clearly, the ASPE of the non-additive penalized model is stochastically dominated by the ASPE of any of the remaining models. This indicates that the non-additive penalized model outperforms all others in terms of predictive ability. It is also evident that the parametric model underperforms with respect to the nonparametric ones.

Finally, we compare the different specifications using the test of revealed performance (TRP) proposed by Racine and Parmeter (2014).[4] The results of these paired t-tests are presented in Table 1. In all cases, the null hypothesis that the difference in means of the ASPEs is zero is rejected. Thus, the tests complement the above presented results, indicating that this difference is statistically significant in all cases.

[3] See also Baltagi et al. (2003) who contrast the out-of-sample forecast performance of alternative parametric panel data estimators.

[4] The TRP involves estimating the distribution of the true errors for the different models and testing whether their expectations are statistically different. The true error is associated with out-of-sample measures of fit, contrasted to the apparent error, which is associated with within sample measures. Typically, the latter is smaller than the former and frequently overly optimistic (see e.g. Efron, 1982).

3.3.2 Estimation results

In this subsection, we present the main estimation results and specifically focus attention on the nonparametric specifications. We only consider PRS, since they outperform their unpenalized counterparts. We first provide the results obtained using the additive specification (3.7) because, due to the additive structure, the results are directly comparable to those ones of the parametric specifications adopted in previous studies. Then, we present the results of the non-additive specification (3.8), which, according to our findings, provides the best performance. Specifically, we focus on the interaction between domestic and foreign R&D.

The results concerning the nonparametric part of the additive penalized specification are presented in figure 3.3. The three graphs depict the estimated univariate smooth functions, which all appear to be highly significant, with extremely low p-values associated with the Wald test (Wood, 2012) that the function equals zero. It is worth mentioning that because the response as well as the explanatory variables are in logs, the slope of the estimated smooth functions represents the estimated elasticity. The first plot shows the effect of domestic R&D on TFP. It appears that for low values of R&D, where data are sparse and large confidence interval bands are present, the relation is flat. Then, for intermediate values of domestic R&D, the function is monotonic increasing, with a steep rise in approximately the last two deciles. The policy implications resulting from the above are clear: an increase in domestic R&D has an effect on productivity only above a threshold, thus suggesting that a critical mass of investments in R&D is crucial for R&D to become effective. After this threshold, the estimated output elasticity becomes positive and increases even more for very high levels of domestic R&D. This can be seen as a refinement of the results of the existing empirical literature on R&D spillovers, which is based on parametric models and generally distinguishes between G7 and non-G7 countries. Indeed, Ertur and Musolesi (2017), employing the CCE approach, show that the estimated output elasticity of domestic R&D is positive and significant for G7 countries, while it is non-significant for non-G7 countries. Similar results are also found by Coe et al. (2009), who adopt the dynamic OLS for cointegrated panels, and by Barrio-Castro et al. (2002), who use a standard fixed effects approach.

The second graph shows the effect of foreign R&D on TFP. Again, for low levels of the variable, data are scarce, making it difficult to identify a clear pattern. Then, the relation is positive and roughly concave for intermediate values, while it becomes flat for high levels of foreign R&D. The results show that an increase in foreign R&D affects TFP positively, but only up to a certain level. They complement previous empirical literature such as Coe et al. (2009), who indicate that trade-related foreign R&D is a significant determinant of TFP. More specifically, our findings improve the results of Ertur and Musolesi (2017), among others, who find a small, positive and significant effect of R&D on TFP in non-G7 countries, but no significant effect in the case of the G7. Nevertheless, in all previous studies, the linearity assumption obscures the fact that the output elasticity of foreign R&D is not constant but varies with respect to the different levels of foreign R&D. Indeed, looking at the bottom panel of figure 3.3, it can be seen that the estimated elasticity constantly decreases over the range of foreign R&D up to a level where it becomes not significantly different from zero.

The third graph in figure 3.3 depicts the effect of human capital on TFP. It again shows scarce data and large confidence bands for low levels. Then, the relation between human capital and TFP is approximately flat for intermediate values, while for high values, it seems to be monotonic increasing, with a steep rise in approximately the last two deciles. In terms of policy perspectives, the results suggest a threshold that occurs at very high levels of human capital, above which the estimated elasticity becomes positive. Investing in human capital becomes effective only after a certain level is reached. These findings add new insights to Ertur and Musolesi (2017), who find no significant effect of human capital on TFP for both G7 and non-G7 countries and explain their result on the grounds that the quantity of education no longer has a significant effect when omitted variable bias is addressed. We find confirmation of such results for most of the domain of human capital, but we also show that allowing for nonlinearity in the relation between human capital and TFP is crucial in order to highlight a positive effect for the highest levels of human capital.

Next, we turn to the estimates of the non-additive specification. Also, in this case, the estimated (multivariate) smooth function appears to be highly significant. In particular, we focus on the effect of the interaction between domestic and foreign R&D on TFP. The results are presented in figure 3.4, which shows the impact on TFP for a level of human capital fixed to the first, fifth (the median) and ninth decile. As depicted in the first graph, for low levels of human capital and irrespective of the level of domestic R&D, foreign R&D has almost no effect on TFP. In terms of policy implications, these findings suggest that foreign R&D spillovers cannot be effective if the level of human capital in a country remains low. Moreover, the effect of domestic R&D on TFP seems not to be linked to the level of foreign R&D, which implies an additive pattern when the level of human capital is low. Similar to the additive model presented above, there is a threshold above which domestic R&D becomes effective.

The second and third graphs in figure 3.4 show the effect on TFP when human capital is fixed to the median and to the ninth decile, respectively. The results in both graphs suggest a complementarity between domestic R&D and foreign R&D. For low levels of domestic R&D, the effect of foreign R&D on TFP is low, and vice versa. Domestic R&D becomes more effective when the levels of both domestic and foreign R&D are increasing. This is also true for foreign R&D. These findings have interesting policy implications; in countries with intermediate or high levels of human capital, investments in R&D are not very effective if the level of foreign R&D is low. Further, the benefits of foreign R&D spillovers cannot be exploited unless both human capital and domestic R&D are above a critical mass. The above results contrast with results from some previous studies such as in Coe et al. (2009), who report that their estimations considering interactions between human capital and domestic and foreign R&D do not yield correctly signed and significant results.

3.4 Concluding remarks

This paper revisits the analysis of international technology diffusion by adopting the approach proposed by Su and Jin (2012), which extends the multifactor linear specification proposed by

Pesaran (2006) to nonparametric specifications. We first show that a shift from a parametric to a nonparametric framework provides a significant improvement in terms of predictive ability. Moreover, it is also documented that penalized regression splines perform significantly better than their unpenalized counterparts, especially in the case of a non-additive model. Turning to the estimation results, our findings suggest the presence of threshold effects and nonlinearities. Then, the estimation of a non-additive specification provides further insights into the interactions among explanatory variables without imposing any parametric restrictions and definitively indicating that a critical mass of human capital is necessary to benefit from R&D spillovers and to observe an interactive effect between domestic and foreign R&D. In general, our findings strongly highlight that the presence of nonlinearities and complex interactions is an important feature of the data; these are obviously hidden within a parametric framework and have relevant implications for policy. Finally, it is worth mentioning that a further extension of the present study may account for heterogeneity across countries. This work is outside the realm of the nonparametric estimations presented in this paper and could be accomplished, for instance, by resorting to Bayesian modeling (Kiefer and Racine, 2017) to address the curse of dimensionality problem raised by heterogeneity.

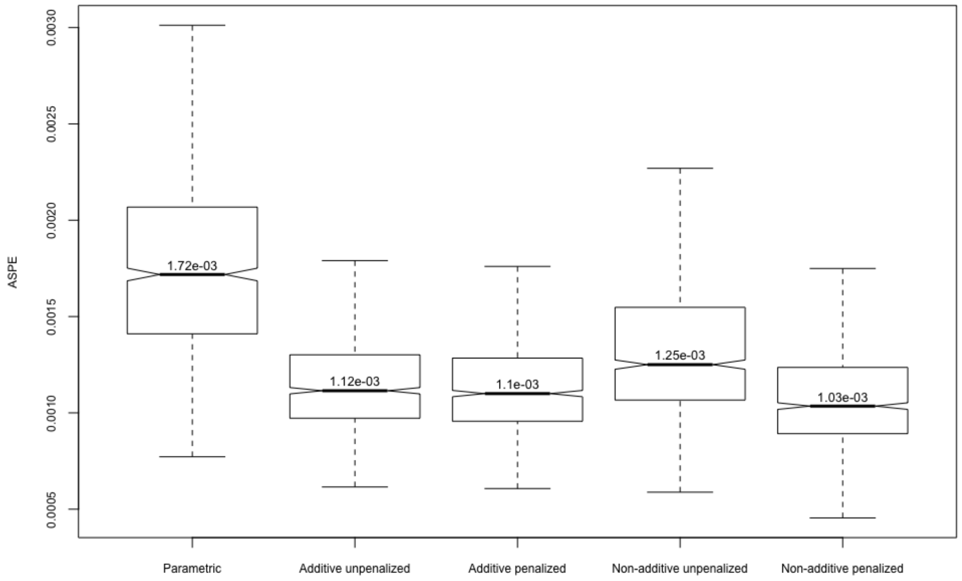


Figure 3.1: Out-of-sample average square prediction error (ASPE) box plots for different factor models: the parametric, the additive and the non-additive.

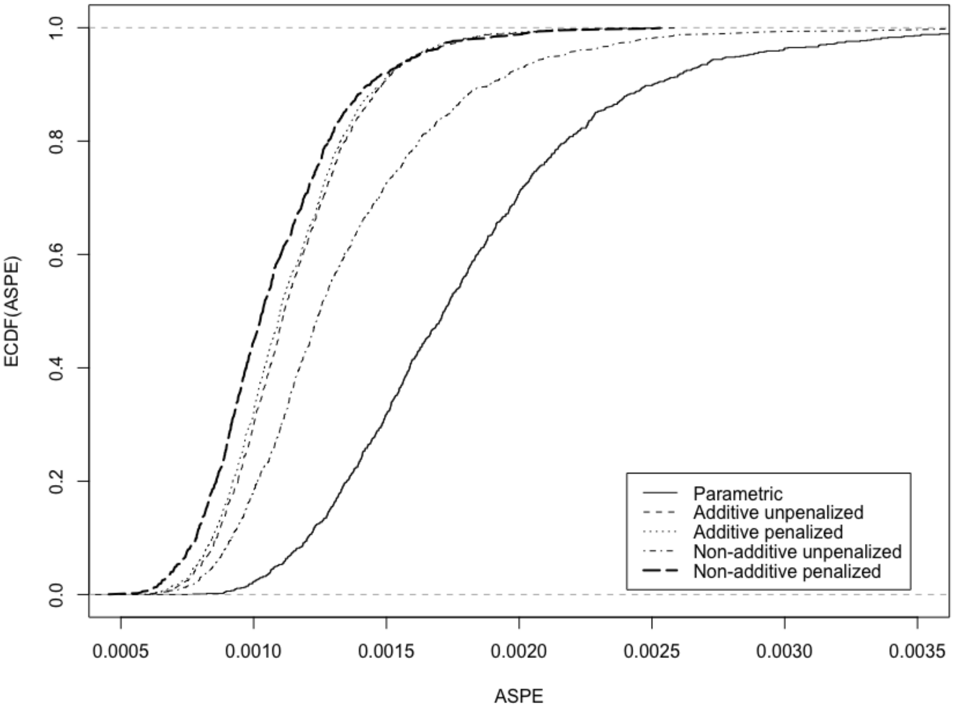


Figure 3.2: Empirical Cumulative Distribution Functions (ECDFs) of the ASPE for different factor models: the linear, the additive and the non-additive models for the OECD data.

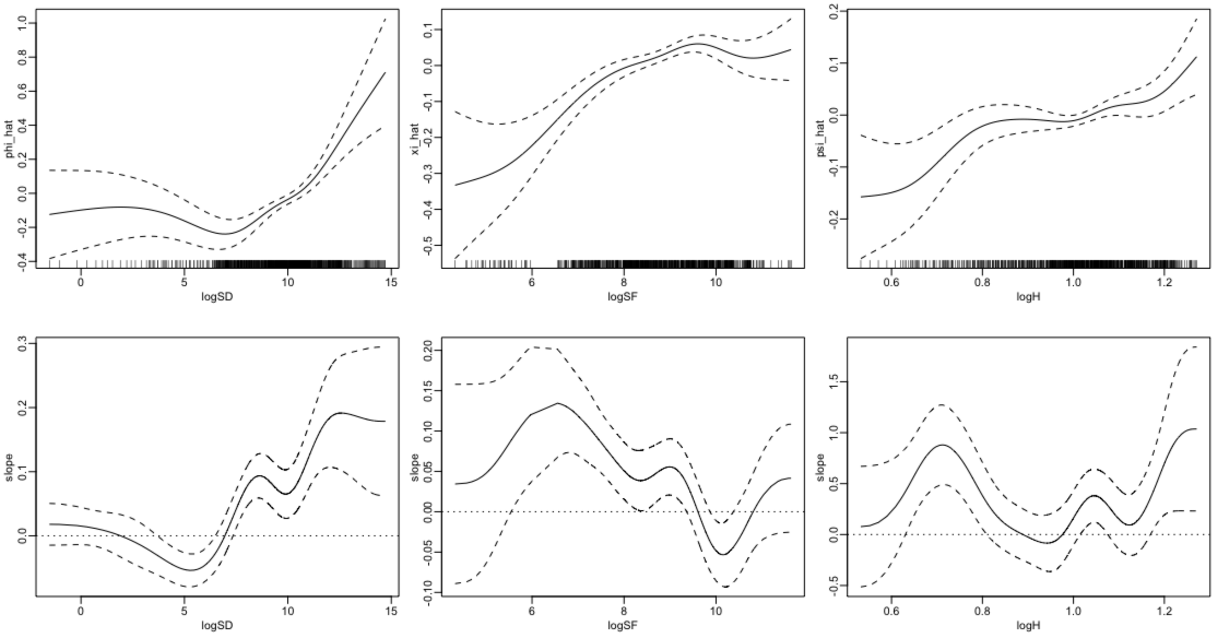
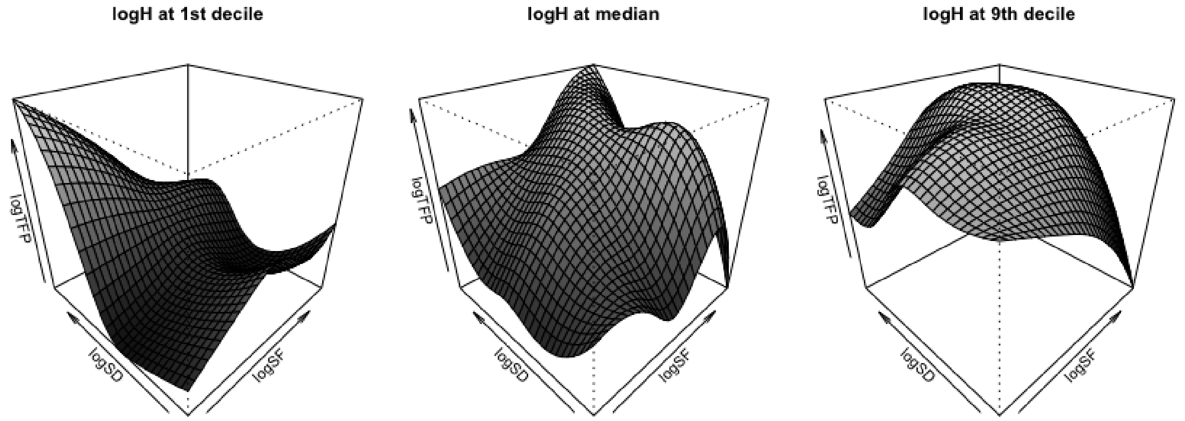


Figure 3.3: Additive Model. Estimated smooths (top panel) and corresponding derivatives (bottom panel) for the additive penalized regression model. Component smooths are shown with confidence intervals obtained by computing a Bayesian posterior covariance matrix.


logH at 1st decile

logH at median

logH at 9th decile

Figure 3.4: Non-additive model. The effect of domestic and foreign R&D on TFP for different levels of human capital. The log of human capital is fixed to the first, fifth and ninth decile, respectively.

Table 1 - Paired t-tests of factor models

models	Additive unpenalized	Additive penalized	Non-additive unpenalized	Non-additive penalized
Parametric	43.683***	45.461***	27.042***	47.992***
Additive unpenalized		9.849***	-18.493***	13.138***
Additive penalized			-20.492***	10.697***
Non-additive unpenalized				32.642***

Null hypothesis: The true difference in means of the ASPEs of the compared models is zero.
The training sample is 90% of the data-sample; number of resampling iterations B: 1.000
Classification of p-value: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Chapter 4

Review in the nonparametric kernel regression

4.1 Nonparametric kernel regression

The use of regression models is widespread in economic literature. Assuming additivity of the error term u, these models can be described by the following equation:

$$Y = f\left(\mathbf{Z}\right) + u.$$

The expected value of a dependent variable Y is modeled given a vector of explanatory variables $\mathbf{Z}$. f is the function of interest. There are two general approaches to the problem. The first treats f as a function of known shape with unknown parameters $\boldsymbol{\beta}$. The aim is then to estimate $\boldsymbol{\beta}$. As an example, consider the well-known linear regression model which is described by:

$$Y = \mathbf{Z}^T\boldsymbol{\beta} + u.$$

In the above expression, we have imposed a linear relation between $\mathbf{Z}$ and Y. $\boldsymbol{\beta}$ is a vector of unknown parameters that is estimated using methods such as Ordinary Least Squares (OLS). Nevertheless, if the model is not correctly specified, then the parameter estimates are not consistent and inference is not possible. Moreover, tests for the misspecification of the presumed parametric model do not give information about the correct one. An alternative approach which does not impose any functional form to the model is nonparametric regression.

4.1.1 Nonparametric kernel regression with continuous data

Consider the following regression model:

$$Y = f(\mathbf{Z}) + u, \tag{4.1}$$

where f is a function of unknown form to be estimated. $\mathbf{Z}$ is a q-dimensional vector of explanatory variables. We assume that $\mathbf{Z}$ is a continuous random vector. In section 4.1.2 this assumption is relaxed towards cases of mixed categorical and continuous regressors. Consider the case of an i.i.d. sample $(Y_i, \mathbf{Z}_i),\ \ i = 1, 2, \ldots, n$. If $\mathbf{Z}$ is exogenous, that is $E(u|\mathbf{Z}) = 0$, then by taking the mean conditional upon $\mathbf{Z}$ in both sides of (4.1) results in:

$$E(Y|\mathbf{Z} = \mathbf{z}) = f(\mathbf{z}). \tag{4.2}$$

Therefore, we can estimate f by computing the conditional mean of Y. It is proven (see Li and Racine, 2007, p.59) that $E(Y|\mathbf{Z} = \mathbf{z})$ is the function that, among all Borel measurable functions, minimizes the Mean Squared Error (MSE) of Y. This is stated in the following theorem:

Theorem 1 *Let $\mathscr{G}$ be the class of Borel measurable (or continuous) functions having finite second moment. Assume that $f(\mathbf{z}) = E(Y|\mathbf{Z} = \mathbf{z})$ belongs to $\mathscr{G}$ and that $E(Y^2) < \infty$. Then, $E(Y|\mathbf{Z} = \mathbf{z})$ is the optimal predictor of Y given $\mathbf{Z}$, according to the following inequality:*

$$E\{[Y - g(\mathbf{Z})]^2\} \geq E\{[Y - E(Y|\mathbf{Z} = \mathbf{z})]^2\}, \ \ \textit{for all} \ \ g \in \mathscr{G} \tag{4.3}$$

Nadaraya (1964) and Watson (1964) are the first to propose an estimation of $E(Y|\mathbf{Z})$ according to the following steps. By definition, the conditional mean of Y is:

$$E(Y|\mathbf{Z}=\mathbf{z}) \overset{\text{def}}{=} \int y\phi_{Y|Z}(y|\mathbf{z})dy = \int y\frac{\phi_{Y,Z}(y,\mathbf{z})}{\phi(\mathbf{z})}dy = \frac{1}{\phi(\mathbf{z})}\int y\phi_{Y,Z}(y,\mathbf{z})dy. \tag{4.4}$$

ϕ denotes the marginal distribution of $\mathbf{Z}$, while $\phi_{Y,Z}$ and $\phi_{Y|Z}$ are the joint distribution of $(Y, \mathbf{Z})$ and the conditional distribution of Y given $\mathbf{Z}$ respectively. Based on (4.2), an estimator of f is derived by replacing ϕ and $\phi_{Y,Z}$ in (4.4) with their kernel estimations. Therefore, the estimator $\hat{f}$ is described as follows:

$$\hat{f}(\mathbf{z}) = \frac{1}{\hat{\phi}(\mathbf{z})}\int y\hat{\phi}_{Y,Z}(y,\mathbf{z})dy. \tag{4.5}$$

The kernel estimator of ϕ is expressed as:

$$\hat{\phi}(\mathbf{z}) = \frac{n^{-1}}{h_1 h_2 \dots h_q}\sum_{i=1}^{n} K(h_Z, \mathbf{Z}_i, \mathbf{z}). \tag{4.6}$$

$K(h_Z, \mathbf{Z}_i, \mathbf{z})$ is the product kernel at $\mathbf{Z}_i$ which is defined in Li and Racine (2003) and in Racine and Li (2004). It is constructed by defining a kernel function k for each of the discrete or continuous explanatory variables and then multiplying these functions together to produce the product kernel. Let Z_{is} and z_s, $s = 1, \dots, q$ be the s-th component of $\mathbf{Z}_i$ and $\mathbf{z}$ respectively. Denote by $\mathbf{h}_Z = (h_1, h_2, \dots, h_q)$ the vector of the smoothing parameters h_s, $s = 1, \dots, q$ that correspond to the Z_s variables. Let $k(h_s, Z_{is}, z_s) = k\left(\frac{Z_{is}-z_s}{h_s}\right)$ is the kernel function associated with Z_{is}. Then:

$$K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}) = \prod_{s=1}^{q} k(h_s, Z_{is}, z_s). \tag{4.7}$$

Similarly, the kernel estimator of $\phi_{Y,Z}$ is given by:

$$\hat{\phi}_{Y,Z}(y,\mathbf{z}) = \frac{n^{-1}}{h_Y h_1 \dots h_q}\sum_{i=1}^{n} K(\mathbf{h}_V, \mathbf{V}_i, \boldsymbol{\nu}). \tag{4.8}$$

$\mathbf{h}_V = (h_Y, h_1, \dots, h_q)$ is the vector of smoothing parameters corresponding to $\mathbf{V} = (Y, \mathbf{Z})$ and h_Y is the smoothing parameter of Y. Let $k(h_Y, Y_i, y) = k\left(\frac{Y_i - y}{h_Y}\right)$ be the kernel function of Y. Then $K(\mathbf{h}_V, \mathbf{V}_i, \boldsymbol{\nu})$ is the product kernel at $\mathbf{V}_i = (Y_i, \mathbf{Z}_i)$ and is constructed as the product of $K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})$ and $k(h_Y, Y_i, y)$. Therefore:

$$K(\mathbf{h}_V, \mathbf{V}_i, \boldsymbol{\nu}) = k(h_Y, Y_i, y)K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}). \tag{4.9}$$

As implied in (4.5), an estimator for $\int y\phi_{Y,Z}(y,\mathbf{z})dy$ is derived by replacing $\phi_{Y,Z}$ with $\hat{\phi}_{Y,Z}$. Indeed, substituting $\hat{\phi}_{Y,Z}$ with the expression in (4.8) results in:

$$\int y\hat{\phi}_{Y,Z}(y,\mathbf{z})dy = \frac{n^{-1}}{h_Y h_1 \dots h_q}\int y \sum_{i=1}^{n} K(\mathbf{h}_V, \mathbf{V}_i, \boldsymbol{\nu})dy =$$

$$\frac{n^{-1}}{h_Y h_1 \dots h_q}\int y \sum_{i=1}^{n}(k(h_Y, Y_i, y)\cdot K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}))dy =$$

$$\frac{n^{-1}}{h_Y h_1 \dots h_q}\sum_{i=1}^{n}\left(K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})\int yk(h_Y, Y_i, y)dy\right).$$

By construction, it holds that:

$$\frac{1}{h_Y}\int yk(h_Y, Y_i, y)dy = Y_i.$$

Therefore,

$$\int y\hat{\phi}_{Y,Z}(y,\mathbf{z})dy = \frac{n^{-1}}{h_1 \dots h_q}\sum_{i=1}^{n} K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})Y_i. \tag{4.10}$$

Substituting the numerator and the denominator in (4.5) with the expressions in (4.6) and (4.10) respectively results in the Nadaraya - Watson estimator:

$$\hat{f}(\mathbf{z}) = \frac{\sum_{i=1}^{n} K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})Y_i}{\sum_{i=1}^{n} K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})}. \tag{4.11}$$

Define:

$$w_i(\mathbf{z}) = \frac{K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})}{\sum_{i=1}^{n} K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})}. \tag{4.12}$$

Then, (4.11) is expressed as:

$$\hat{f}(\mathbf{z}) = \sum_{i=1}^{n} w_i(\mathbf{z})Y_i.$$

$w_i(\mathbf{z}),\ i = 1, \dots, n$ are functions that for a given $\mathbf{z}$, $\sum_{i=1}^{n} w_i(\mathbf{z}) = 1$ and $w_i(\mathbf{z}) \geq 0,\ \forall i$. Therefore, the Nadaraya - Watson estimator is simply a weighted average of $Y_i,\ i = 1, \dots, n$. (4.11) also implies that $\hat{f}(\mathbf{z})$ is a local average of the sample values of Y. Indeed, for a kernel function k, it holds that $k(h_s, Z_{is}, z_s) \to 0$ as $Z_{is} - z_s \to \pm\infty$. The value of $w_i(\mathbf{z})$ is large only if Z_{is} is close to z_s for all $s = 1, \dots, q$. Otherwise, the kernel value is close to zero and the respective Y_is do not contribute to the computation of $\hat{f}(\mathbf{z})$. Therefore, the Nadaraya - Watson estimation of the conditional mean at a point $\mathbf{z}$ is a local averaging of the Y_is for which their corresponding $\mathbf{Z}_i$s are close to $\mathbf{z}$.

For illustration purposes, consider the simple case of one continuous explanatory variable Z and the use of a uniform kernel, i.e.

$$k\left(\frac{Z_i - z}{h_Z}\right) = \begin{cases} 1/2 & \text{if } |Z_i - z| < h_Z \\ 0 & \text{otherwise} \end{cases}.$$

Then,

$$\hat{f}(z) = \frac{\sum_{|Z_i - z| < h_Z} Y_i(1/2)}{\sum_{|Z_i - z| < h_Z}(1/2)} = \frac{\sum_{|Z_i - z| < h_Z} Y_i}{\sum_{|Z_i - z| < h_Z} 1}.$$

The above equation shows that for the computation of $\hat{f}$ at a point z, only the Z_i-s which are within h_Z distance from z are considered. The local information used for the computation of $\hat{f}(z)$ is controlled by the smoothing parameter. In the extremes, if h is small (close to zero), less sample points are considered for the estimation of f at z. On the other hand, if h is large, $\hat{f}(z)$ at any point z is just the average of Y_is, and Z becomes an irrelevant regressor of Y. According to the above, the value of h is important because it controls the number of observations which enter $\hat{f}$. For this reason, Bowman and Azzalini (1997) refer to the quantity nh as the local sample size. nh is a measure of the number of observations locally. As the sample size increases, that is as $n \to \infty$, the smoothing parameter should allow local averaging to become more informative by allowing $nh \to \infty$. At the same time, the bandwidth $h \to 0$ should shrink to zero for consistency of the estimation so that $\lim\limits_{n\to\infty} \hat{f}(z) = f(z)$.

The Nadaraya - Watson estimator is also known as the local constant estimator. To understand this, we estimate $f(\mathbf{z})$ by minimizing the kernel weighted least squares objective function which follows:

$$\min_c \sum_{i=1}^n (Y_i - c)^2 K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}). \tag{4.13}$$

Minimizing (4.13) results the local constant estimator $\tilde{c} = \tilde{f}(\mathbf{z})$. The first order condition of (4.13) requires that the first derivative with respect to c at point $\tilde{c}$ is zero, that is:

$$\frac{d}{dc}\left(\sum_{i=1}^n (Y_i - c)^2 K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})\right)\Bigg|_{c=\tilde{c}} = -2\sum_{i=1}^n (Y_i - \tilde{c}) K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}) = 0.$$

Solving this equation for $\tilde{c}$ gives:

$$\tilde{f}(\mathbf{z}) \equiv \tilde{c} = \frac{\sum_{i=1}^n K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}) Y_i}{\sum_{i=1}^n K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})}. \tag{4.14}$$

(4.14) is the same as (4.11). Therefore, the local constant estimator $\tilde{f}(\mathbf{z})$ is the Nadaraya - Watson estimator $\hat{f}(\mathbf{z})$.

The Nadaraya - Watson estimator in (4.11) requires a value for the smoothing parameter $\mathbf{h}_Z$. In the following we present three of the well-known approaches for bandwidth selection. We restrict ourselves to the case of continuous variables $\mathbf{Z} \in \mathbb{R}^q$ and to the Nadaraya - Watson estimation of the regression function. The more general case which accounts for both continuous and discrete variables is presented in section 4.1.2.

Plug-in method

The plug-in method minimizes a Weighted Integrated Mean Squared Error (WIMSE) function of the form $\int E[\hat{f}(\mathbf{z}) - f(\mathbf{z})]^2 v(\mathbf{z}) d\mathbf{z}$. $v(\mathbf{z})$ is a nonnegative weight function that ensures that, asymptotically, WIMSE is finite. The leading term of WIMSE is described by:

$$\int \left\{ \left[\sum_{s=1}^q h_s^2 B_s(\mathbf{z})\right]^2 + \frac{\kappa^q}{nh_1 \ldots h_q} \frac{\sigma^2(\mathbf{z})}{\phi(\mathbf{z})} \right\} v(\mathbf{z}) d\mathbf{z} = O\left(\left(\sum_{s=1}^q h_s^2\right)^2 + (nh_1 \ldots h_q)^{-1}\right). \tag{4.15}$$

In (4.15), $h_s,\ s = 1, \ldots, q$ is the smoothing parameter that corresponds to the s-th component of $\mathbf{Z}$. κ and σ are defined by $\kappa = \int k^2(x)dx$ and $\sigma^2(\mathbf{z}) = E(u_i^2|\mathbf{Z}_i = \mathbf{z})$ respectively. $B_s,\ s = 1, \ldots, q$ is defined by the equation:

$$B_s(\mathbf{z}) = \frac{\kappa_2}{2} \frac{2\phi_s(\mathbf{z}) f_s(\mathbf{z}) + \phi(\mathbf{z}) f_{ss}(\mathbf{z})}{\phi(\mathbf{z})}. \tag{4.16}$$

In this expression, $\kappa_2 = \int x^2 k(x)dx$ is the second moment of the kernel function k. Also, we denote by r_s and r_{ss} ($r = f$ or ϕ) the first and second derivative of $r(\mathbf{z})$ with respect to z_s.

Define α_s through $h_s = \alpha_s n^{-1/(q+4)},\ s = 1, \ldots, q$. Then, the leading term of WIMSE in (4.15) becomes:

$$n^{-4/(q+4)} \chi_v(\alpha_1, \ldots, \alpha_q), \tag{4.17}$$

where:

$$\chi_v(\alpha_1, \ldots, \alpha_q) = \int \left\{ \left[\sum_{s=1}^{q} \alpha_s^2 B_s(\mathbf{z}) \right]^2 + \frac{\kappa^q}{\alpha_1 \ldots \alpha_q} \frac{\sigma^2(\mathbf{z})}{\phi(\mathbf{z})} \right\} v(\mathbf{z}) d\mathbf{z}. \tag{4.18}$$

Denote by $\alpha_1^0, \ldots, \alpha_q^0$ the values of $\alpha_1, \ldots, \alpha_q$ that minimize $\chi_v(\alpha_1, \ldots, \alpha_q)$ and by $h_s^0,\ s = 1, \ldots, q$ the smoothing parameters that minimize (4.15). Obviously, it holds that:

$$h_s^0 = n^{-1/(q+4)} \alpha_s^0,\ s = 1, \ldots, q. \tag{4.19}$$

If one parameter α_k^0 is zero, then another one α_t^0, $t \neq k$ should equal infinity, otherwise (4.18) is not minimized. If $h_t^0 \to \infty$, then the kernel function $k(h_t^0, Z_{it}, z_t) = k(\frac{Z_{it} - z_t}{h_t^0}) \to k(0)$ becomes a constant and cancels out from the numerator and the denominator of (4.11). This means that the variable Z_t is irrelevant for the regression. We restrict ourselves to the cases where all regressors are relevant. This restriction is relaxed in section 4.1.2. Then, we assume that:

Assumption 1 *Each α_s^0, $s = 1, \ldots, q$ is uniquely defined, positive and finite.*

To obtain an estimator for α_s^0, one can substitute $B_s(\mathbf{z})$ in (4.18) with its estimator $\hat{B}_s(\mathbf{z})$, then compute the integration in χ_v and, at last, minimize the resulting expression with respect to $\alpha_1, \ldots, \alpha_q$. The plug-in method constructs an estimator of h_s^0 by plugging $\hat{\alpha}_s^0$ in (4.19), so that:

$$\hat{h}^0 = n^{-1/(q+4)} \hat{\alpha}^0,\ s = 1, \ldots, q. \tag{4.20}$$

The plug-in method requires an initial estimate of $B_s(\mathbf{z})$ and $\sigma^2(\mathbf{z})$. Therefore, an initial smoothing parameter should be selected. If this pilot bandwidth is far from h_s^0, then also the $\hat{h}_s^0$ may not be an accurate estimator of h_s^0. Moreover, if the assumption concerning α_s^0 does not hold, the plug-in method is not well defined.

Least Squares Cross Validation

Least Squares Cross Validation (LSCV) is a totally data driven method of bandwidth selection which relies in minimizing the following objective function:

$$CV_{lc}(h_1, \ldots, h_q) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_{-i}(\mathbf{Z}_i))^2 M(\mathbf{Z}_i). \tag{4.21}$$

In (4.21), $\hat{f}_{-i}(\mathbf{Z}_i)$ is the leave-one-out Nadaraya - Watson estimator, computed at $\mathbf{Z}_i$ and described by the expression:

$$\hat{f}_{-i}(\mathbf{Z}_i) = \frac{\sum\limits_{j \neq i}^{n} Y_i K(\mathbf{h}_Z, \mathbf{Z}_j, \mathbf{Z}_i)}{\sum\limits_{j \neq i}^{n} K(\mathbf{h}_Z, \mathbf{Z}_j, \mathbf{Z}_i)}.$$

$M(\mathbf{Z}_i)$ is a weight function which takes values between zero and one ($0 \leq M(\cdot) \leq 1$). It prevents difficulties caused by dividing by zero or by the slow convergence rate caused by boundary effects. (4.21) is expressed as follows (Hall et al., 2007, p.787):

$$CV_{lc}(h_1, \dots, h_q) = \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{Z}_i) - \hat{f}_{-i}(\mathbf{Z}_i))^2 M(\mathbf{Z}_i) + \frac{2}{N}\sum_{i=1}^{n} u_i(f(\mathbf{Z}_i) - \hat{f}_{-i}(\mathbf{Z}_i))M(\mathbf{Z}_i) + \frac{1}{n}\sum_{i=1}^{n} u_i^2 M(\mathbf{Z}_i). \tag{4.22}$$

u_i is the error term in (4.1) at $(Y_i, \mathbf{Z}_i)$. The third term of the right hand side of (4.22) is not related to $\mathbf{h}$ and, therefore, does not participate in the minimization of CV_{lc}. The second term is of order smaller than the first one. For this reason, minimizing CV_{lc} is asymptotically equivalent to minimizing the leading term $CV_{lc,0}$ of CV_{lc}:

$$CV_{lc,0} = 1/n\sum_{i=1}^{n}(f(\mathbf{Z}_i) - \hat{f}_{-i}(\mathbf{Z}_i))^2 M(\mathbf{Z}_i).$$

Moreover, it holds that (see Li and Racine, 2007, p.99):

$$CV_{lc,0} = E[CV_{lc,0}] + (\text{terms of smaller order}).$$

Therefore, minimizing CV_{lc} is equivalent to minimizing $E[CV_{lc,0}]$, which is expressed as follows (see Li and Racine, 2007, p.100):

$$E[CV_{lc,0}(h_1, \dots, h_q)] = n^{-4/(q+4)}\chi(\alpha_1, \dots, \alpha_q) + o(n^{-4/(q+4)}). \tag{4.23}$$

$\alpha_s,\ s = 1, \dots, q$ are defined by $h_s = \alpha_s n^{-1/(q+4)}$. χ is described by the equation (see Li and Racine, 2007, p.69):

$$\chi(\alpha_1, \dots, \alpha_q) = \int \left\{\sum_{s=1}^{q} B_s(\mathbf{z})\alpha_s^2\right\}^2 \phi(\mathbf{z})M(\mathbf{z})d\mathbf{z} + \frac{\kappa^q}{\alpha_1 \dots \alpha_q}\int \sigma^2(\mathbf{z})M(\mathbf{z})d\mathbf{z}, \tag{4.24}$$

where B_s, κ and σ are defined previously.

A comparison between (4.18) and (4.24) shows that if $v(\mathbf{z}) = f(\mathbf{z})M(\mathbf{z})$, then $\chi_v = \chi$.

Indeed, it holds that:

$$\begin{aligned} E[CV_{lc,0}] &= E\{[\hat{f}_{-i}(\mathbf{Z}_i) - f(\mathbf{Z}_i)]^2 M(\mathbf{Z}_i)\} = E\{E\{[\hat{f}_{-i}(\mathbf{Z}_i) - f(\mathbf{Z}_i)]^2 M(\mathbf{Z}_i)]|\mathbf{Z}_i\}\} \\ &= \int E\{[\hat{f}_{-i}(\mathbf{Z}_i) - f(\mathbf{Z}_i)]^2 M(\mathbf{Z}_i)|\mathbf{Z}_i = \mathbf{z}\}\phi(\mathbf{z})d\mathbf{z} \\ &= \int E\{[\hat{f}_{-i}(\mathbf{Z}_i) - f(\mathbf{Z}_i)]^2 \phi(\mathbf{z})M(\mathbf{z})d\mathbf{z}, \end{aligned}$$

which is the same as WIMSE for $v(\mathbf{z}) = \phi(\mathbf{z})M(\mathbf{z})$.

Let $\alpha_1^0, \ldots, \alpha_q^0$ be the values of $\alpha_1, \ldots, \alpha_q$ that minimize χ. As in the plug-in method, we assume that $\alpha_s^0,\ s = 1, \ldots, q$ are uniquely defined, positive and finite. Denote by h_s^0, $s = 1, \ldots, q$ the values of h_s that minimize the leading term of $E[CV_{lc,0}(\mathbf{h})]$ in (4.23) and by $\hat{h}_s$, $s = 1, \ldots, q$ the values that minimize the cross validation objective function in (4.21). Racine and Li (2004) show that:

$$\hat{h}_s = h_s^0 + o_p(h_s^0), \quad s = 1, \ldots, q.$$

Therefore, through cross validation the computed $\hat{h}_s$ is asymptotically equal to h_s^0. The above discussion shows that the idea behind LSCV is straightforward; instead of minimizing χ, cross validation minimizes (4.21) which is fully data driven and requires just a standard optimization process. This is in contrast to the plug-in method, which minimizes χ_v. The asymptotic properties of LSCV are summarized in the following theorem (for details, see Li and Racine, 2007, p.70).

Theorem 2 *Assume that:*

1. $\alpha_s^0,\ s = 1, \ldots, q$ *are uniquely defined, positive and finite*
2. f, ϕ *and* σ^2 *have two continuous derivatives*
3. w*, as defined in eq.(4.12), is continuous, nonnegative and has compact support*
4. ϕ *is bounded away from zero for* $\mathbf{z}$ *in the support of* w

Then, the following limits hold:

$$n^{1/(q+4)}\hat{h}_s \xrightarrow{P} \alpha_s^0, \quad s = 1, \ldots, q$$

$$n^{4/(q+4)} \left\{ CV_{lc}(\hat{h}_1, \ldots, \hat{h}_q) - n^{-1} \sum_{i=1}^{N} u_i^2 M(Z_i) \right\} \xrightarrow{P} \inf_{\alpha_1, \ldots, \alpha_q} \chi(\alpha_1, \ldots, \alpha_q).$$

Akaike Information Criterion

Hurvich et al. (1998) propose a bandwidth selection method based on the Akaike Information Criterion. The optimization criterion is given by:

$$AIC_c = ln(\hat{\sigma}^2) + \frac{1 + tr(\mathbf{Q})/n}{1 - \{tr(\mathbf{Q}) + 2\}/n}, \tag{4.26}$$

where:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{f}(\mathbf{Z}_i)\}^2 = \frac{1}{n}\mathbf{Y}^T(\mathbf{I}-\mathbf{Q})^T(\mathbf{I}-\mathbf{Q})\mathbf{Y}.$$

$\mathbf{Y}$ is the $n \times 1$ vector of Y_is. $\mathbf{I}$ is the $n \times n$ identity matrix. $\mathbf{Q}$ is a $n \times n$ matrix with its (i,j)-th element given by:

$$w_{ij} = \frac{K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{Z}_j)}{\sum_{l=1}^{n} K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{Z}_l)}.$$

Li and Racine (2004) show that this approach appears to be asymptotically equivalent to the cross validation method and has good finite sample performance.

4.1.2 Nonparametric kernel regression with mixed data

In section 4.1.1 we consider the case of only continuous variables. Nevertheless, in practice, mixed data are common. In this section, the nonparametric regression accounts for both categorical and continuous regressors. Moreover, assumption 1 that all explanatory variables are relevant is relaxed; the presence of irrelevant regressors is considered.

Let $\mathbf{D}$ be the $d \times 1$ vector of discrete regressors and $\mathbf{C}$ the $c \times 1$ vector of continuous ones. $D_s,\ s = 1,\ldots,d$ and $C_s,\ s = 1,\ldots,c$ denote the s-th component of $\mathbf{D}$ and $\mathbf{C}$ respectively. Assume that $\mathbf{C} \in \mathbb{R}^c$ and $\mathbf{D} \in \prod_{s=1}^{d}\{0,1,\ldots,\kappa_s - 1\}$. κ_s is the number of different values that D_s can take. Denote also by $\mathbf{Z} = (\mathbf{D},\mathbf{C})$ the vector of $q = d + c$ explanatory variables and $\mathbf{z} = (\textsc{d},\textsc{c})$ a respective point at the range of $\mathbf{Z}$.

Moreover, let $\bar{\mathbf{D}}$ and $\tilde{\mathbf{D}}$ be the vectors of $\bar{d}$ relevant and $\tilde{d}$ irrelevant categorical regressors respectively. Similarly, denote by $\bar{\mathbf{C}}$ and $\tilde{\mathbf{C}}$ the vectors of $\bar{c}$ relevant and $\tilde{c}$ irrelevant continuous regressors. Obviously, $\bar{d} + \tilde{d} = d$ and $\bar{c} + \tilde{c} = c$. We assume that we do not know in advance $\tilde{d}$ and $\tilde{c}$. Without loss of generality, we rearrange variables so that the first $\bar{d}$ variables of the $\mathbf{D}$ vector and the first $\bar{c}$ variables of the $\mathbf{C}$ vector be the relevant ones. The relevant and irrelevant variables are defined according to the following assumption.

Assumption 2 (Relevance of regressors) *The random vector* $(Y, \bar{\mathbf{Z}})$ *is independent of* $\tilde{\mathbf{Z}}$*, where* $\bar{\mathbf{Z}} = (\bar{\mathbf{D}}, \bar{\mathbf{C}})$ *and* $\tilde{\mathbf{Z}} = (\tilde{\mathbf{D}}, \tilde{\mathbf{C}})$*.*

Assumption 2 is quite limitative; irrelevant variables must be independent not only from Y, but also from the relevant variables $\bar{\mathbf{Z}}$. In simulations, Hall et al. (2007) replace this condition with the less restrictive one:

$$E[Y|\mathbf{Z}] = E[Y|\bar{\mathbf{Z}}],\ \text{almost surely.} \tag{4.27}$$

Though not theoretically established, the results show that theorems 3 and 4 hold also under (4.27).

The regression model is given by (4.1). In the case of mixed data, the Nadaraya - Watson estimator of f is described by:

$$\hat{f}(\mathbf{z}) = \frac{\sum_{i=1}^{n}\Psi((\mathbf{h},\boldsymbol{\eta})_Z, \mathbf{Z}_i, \mathbf{z})Y_i}{\sum_{i=1}^{n}\Psi((\mathbf{h},\boldsymbol{\eta})_Z, \mathbf{Z}_i, \mathbf{z})}. \tag{4.28}$$

This expression is similar to (4.11). Unlike $K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})$ in (4.11), the product kernel in (4.28) accounts also for the presence of categorical regressors. $\Psi((\mathbf{h}, \boldsymbol{\eta})_Z, \mathbf{Z}_i, \mathbf{z})$ is constructed according to the following steps. A kernel function is assigned to each discrete variable D_s. Aitchison and Aitken (1976) present a kernel function for categorical unordered variables, which is described by the equation:

$$\tilde{l}(\eta_s, D_{is}, \mathsf{D}_s) = \begin{cases} \eta_s, & \text{if} \quad D_{is} = \mathsf{D}_s \\ \frac{1-\eta_s}{\kappa_s - 1}, & \text{if} \quad D_{is} \neq \mathsf{D}_s \end{cases},$$

where D_{is} and D_s are the sth components of $\mathbf{D}_i$ and D respectively. The bandwidth η_s takes values in $[1/\kappa_s, 1]$. Instead of the Aitkinson - Aitken kernel $\tilde{l}$, the following kernel function, proposed by Racine and Li (2004), can be used:

$$l(\eta_s, D_{is}, \mathsf{D}_s) = \begin{cases} 1, & \text{if} \quad D_{is} = \mathsf{D}_s \\ \eta_s, & \text{if} \quad D_{is} \neq \mathsf{D}_s \end{cases} \quad \text{(for unordered variables)}. \tag{4.29}$$

In the case of categorical ordered variables, the kernel function below can be used:

$$l(\eta_s, D_{is}, \mathsf{D}_s) = \eta_s^{|D_{is} - \mathsf{D}_s|} \quad \text{(for ordered variables)}. \tag{4.30}$$

In both cases of unordered or ordered data, $\eta_s,\ s = 1, \ldots, d$ is the smoothing parameter corresponding to D_s and is restricted in $[0, 1]$. In the extreme case where $\eta_s = 0$, the kernel function becomes a frequency estimator (indicator function) and the data are split into cells according to the categories. In the opposite extreme, if $\eta_s = 1$, the kernel function becomes a constant and the corresponding variable D_s is smoothed out of the regression. The product kernel L for the discrete variables is defined according to the following equation:

$$L(\boldsymbol{\eta}, \mathbf{D}_i, \mathsf{D}) = \prod_{s=1}^{d} l(\eta_s, D_{is}, \mathsf{D}_s). \tag{4.31}$$

We also assign a kernel function for each continuous regressor $C_s,\ s = 1, \ldots, c$. Denote by C_{is} and c_s the sth component of $\mathbf{C}_i$ and c respectively. The kernel function used is described by:

$$k(h_s, C_{is}, \mathsf{c}_s) = k\left(\frac{C_{is} - \mathsf{c}_s}{h_s}\right). \tag{4.32}$$

k is a univariate and symmetric density function, such as Gaussian, Epanechnikov or Biweight (for details, see Henderson and Parmeter, 2015, sec.2.4). Then, the product kernel corresponding to the continuous vector $\mathbf{C}$ is:

$$K(\mathbf{h}, \mathbf{C}_i, \mathsf{c}) = \prod_{s=1}^{c} k(h_s, C_{is}, \mathsf{c}_s). \tag{4.33}$$

The generalized product kernel Ψ in (4.28) is the product of the kernels in (4.31) and (4.33), that is:

$$\Psi((\mathbf{h}, \boldsymbol{\eta})_Z, \mathbf{Z}_i, \mathbf{z}) = L(\boldsymbol{\eta}, \mathbf{D}_i, \mathsf{D}) K(\mathbf{h}, \mathbf{C}_i, \mathsf{c}). \tag{4.34}$$

According to the above, the estimator of f in (4.28) is rewritten as:

$$\hat{f}(\mathbf{z}) = \frac{\sum\limits_{i=1}^{n}\left(Y_i \prod\limits_{s=1}^{\bar{c}} k(h_s, C_{is}, \mathsf{c}_s) \prod\limits_{s=\bar{c}+1}^{c} k(h_s, C_{is}, \mathsf{c}_s) \prod\limits_{s=1}^{\bar{d}} l(\eta_s, D_{is}, \mathsf{D}_s) \prod\limits_{s=\bar{d}+1}^{d} l(\eta_s, D_{is}, \mathsf{D}_s)\right)}{\sum\limits_{i=1}^{n}\left(\prod\limits_{s=1}^{\bar{c}} k(h_s, C_{is}, \mathsf{c}_s) \prod\limits_{s=\bar{c}+1}^{c} k(h_s, C_{is}, \mathsf{c}_s) \prod\limits_{s=1}^{\bar{d}} l(\eta_s, D_{is}, \mathsf{D}_s) \prod\limits_{s=\bar{d}+1}^{d} l(\eta_s, D_{is}, \mathsf{D}_s)\right)}. \quad (4.35)$$

The choice of $(\mathbf{h}, \boldsymbol{\eta})$ can be made by minimizing the following Cross Validation criterion:

$$CV_{lc}(\mathbf{h}, \boldsymbol{\eta}) = CV_{lc}(h_1, \ldots, h_c, \eta_1, \ldots, \eta_d) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{-i}(\mathbf{Z}_i))^2 M(\mathbf{Z}_i). \quad (4.36)$$

Similar to (4.21), $M(\mathbf{Z}_i)$ is a weight function. $\hat{f}_{-i}$ is the leave-one-out local constant kernel estimator of f, described by the equation:

$$\hat{f}_{-i}(\mathbf{Z}_i) = \frac{\sum\limits_{j \neq i}^{n} Y_j \Psi((\mathbf{h}, \boldsymbol{\eta})_Z, \mathbf{Z}_j, \mathbf{Z}_i)}{\sum\limits_{j \neq i}^{n} \Psi((\mathbf{h}, \boldsymbol{\eta})_Z, \mathbf{Z}_j, \mathbf{Z}_i)}. \quad (4.37)$$

The estimated values $(\hat{\mathbf{h}}, \hat{\boldsymbol{\eta}})$ that result from minimizing (4.36), enter (4.28) for the derivation of $\hat{f}(\mathbf{z})$. Hall et al. (2007) show that the leading term of (4.36) is the quantity below:

$$\begin{aligned}&\int \frac{\kappa^{\bar{c}}\sigma^2(\bar{\mathbf{z}})}{nh_1 \ldots h_{\bar{c}}} M(\mathbf{z})R(\mathbf{z})\tilde{\phi}(\tilde{\mathbf{z}})d\mathbf{z} + \int \left(\sum_{s=1}^{\bar{d}} \eta_s \sum_{\mathbf{v}} [I_s(\mathbf{v}, \bar{\mathsf{D}})\{\bar{f}(\bar{\mathsf{c}}, \mathbf{v}) - \bar{f}(\bar{\mathbf{z}})\}\bar{\phi}(\bar{\mathsf{c}}, \mathbf{v})] + \right.\\ &\left. 1/2\kappa_2 \sum_{s=1}^{\bar{c}} h_s^2 \{\bar{f}_{ss}(\bar{\mathbf{z}})\bar{\phi}(\bar{\mathbf{z}}) + 2\bar{f}_s(\bar{\mathbf{z}})\bar{\phi}_s(\bar{\mathbf{z}})\}\right)^2 \frac{\bar{M}(\bar{\mathbf{z}})}{\bar{\phi}(\bar{\mathbf{z}})} d\bar{\mathbf{z}},\end{aligned} \quad (4.38)$$

where $\kappa = \int k^2(x)dx$ and $\kappa_2 = \int x^2 k(x)dx$. $\bar{f}$ denotes a regression function of $\bar{\mathbf{Z}}$ on Y. $\bar{\phi}$ and $\tilde{\phi}$ denote the marginal densities of the relevant regressors $\bar{\mathbf{Z}}$ and the irrelevant regressors $\tilde{\mathbf{Z}}$ respectively. The subscripts s and ss in $\bar{f}$ and $\bar{\phi}$ denote first and second derivative with respect to the sth component of $\bar{\mathbf{z}}$. $\bar{M}$ is defined according to the equation $\bar{M}(\bar{\mathbf{z}}) = \int \tilde{\phi}(\tilde{\mathbf{z}})M(\mathbf{z})d\tilde{\mathbf{z}}$. $\mathbf{v}$ is a point in the range of the relevant, discrete variables, while I_s is an indicator function which is defined according to the expression:

$$I_s(\mathbf{x}, \mathbf{w}) = I(x_s \neq w_s)\prod_{t \neq s}^{\bar{d}} I(x_t = w_t),$$

for $\mathbf{x}, \mathbf{w} \in \prod_{s=1}^{\bar{d}}\{0, 1, \ldots, \kappa_s - 1\}$. Therefore, $I_s(\mathbf{x}, \mathbf{w})$ equals unity when only the sth components of $\mathbf{x}$ and $\mathbf{w}$ are different. Moreover, R is given by the expression:

$$R(\mathbf{z}) = \frac{v_2(\mathbf{z})}{v_1^2(\mathbf{z})},$$

where $v_j(\mathbf{z}),\ j = 1, 2$ is defined as follows (see Hall et al., 2007):

$$v_j(\mathbf{z}) = E\left(\left[\prod_{s=\bar{c}+1}^{c} h_s^{-1} k(\frac{C_{is} - \mathsf{c}_s}{h_s}) \prod_{s=\bar{d}+1}^{d} \eta_s^{I(D_{is} \neq \mathsf{D}_s)}\right]^j\right).$$

By use of Holder's inequality, it is proven that:

$$R(\mathbf{z}) \geq 1,\ \forall \mathbf{z}, h_{\bar{c}+1}, \ldots, h_c \text{ and } \forall \eta_{\bar{d}+1}, \ldots, \eta_d.$$

Obviously, the first term in (4.38) is minimized when $R(\mathbf{z})$ becomes unity. Thus, the minimization of the cross validation objective function in (4.36) requires $R(\mathbf{z})$ to be unity. It holds that $R \to 1$ if and only if the bandwidths of the irrelevant regressors approach their extreme upper values, that is if $h_s \to \infty$ and $\eta_t \to 1$, for $s \in \{\bar{c}+1, \ldots, c\}$ and $t \in \{\bar{d}+1, \ldots, d\}$. Therefore, minimization of the local constant cross validation function leads to a choice of extreme values for the smoothing parameters of the irrelevant regressors. In other words:

$$\lim_{\eta_s \to 1} l(\eta_s, D_{is}, \mathsf{D}_s) = 1,\ \text{for } \bar{d}+1 \leq s \leq d \text{ and} \tag{4.39a}$$

$$\lim_{h_s \to \infty} k(h_s, C_{is}, \mathsf{c}_s) = \lim_{h_s \to \infty} k\left(\frac{C_{is} - \mathsf{c}_s}{h_s}\right) = k(0),\ \text{for } \bar{c}+1 \leq s \leq c. \tag{4.39b}$$

By use of (4.39), the estimator of the regression function in (4.35) becomes:

$$\hat{f}(\mathbf{z}) = \frac{\sum\limits_{i=1}^{n}\left(Y_i \prod\limits_{s=1}^{\bar{c}} k(h_s, C_{is}, \mathsf{c}_s) \prod\limits_{s=1}^{\bar{d}} l(\eta_s, D_{is}, \mathsf{D}_s)\right)}{\sum\limits_{i=1}^{n}\left(\prod\limits_{s=1}^{\bar{c}} k(h_s, C_{is}, \mathsf{c}_s) \prod\limits_{s=1}^{\bar{d}} l(\eta_s, D_{is}, \mathsf{D}_s)\right)}. \tag{4.40}$$

Therefore, using the local constant cross validation method, only the relevant regressors enter the estimation of f. LSCV and the local constant estimator detect and smooth out the variables which are irrelevant to the regression. The estimation using the local constant cross validation method is the same as in the case of only relevant variables in the model:

$$Y = f(\bar{\mathbf{Z}}) + u. \tag{4.41}$$

For $R \to 1$, the terms described in eq.(4.38) become:

$$\begin{aligned}&\int \frac{\kappa^{\bar{c}}\sigma^2(\bar{\mathbf{z}})}{nh_1 \ldots h_{\bar{c}}} \bar{M}(\bar{\mathbf{z}}) d\bar{\mathbf{z}} + \int \left(\sum_{s=1}^{\bar{d}} \eta_s \sum_{\mathbf{v}} [I_s(\mathbf{v}, \bar{\mathbf{D}}) \{\bar{f}(\bar{\mathbf{c}}, \mathbf{v}) - \bar{f}(\bar{\mathbf{z}})\} \bar{\phi}(\bar{\mathbf{c}}, \mathbf{v})] + \right. \\ &\left. 1/2\kappa_2 \sum_{s=1}^{\bar{c}} h_s^2 \{\bar{f}_{ss}(\bar{\mathbf{z}}) \bar{\phi}(\bar{\mathbf{z}}) + 2\bar{f}_s(\bar{\mathbf{z}}) \bar{\phi}_s(\bar{\mathbf{z}})\}\right)^2 \frac{\bar{M}(\bar{\mathbf{z}})}{\bar{\phi}(\bar{\mathbf{z}})} d\bar{\mathbf{z}}.\end{aligned} \tag{4.42}$$

Define $\alpha_s = h_s n^{1/(\bar{c}+4)},\ s = 1, \ldots, \bar{c}$ and $\beta_s = \eta_s n^{2/(\bar{c}+4)},\ s = 1, \ldots, \bar{d}$ (Hall et al., 2007). Then, the leading term of CV as expressed in (4.42) is rewritten as:

$$n^{-4/(\bar{c}+4)}\bar{\chi}(\alpha_1,\ldots,\alpha_{\bar{c}},\beta_1,\ldots,\beta_{\bar{d}}),$$

where $\bar{\chi}$ is given by:

$$\begin{aligned}&\bar{\chi}(\alpha_1,\ldots,\alpha_{\bar{c}},\beta_1,\ldots,\beta_{\bar{d}}) = \\ &\int \frac{\kappa^{\bar{c}}\sigma^2(\bar{\mathbf{z}})}{\alpha_1\ldots\alpha_{\bar{c}}}\bar{M}(\bar{\mathbf{z}})d\bar{\mathbf{z}} + \int \left(\sum_{s=1}^{\bar{d}}\beta_s\sum_{\mathbf{v}}[I_s(\mathbf{v},\bar{\mathbf{D}})\{\bar{f}(\bar{\mathbf{c}},\mathbf{v})-\bar{f}(\bar{\mathbf{z}})\}\bar{\phi}(\bar{\mathbf{c}},\mathbf{v})]+\right. \\ &\left. 1/2\kappa_2\sum_{s=1}^{\bar{c}}\alpha_s^2\{\bar{f}_{ss}(\bar{\mathbf{z}})\bar{\phi}(\bar{\mathbf{z}})+2\bar{f}_s(\bar{\mathbf{z}})\bar{\phi}_s(\bar{\mathbf{z}})\}\right)^2\frac{\bar{M}(\bar{\mathbf{z}})}{\bar{\phi}(\bar{\mathbf{z}})}d\bar{\mathbf{z}}.\end{aligned} \tag{4.43}$$

Denote by $\alpha_1^0,\ldots,\alpha_{\bar{c}}^0,\beta_1^0,\ldots,\beta_{\bar{d}}^0$ the values of $\alpha_1,\ldots,\alpha_{\bar{c}},\beta_1,\ldots,\beta_{\bar{d}}$ in (4.43) that minimize $\bar{\chi}$. Then, the properties of LSCV are summarized in the following theorem (Hall et al., 2007):

Theorem 3 *Let $\hat{h}_1,\ldots,\hat{h}_c,\hat{\eta}_1,\ldots,\hat{\eta}_d$ be the values of $h_1,\ldots,h_c,\eta_1,\ldots,\eta_d$ that minimize eq.(4.36). Under assumption 2 and assumptions 4, 5, 6 and 7, presented at the end of the Appendix, the following asymptotic properties hold:*

1. *$n^{1/(\bar{c}+4)}\hat{h}_s \to \alpha_s^0$ in probability, for all $s = 1,\ldots,\bar{c}$.*
2. *$P(\hat{h}_s > C) \to 1$ for all $s = \bar{c}+1,\ldots,c$ and for all $C > 0$.*
3. *$n^{2/(\bar{c}+4)}\hat{\eta}_s \to \hat{\beta}_s^0$ in probability, for all $s = 1,\ldots,\bar{d}$.*
4. *$\hat{\eta}_s \to 1$ in probability, for all $s = \bar{d}+1,\ldots,d$.*

The bandwidths of the irrelevant variables diverge to their extreme values, while the smoothing parameters of the relevant regressors shrink to zero. These properties result to the dimensionality reduction of the model because the number of regressors diminishes from q in eq.(4.1) to $\bar{q} = \bar{c}+\bar{d}$ in eq.(4.41). Furthermore, the asymptotic normality of $\hat{f}$ is presented in the following theorem (Hall et al., 2007):

Theorem 4 *Suppose that assumption 2 and assumptions 4, 5, 6 and 7 hold. If $\mathbf{z}$ is an interior point with $\phi(\mathbf{z}) > 0$, then:*

$$(n\hat{h}_1,\ldots,\hat{h}_{\bar{c}})^{1/2}\left[\hat{f}(\mathbf{z})-\bar{f}(\bar{\mathbf{z}})-\sum_{s=1}^{\bar{d}}X_s(\bar{\mathbf{z}})\hat{\eta}_s^2-\sum_{s=1}^{\bar{c}}\Xi_s(\bar{\mathbf{z}})\hat{h}_s^2\right]\xrightarrow{d}N(0,\Omega(\bar{\mathbf{z}})), \tag{4.44}$$

where:

$$
\begin{aligned}
X_s(\bar{\mathbf{z}}) &= \sum_{\mathbf{v}} I_s(\mathbf{v}, \bar{\boldsymbol{D}}) \left(\bar{f}(\bar{\boldsymbol{C}}, \mathbf{v}) - \bar{f}(\bar{\mathbf{z}})\right) \bar{\phi}(\bar{\boldsymbol{C}}, \mathbf{v}) \bar{\phi}(\bar{\mathbf{z}})^{-1}, \\
\Xi_s(\bar{\mathbf{z}}) &= 1/2\kappa_2 \left(\bar{f}_{ss}(\bar{\mathbf{z}}) + 2\frac{\bar{\phi}_s(\bar{\mathbf{z}})\bar{f}_s(\bar{\mathbf{z}})}{\bar{\phi}(\bar{\mathbf{z}})} \right) \textit{ and} \\
\Omega(\bar{\mathbf{z}}) &= \kappa^{\bar{c}} \frac{\sigma^2(\bar{\mathbf{z}})}{\bar{\phi}(\bar{\mathbf{z}})}
\end{aligned}
$$

are quantities related to the asymptotic bias and variance.

From Theorem 4 the bias and the variance of $\hat{f}$ is calculated (see also Henderson and Parmeter, 2015, sec.5.3). For simplicity reasons we restrict ourselves to the case of only continuous variables. Then, bias in the local constant least squares (LCLS) estimation is given by the expression below:

$$
Bias[\hat{f}(\mathbf{z})] \approx \frac{\kappa_2}{2\phi(\mathbf{z})} \sum_{s=1}^{q} h_s^2 B_s(\mathbf{z}) = \kappa_2 \sum_{s=1}^{q} h_s^2 \frac{f_s(\mathbf{z})\phi_s(\mathbf{z})}{\phi(\mathbf{z})} + \frac{\kappa_2}{2} \sum_{s=1}^{q} h_s^2 f_{ss}(\mathbf{z}), \tag{4.46}
$$

where $B_s(\mathbf{z}) = 2f_s(\mathbf{z})\phi_s(\mathbf{z}) + f_{ss}(\mathbf{z})\phi(\mathbf{z})$. Bias is not directly dependent on the sample size n, but only indirectly through the smoothing parameters. Therefore, a larger sample size does not imply less bias. Moreover, it is proportional to the square of the bandwidth. The less the smoothing parameter, the less biased the estimation is. Furthermore, bias is dependent upon the distribution ϕ of the data. For this reason sparseness of the data also raises bias. From (4.46) it is obvious that if the underlying function f is constant (so that $f_s = f_{ss} = 0$) or if the underlying function is linear and the distribution is uniform (so that $f_{ss} = 0$ and $\phi_s = 0$), then $B_s = 0$ and the estimation is unbiased.

The variance of $\hat{f}$ is given by the following equation:

$$
Var[\hat{f}(\mathbf{z})] \approx \frac{\sigma^2 \kappa^q}{\phi(\mathbf{z})(nh_1 \dots h_q)}. \tag{4.47}
$$

Variance depends on σ^2 but not on f. Moreover, it is inversely proportional to n; the bigger the sample size, the less the variance of the estimation. Unlike bias, variance is inversely proportional to the bandwidth; the less the smoothing parameter, the more the variance of $\hat{f}$. Therefore, there is a trade-off between bias and variance, because if the bandwidth is large, variance is smaller but bias becomes large. Inversely, if the smoothing parameter is small, bias is small but variance is larger.

4.1.3 Local linear nonparametric regression

Fan and Gijbels (1996) highlight that the Nadaraya - Watson estimator has zero minimax efficiency. Moreover, as described in Hastie and Loader (1993) and in Chu and Marron (1991), this approach suffers from boundary bias and bias caused by the asymmetry of observations. Fan and Gijbels (1992) mention that local constant estimators converge more slowly at the boundary.

Both biases appearing in LCLS estimation are caused by the asymmetry of the observations around the point of interest, in combination with a non-zero slope of f at that point. For instance, in a boundary point $\mathbf{z}_b$, only the observations on one side of its neighborhood contribute to the estimation of $f(\mathbf{z}_b)$. If f has a nonzero slope at $\mathbf{z}_b$, then the estimation derived by (4.28) is biased (see Hastie and Loader, 1993). Approaches such as boundary kernel methods that handle the boundary problem in local constant estimation exist but are not efficient.

One of the well-known approaches for treating the boundary bias problem is the local linear estimator (Stone, 1977; Cleveland, 1979). As described by Fan (1993), the method is minimax efficient in correcting boundary bias. Li and Racine (2004) also show through simulations the efficiency gains of the local linear least squares (LLLS) estimator relative to the LCLS one.

Consider the case of continuous data, as described in section 4.1.1. Then, the local linear estimator of f in (4.1) for a fixed point $\mathbf{z}$ results from the solution of the following minimization problem:

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \alpha - (\mathbf{Z}_i - \mathbf{z})^T \boldsymbol{\beta})^2 K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}). \tag{4.48}$$

In (4.48), $K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z})$ is the product kernel as described in (4.7). $(\mathbf{Z}_i - \mathbf{z})$ is a q-dimensional vector. Denote by $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ the solution of (4.48). Then:

$$\hat{\alpha} = \hat{f}(\mathbf{z}),$$
$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{z}),$$

where $\boldsymbol{\beta}(\mathbf{z}) = \nabla f(\mathbf{z}) = \left(\frac{\partial f}{\partial z_1}(\mathbf{z}), \ldots, \frac{\partial f}{\partial z_q}(\mathbf{z})\right)^T$. Therefore, $\hat{\alpha}$ is the local linear estimator of f at $\mathbf{z}$ and $\hat{\boldsymbol{\beta}}$ is the vector of the local linear estimators of the derivatives of f at $\mathbf{z}$. Unlike the LCLS estimator, the local linear approach estimates automatically not only the regression function itself, but also its derivatives. Minimization of (4.48) results in the following estimators:

$$\begin{aligned}\hat{\gamma}(\mathbf{z}) &= \begin{pmatrix} \hat{f}(\mathbf{z}) \\ \hat{\boldsymbol{\beta}}(\mathbf{z}) \end{pmatrix} \\ &= \left[\sum_{i=1}^{n} K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}) \begin{pmatrix} 1 & (\mathbf{Z}_i - \mathbf{z})^T \\ \mathbf{Z}_i - \mathbf{z} & (\mathbf{Z}_i - \mathbf{z})(\mathbf{Z}_i - \mathbf{z})^T \end{pmatrix}\right]^{-1} \sum_{i=1}^{n} K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}) \begin{pmatrix} 1 \\ \mathbf{Z}_i - \mathbf{z} \end{pmatrix} Y_i.\end{aligned} \tag{4.50}$$

The smoothing parameter $\mathbf{h}_z$ in (4.50) is estimated by minimizing the following LSCV function:

$$CV_{ll}(h_1, \ldots, h_q) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{f}_{-i}(\mathbf{Z}_i)]^2. \tag{4.51}$$

Xia and Li (2002) use a weight function M to the cross validation objective function in order to reduce boundary effects. $\hat{f}_{-i}$ is the leave-one-out local linear estimator, given by $\hat{f}_{-i}(\mathbf{Z}_i) =$

$(1, 0, \ldots, 0) \cdot \hat{\boldsymbol{\gamma}}_{-i}(\mathbf{Z}_i)$, where:

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{-i}(\mathbf{Z}_i) &= \begin{pmatrix} \hat{f}_{-i}(\mathbf{Z}_i) \\ \hat{\boldsymbol{\beta}}_{-i}(\mathbf{Z}_i) \end{pmatrix} \\ &= \left[\sum_{j \neq i}^{n} K(\mathbf{h}_Z, \mathbf{Z}_j, \mathbf{Z}_i) \begin{pmatrix} 1 & (\mathbf{Z}_j - \mathbf{Z}_i)^T \\ \mathbf{Z}_j - \mathbf{Z}_i & (\mathbf{Z}_j - \mathbf{Z}_i)(\mathbf{Z}_j - \mathbf{Z}_i)^T \end{pmatrix} \right]^{-1} \times \\ &\quad \sum_{j \neq i}^{n} K(\mathbf{h}_Z, \mathbf{Z}_j, \mathbf{Z}_i) \begin{pmatrix} 1 \\ \mathbf{Z}_j - \mathbf{Z}_i \end{pmatrix} Y_j. \end{aligned} \tag{4.52}$$

In the more general case of mixed continuous and categorical data, the estimator is given by:

$$\begin{aligned} \hat{\boldsymbol{\gamma}}(\mathbf{z}) &= \begin{pmatrix} \hat{f}(\mathbf{z}) \\ \hat{\boldsymbol{\beta}}(\mathbf{z}) \end{pmatrix} \\ &= \left[\sum_{i=1}^{n} \Psi((\mathbf{h}, \boldsymbol{\lambda})_Z, \mathbf{Z}_i, \mathbf{z}) \begin{pmatrix} 1 & (\mathbf{C}_i - \mathbf{c})^T \\ \mathbf{C}_i - \mathbf{c} & (\mathbf{C}_i - \mathbf{c})(\mathbf{C}_i - \mathbf{c})^T \end{pmatrix} \right]^{-1} \times \\ &\quad \sum_{i=1}^{n} \Psi((\mathbf{h}, \boldsymbol{\lambda})_Z, \mathbf{Z}_i, \mathbf{z}) \begin{pmatrix} 1 \\ \mathbf{C}_i - \mathbf{c} \end{pmatrix} Y_i. \end{aligned} \tag{4.53}$$

The notation is defined in section 4.1.2. Ψ is the generalized product kernel as described in (4.34). (4.53) is a local linear estimator for the continuous variables $\mathbf{C}$ and a local constant estimator for the discrete variables $\mathbf{D}$. The smoothing parameters $(\mathbf{h}, \boldsymbol{\eta})_Z = (h_1, \ldots, h_c, \eta_1, \ldots, \eta_d)$ are selected by minimizing the following cross validation function:

$$CV_{ll}(h_1, \ldots, h_c, \eta_1, \ldots, \eta_d) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{f}_{-i}(\mathbf{Z}_i)]^2. \tag{4.54}$$

Again, $\hat{f}_{-i}$ is described by $\hat{f}_{-i}(\mathbf{Z}_i) = (1, 0, \ldots, 0) \cdot \hat{\boldsymbol{\gamma}}_{-i}(\mathbf{Z}_i)$, where:

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{-i}(\mathbf{Z}_i) &= \begin{pmatrix} \hat{f}_{-i}(\mathbf{Z}_i) \\ \hat{\boldsymbol{\beta}}_{-i}(\mathbf{Z}_i) \end{pmatrix} \\ &= \left[\sum_{j \neq i}^{n} \Psi((\mathbf{h}, \boldsymbol{\eta})_Z, \mathbf{Z}_j, \mathbf{z}) \begin{pmatrix} 1 & (\mathbf{Z}_j - \mathbf{Z}_i)^T \\ \mathbf{Z}_j - \mathbf{Z}_i & (\mathbf{Z}_j - \mathbf{Z}_i)(\mathbf{Z}_j - \mathbf{Z}_i)^T \end{pmatrix} \right]^{-1} \times \\ &\quad \sum_{j \neq i}^{n} \Psi((\mathbf{h}, \boldsymbol{\eta})_Z, \mathbf{Z}_j, \mathbf{Z}_i) \begin{pmatrix} 1 \\ \mathbf{Z}_j - \mathbf{Z}_i \end{pmatrix} Y_j. \end{aligned} \tag{4.55}$$

To establish the asymptotic properties of the local linear estimator, Li and Racine (2004) show that the leading term in (4.54) is:

$$CV_{ll,0} = \sum_{\mathbf{D}} \int \left\{ \frac{\kappa_2}{2} \sum_{s=1}^{c} f_{ss}(\mathbf{z}) h_s^2 + \sum_{s=1}^{d} X_s'(\mathbf{z}) \lambda_s \right\}^2 f(\mathbf{z}) d\mathbf{c} + \frac{\Omega'}{n h_1 \ldots h_c}, \tag{4.56}$$

where:

$$X'_s(\mathbf{z}) = \sum_{\mathbf{v}} [I_s(\mathbf{v}, \mathbf{D}) f(\mathbf{C}, \mathbf{v}) - f(\mathbf{z})]\phi(\mathbf{C}, \mathbf{v}) \text{ and}$$

$$\Omega' = \kappa^q \sum_{\mathbf{D}} \int \sigma^2(\mathbf{z}) d\mathbf{C}.$$

Define $\alpha_s,\ s = 1, \dots, c$ and $\beta_s,\ s = 1, \dots, d$ according to:

$$h_s = \alpha_s n^{-1/(c+4)},\ s = 1, \dots, c \tag{4.58a}$$

$$\eta_s = \beta_s n^{-2/(c+4)},\ s = 1, \dots, d \tag{4.58b}$$

Then, (4.56) can be rewritten as $CV_{ll,0} = \chi_{\alpha,\beta} n^{-4/(c+4)}$, where:

$$\chi_{\alpha,\beta} = \sum_{\mathbf{D}} \int \left\{ \frac{\kappa_2}{2} \sum_{s=1}^{c} f_{ss}(\mathbf{z}) \alpha_s^2 + \sum_{s=1}^{d} X'_s(\mathbf{z}) \beta_s \right\}^2 \phi(\mathbf{z}) d\mathbf{C} + \frac{\Omega'}{n\alpha_1 \dots \alpha_c}. \tag{4.59}$$

As in section 4.1.2, denote by $\alpha_1^0, \dots, \alpha_c^0, \beta_1^0, \dots, \beta_d^0$ the values of $\alpha_1, \dots, \alpha_c, \beta_1, \dots, \beta_d$ that minimize $\chi_{\alpha,\beta}$. Similarly, denote by $h_1^0, \dots, h_c^0, \eta_1^0, \dots, \eta_d^0$ the values of $h_1, \dots, h_c, \eta_1, \dots, \eta_d$ that minimize (4.56). Then, the convergence rates of the smoothing parameters are given by the following theorem (for details, see Li and Racine, 2004).

Theorem 5 *Suppose that assumptions 8, 9 and 10, described at the end of the Appendix, hold. Then:*

$$\frac{\hat{h}_s - h_s^0}{h_s^0} = O_p(n^{-\epsilon_1/(4+c)}),\ \forall s = 1, \dots, c \ \textit{ where } \epsilon_1 = \min\{c/2, 2\}$$

$$\hat{\eta}_s - \eta_s^0 = O_p(n^{-\epsilon_2}),\ \forall s = 1, \dots, d \ \textit{ where } \epsilon_2 = \min\{4/(4+c), 1/2\}$$

Theorem 5 proves that the smoothing parameters derived by the LLLS estimation converge to the optimal smoothing parameters. These convergence rates are the same as the ones presented by Racine and Li (2004) for the case of the LCLS estimation. Therefore, if the regression is non-linear, the bandwidths by LLLS and LCLS estimations converge to their optimal values with the same rate. Based on theorem 5, Li and Racine (2004) establish the asymptotic normality of the local linear estimator of f by the following theorem.

Theorem 6 *Under assumptions 8, 9 and 10, it holds that:*

$$\sqrt{n\hat{h}_1 \dots \hat{h}_c} \left(\hat{f}(\mathbf{z}) - f(\mathbf{z}) - \sum_{s=1}^{c} \frac{\kappa_2}{2} f_{ss}(\mathbf{z}) \hat{h}_s^2 - \sum_{s=1}^{d} \hat{\eta}_s X'_s(\mathbf{z}) \right) \xrightarrow{d} N(0, \Omega''),$$

$$\textit{where } \Omega'' = \kappa^c \sigma^2(\mathbf{z})/\phi(\mathbf{z}) \textit{ and } X'_s \textit{ is defined previously.} \tag{4.61}$$

Theorems 5 and 6 do not account for linear regressions and do not observe or smooth out irrelevant variables. Nevertheless, Li and Racine (2004) show through Monte Carlo simulations that when a continuous variable Z_s enters the regression linearly, then local linear least squares selects a large value for the bandwidth h_s of this variable. Moreover, in the presence of an irrelevant discrete variable, the method chooses a bandwidth η_s close to the upper extreme of its support. Therefore, though not theoretically established, the local linear cross validation (LLCV) method smooths out discrete variables which are not relevant to the regression. Further, a large value of h_s is chosen when the respective continuous regressor Z_s enters the regression linearly.

Theorem 6 provides the bias and variance of $\hat{f}$. For simplicity reasons we consider the case of only continuous data. Then, the bias of $\hat{f}(\mathbf{z})$ becomes:

$$Bias[\hat{f}(\mathbf{z})] \approx \frac{\kappa_2}{2} \sum_{s=1}^{q} h_s^2 f_{ss}(\mathbf{z}). \tag{4.62}$$

According to (4.62), bias is not dependent on the sample size n. Moreover, if the regression is linear to all its components, that is $f_{ss} = 0,\ \forall s = 1, \ldots, q$, then the estimator of f is unbiased. This is in contrast to the LCLS estimator, which generally remains biased when the underlying function is linear[1]. Contrary to the LCLS method, the bias of the LLLS estimator does not depend on the distribution $\phi(\mathbf{z})$ of $\mathbf{Z}$.

The variance of the local linear estimator is given by:

$$Var[\hat{f}(\mathbf{z})] \approx \frac{\sigma^2 \kappa^q}{\phi(\mathbf{z})(nh_1 \ldots h_q)}. \tag{4.63}$$

Obviously, the variance of $\hat{f}$ using the LLLS method is the same as the one of the LCLS estimation.

4.1.4 Local polynomial nonparametric regression

The local polynomial estimator in nonparametric regression is obtained by using a higher order polynomial in (4.13) or (4.48), instead of a constant or a linear function. It is a generalization of the Nadaraya - Watson and the local linear estimator. Consider the case of q continuous regressors. Then, for a fixed point $\mathbf{z}$, the local polynomial estimator $\hat{f}(\mathbf{z})$ of order p results from minimizing the following objective function (see Li and Racine, 2007, p.89):

$$\min_{\alpha_t} \sum_{i=1}^{n} \left\{ Y_i - \sum_{0 \leq \bar{t} \leq p} \alpha_t (\mathbf{Z}_i - \mathbf{z})^t \right\}^2 K(\mathbf{h}_Z, \mathbf{Z}_i, \mathbf{z}). \tag{4.64}$$

[1] For a comparison see eqs.(4.46) and (4.62).

(4.64) follows the notation introduced by Masry (1996), in particular:

$$
\begin{aligned}
t &\overset{\text{def}}{=} (t_1, \ldots, t_q), \\
\bar{t} &\overset{\text{def}}{=} \sum_{j=1}^{q} t_j, \\
z^t &\overset{\text{def}}{=} z_1^{t_1} \times \cdots \times z_q^{t_q} \text{ and} \\
\sum_{0 \leq \bar{t} \leq p} &\overset{\text{def}}{=} \sum_{\substack{j=1 \\ \bar{t}=j}}^{p} \sum_{t_1=0}^{j} \cdots \sum_{t_q=0}^{j}.
\end{aligned}
$$

α_t is a function of $\mathbf{z}$. For $p = 0$, expressions (4.13) and (4.64) are the same, while for $p = 1$ (4.64) is the same as (4.48). Denote by $\mathbb{T}$ the collection of all t so that $0 \leq \bar{t} \leq p$, that is $\mathbb{T} = \{t \in \mathbb{Z}^q : 0 \leq \bar{t} \leq p\}$. Assume that $\hat{\alpha}_t,\ t \in \mathbb{T}$ are the values of α_t at point $\mathbf{z}$ that minimize (4.64). Then, the local polynomial estimators of the regression function and its derivatives up to order p are given by:

$$d^t \hat{f}(\mathbf{z}) = t! \hat{\alpha}_t,\ \forall t \in \mathbb{T}, \tag{4.66}$$

where:

$$
\begin{aligned}
t! &\overset{\text{def}}{=} t_1! \times \cdots \times t_q! \text{ and} \\
(d^t f)(\mathbf{z}) &\overset{\text{def}}{=} \frac{\partial^t f(\mathbf{z})}{\partial z_1^{t_1} \ldots \partial z_q^{t_q}}.
\end{aligned}
$$

For t equal to the transpose of the zero vector **0**, (4.66) provides the following estimator of f at $\mathbf{z}$:

$$\hat{f}(\mathbf{z}) = \hat{\alpha}_{0^\intercal}.$$

In both (4.64) and (4.66), $\mathbf{h}_Z$ is the smoothing parameter, which is estimated by minimizing the LSCV function. The leave-one-out estimator of f is computed by:

$$\hat{f}_{-i}(\mathbf{Z}_i) = \hat{\alpha}_{-i,0^\intercal}$$

$\hat{\alpha}_{-i,t},\ t \in T$ are the values of $\alpha_{-i,t}$ that minimize:

$$\min_{\alpha_{-i,t}} \sum_{j \neq i}^{n} \left\{ Y_j - \sum_{0 \leq \bar{t} \leq p} \alpha_{-i,t} (\mathbf{Z}_j - \mathbf{Z}_i)^t \right\}^2 K(\mathbf{h}_Z, \mathbf{Z}_j, \mathbf{Z}_i).$$

The asymptotic normality of estimator $\hat{f}$ is established through the following requirements:

Assumption 3 *It is assumed that:*

1. *the regression function $f(\mathbf{z})$ is continuous with continuous derivatives up to the $(p+1)$ order.*

2. *k is a bounded second order kernel function with compact support.*

3. $h = O(n^{-1/(q+2p+2)})$

The theorem that follows describes the asymptotic properties of $\hat{f}$. For simplicity reasons, the bandwidths corresponding to the regressors are the same, i.e. $h_1 = h_2 = \cdots = h_q = h$.

Theorem 7 *Assume the conditions in 3. If $\phi(\mathbf{z}) > 0, \forall \mathbf{z}$ then:*

$$nh^{q+2l}\left[\nabla^{(l)}\hat{f}(\mathbf{z}) - \nabla^{(l)}f(\mathbf{z}) - A_{l,p+1}(\mathbf{z})h^{p+1-l}\right] \xrightarrow{d} N\left(0, \frac{\sigma^2(\mathbf{z})}{\phi(\mathbf{z})}V_l\right). \tag{4.68}$$

A_l and V_l are described in Masry (1996), as well as the following uniform almost sure convergence of the local polynomial least squares (LPLS) estimator:

Theorem 8 *For all $l \in \mathbb{Z} : 0 \leq l \leq p$, it holds that:*

$$\sup_{\mathbf{z}\in\mathscr{S}}\left|\nabla^{(l)}\hat{f}(\mathbf{z}) - \nabla^{(l)}f(\mathbf{z})\right| = O\left(\left(\frac{ln(n)}{nh^q + 2l}\right)^{1/2} + h^{p-l+1}\right) \text{ ,almost surely.} \tag{4.69}$$

An increase in the order of the local polynomial results in a smaller order of bias. On the other hand, the order of the polynomial has a negative effect with respect to the variance of the estimator. Fan and Gijbels (1996) describe the rise in variability of $\hat{f}$ as a result of the increase of p. Therefore, interest lies in the choice of the order p of the local polynomial. Hall and Racine (2015) propose a LSCV approach to determine the bandwidth as well as the order of the local polynomial. According to the proposed method, the cross validation function in (4.51) is minimized jointly with respect to the bandwidth $\mathbf{h}_Z = (h_1, \dots, h_q)^T$ and with respect to p. This approach is in contrast to previous ones which suggest as a choice of p the smallest integer above the order l of the derivative of f of interest.

4.2 Nonparametric kernel regression with instrumental variables

In section 4.1 exogeneity of $\mathbf{Z}$ is assumed. Nevertheless, if $\mathbf{Z}$ is endogenous, f in (4.1) can not be estimated by $E(Y|\mathbf{Z})$ because the error term is not independent of the vector of explanatory variables, that is $E(u|\mathbf{Z}) \neq 0$. One approach to this problem is using a vector $\mathbf{W} \in \mathbb{R}^\rho$ of instrumental variables so that $E(u|\mathbf{W}) = 0$. Then, f is estimated as the solution to the following equation:

$$E(Y - f(\mathbf{Z})|\mathbf{W}) = 0. \tag{4.70}$$

The solution f in (4.70) requires solving an inverse problem. Indeed, (4.70) can be rewritten as:

$$E(Y|\mathbf{W}) = E(f(\mathbf{Z})|\mathbf{W}).$$

Assume the following notation:

$$r : \mathbb{R}^\rho \to \mathbb{R}, r(\mathbf{W}) = E(Y|\mathbf{W}), \tag{4.71a}$$

$$T : \mathbb{L}^2(\mathbf{Z}) \to \mathbb{L}^2(\mathbf{W}), Th = E(h(\mathbf{Z})|\mathbf{W}), \tag{4.71b}$$

$$T^* : \mathbb{L}^2(\mathbf{W}) \to \mathbb{L}^2(\mathbf{Z}), T^*k = E(k(\mathbf{W})|\mathbf{Z}). \tag{4.71c}$$

$\mathbb{L}_2(\mathbf{Z})$ and $\mathbb{L}_2(\mathbf{W})$ are the sets of square integrable functions of $\mathbf{Z}$ and $\mathbf{W}$ respectively. T^* is the adjoint operator of T. Properties of the T operator are presented in the end of the Appendix. Then, (4.70) is equivalent to the expression:

$$r = Tf. \tag{4.72}$$

(4.71) and (4.72) show that f is the solution to a Fredholm integral equation of the first kind. In the Appendix an alternative derivation of (4.72) by Horowitz (2011) when $\mathbf{Z}$ and $\mathbf{W}$ are continuous is presented.

4.2.1 Identification of f

Identification of f is related to whether there is a unique solution to (4.72). f and f' both solve (4.72) if and only if:

$$E(f(\mathbf{Z}) - f'(\mathbf{Z})|\mathbf{W}) = 0.$$

For this reason, Newey and Powell (2003) and Newey (2013) highlight that f is identified, i.e. is the only solution to (4.72), if and only if $g(\mathbf{Z}) \equiv 0$ is the only solution to the equality:

$$E(g(\mathbf{Z})|\mathbf{W}) = 0.$$

Obviously, the identification of f and the completeness of the conditional expectation are equivalent. Indeed, according to the definition in D'Haultfoeuille (2011), $\mathbf{Z}$ is complete for $\mathbf{W}$ if, for all measurable real functions g such that $E(|g(\mathbf{Z})|) < \infty$:

$$E(g(\mathbf{Z})|\mathbf{W}) = 0 \text{ a.s.} \Rightarrow g(\mathbf{Z}) = 0 \text{ a.s.} \tag{4.73}$$

Moreover, completeness is equivalent to the injectivity of T, because expression (4.73) implies that the null space of the conditional mean operator $E(\cdot|\mathbf{W})$ is zero.

Canay et al. (2013) show that for nonparametric models and under commonly imposed restrictions, the null hypothesis that the completeness condition does not hold is not testable. Newey (2013) provides the following explanation which is based on the eigenvalues of the conditional expectation operator. Denote by $\{\lambda_i, \xi_i, i = 1, 2, \dots\}$ the eigenvalues λ_i and eigenvectors ξ_i of T. $\lambda_i,\ i = 1, 2, \dots$ are positive. f in (4.70) is not identified when at least one of the eigenvalues is zero. Nonetheless, as mentioned in Horowitz (2011), if T is assumed non-singular and the eigenvalues are sorted in descending order, then they have a unique limit point at 0. Therefore, the problem in testing completeness is that, empirically, models with zero eigenvalues and models with eigenvalues that approach zero cannot be distinguished.

The usual maintained identification assumption is that T is one-to-one or nonsingular (see Darolles et al., 2011; Horowitz, 2011, among others). T has an inverse T^{-1} and the solution f in (4.72) is given by the expression:

$$f = T^{-1}r. \tag{4.74}$$

4.2.2 The ill posed inverse problem

The definition of well posed and ill posed problems was first introduced by Hadamard (2014)[2]. A problem is well posed if it satisfies the following three conditions:

1. Existence of the solution

2. Uniqueness of the solution

3. Continuous dependence of the solution on the data (stability of the solution)

A problem is ill posed if it is not well posed, i.e. if one of the above properties does not hold.

Engl et al. (1996, ch.2) mention that the existence of a solution in (4.72) is equivalent to r being attainable for all $r \in \mathbb{L}^2(\mathbf{W})$. This means that r must be in the range of T. The second condition holds if and only if only zero belongs to the null space of T. This requirement holds assuming that T is one-to-one. Then, the inverse of T exists. If the first two conditions are satisfied, the third one holds if and only if T^{-1} is continuous.

The nonparametric IV estimation leads to an ill posed problem because the mapping from r to f in (4.74) is not continuous. Indeed, under the conditions of Picard theorem (Kress, 1999a, sec.15.4), the solution in (4.72) is given by:

$$f = T^{-1}r = \sum_{i=1}^{\infty} \frac{\langle r, \psi_i \rangle}{\lambda_i} f_i. \tag{4.75}$$

In (4.75), $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{L}^2$, defined as $\langle h_1, h_2 \rangle = \int h_1(z)h_2(z)dz$ for $h_1,\ h_2 \in \mathbb{L}^2$. $(\lambda_i, f_i, \psi_i), i = 1, 2, \dots$ are determined by the Singular Value Decomposition (SVD) of T. λ_i's are the eigenvalues of T and T^* while f_i's and ψ_i's are the corresponding eigenfunctions. (4.75) shows that because $\lambda_i \to 0$ as $i \to \infty$, f is not continuously dependent on the data. In practice, an estimation $\hat{r} = r + \delta\psi_i$ enters (4.72), the error level $\delta > 0$ being arbitrarily small. Darolles et al. (2011) mention that this approximation of r introduces an error which leads to a result $f + \frac{\delta}{\lambda_i} f_i$ infinitely far from the exact solution f. The large change in f due to an arbitrarily small change in r is also explained in Horowitz (2011).

4.2.3 Regularization

Because of the ill-posedness of the problem described in (4.72), a consistent estimator of f can not result from plugging in (4.74) consistent estimators of r and the conditional distribution $\phi(\mathbf{Z}|\mathbf{W})$ (see also Newey, 2013). Nonetheless, a sequence of bounded operators $R_\alpha : \mathbb{L}^2(\mathbf{W}) \to \mathbb{L}^2(\mathbf{Z}),\ \alpha > 0$ can be found to approximate the unbounded inverse operator T^{-1}. Then, the solution f in (4.72) is estimated by $\hat{f}_\alpha = R_\alpha \hat{r}$, $\alpha \to 0$. This approach is called regularization and describes the approximation of an ill-posed problem by a family of well-posed problems, such that their solutions converge to the solution of the ill-posed problem (see Kress, 1999a, sec.15.2). Kress (1999a) provides the following definition.

[2] for a definition see also Kress (1999a, ch.15).

Definition 9 *Let X and Y be normed spaces and $T : X \to Y$ a one-to-one, bounded, linear operator. A family of bounded linear operators $R_\alpha : Y \to X$, $\alpha > 0$, with the property of pointwise convergence*

$$\lim_{\alpha \to 0} R_\alpha T f = f, \;\; f \in X$$

is a regularization scheme for the operator T. α is called the regularization parameter.

The regularization scheme approximates the solution f in (4.72) by $\hat{f}_\alpha = R_\alpha \hat{r}$. Denote by δ the error level imposed by the estimation of r, so that $||\hat{r} - r|| \leq \delta$[3]. The total approximation error is given by the inequality:

$$||\hat{f}_\alpha - f|| \leq \delta ||R_\alpha|| + ||R_\alpha T f - f||. \tag{4.76}$$

As described in Kress (1999a), the first term of the right hand side in (4.76) is the sample error, which increases as $\alpha \to 0$. The second term is the error introduced by the approximation of T^{-1} by R_α and decreases as $\alpha \to 0$. Therefore, the regularization parameter α minimizes the overall error balancing between accuracy and stability. A small value of α corresponds to large sample error and a highly variable estimator. On the other hand, a large value of α results in an over-regularized estimator.

In practice, instead of directly regularizing the problem in (4.72), we reshape the equation as follows. r, T and T^* are defined in (4.71). It is assumed that the completeness condition that T is one-to-one holds and that both operators T and T^* are compact. Projecting (4.72) onto the space of $\mathbf{Z}$ (see, e.g. Centorrino et al., 2017), we obtain:

$$T^* r = T^* T f. \tag{4.77}$$

Denote by $\hat{T}^*$, $\hat{T}$ and $\hat{r}$ the kernel estimators of T^*, T and r respectively. Their smoothing parameters can be estimated according to one of the bandwidth selection methods discussed before. Then, the sample counterpart of (4.77) is:

$$\hat{T}^* \hat{r} = \hat{T}^* \hat{T} \hat{f}. \tag{4.78}$$

Darolles et al. (2011) mention that f is identifiable if and only if T^*T is one-to-one. $(T^*T)^{-1}$ exists but is noncontinuous. Therefore $\hat{f} = (\hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{r}$ is not well-defined. The regularization procedure is applied in (4.78) rather than in (4.72).

In literature a plethora of different ways to regularize exists, including Tikhonov methods, sieve estimations and truncated iterative approaches[4]. A brief description of two well-known regularization methods follows:

1. Tikhonov regularization (Tikhonov, 1943; Darolles et al., 2011)

2. Landweber - Fridman regularization (Landweber, 1951; Fridman, 1956)

[3] $|| \cdot ||$ is the *norm* in $\mathbb{L}^2$ so that for every $h \in \mathbb{L}^2$, $||h|| = \left[\int h^2(z)dz\right]^{1/2}$

[4] for example Kaczmarz iteration or Krylov subspace methods

Tikhonov regularization

The method adds a small positive constant term α to the eigenvalues of $\hat{T}^*\hat{T}$ to obtain a well-posed solution to the problem of (4.78). Therefore, the estimator of f is the solution to:

$$\hat{T}^*\hat{r} = \alpha\hat{f} + \hat{T}^*\hat{T}\hat{f}. \tag{4.79}$$

Denote this estimator by $\hat{f}^\alpha$. $\alpha I + T^*T$ is invertible for any nonnegative α, because T^*T is a self-adjoint nonnegative operator. Nevertheless, $\alpha I + \hat{T}^*\hat{T}$ may not be invertible, because $\hat{T}^*$ is not generally the adjoint operator of $\hat{T}$. However, if $\hat{T}$ and $\hat{T}^*$ are consistent and the sample size n is sufficiently large, then $\alpha I + \hat{T}^*\hat{T}$ is also invertible. Then, the solution to (4.79) is given by:

$$\hat{f}^\alpha = \left(\alpha I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\hat{r}. \tag{4.80}$$

Tikhonov's regularized solution of the problem in (4.77) can be expressed in terms of the sequences λ_i, f_i and ψ_i, $i = 1, 2, \dots$ as follows:

$$f^\alpha = (\alpha I + T^*T)^{-1} T^*r = \sum_{i \geq 0} \frac{\lambda_i}{\alpha + \lambda_i^2}\langle r, \psi_i\rangle f_i. \tag{4.81}$$

By comparison of (4.81) with the result provided by the Picard theorem, we observe that the Tikhonov regularization controls the decrease of eigenvalues λ_i by replacing $\frac{1}{\lambda_i}$ in (4.75) with $\frac{\lambda_i}{\alpha+\lambda_i^2}$ in (4.81). The regularized solution f^α can be written as:

$$f^\alpha = \arg\min_f \left[||r - Tf||^2 + \alpha||f||^2\right]. \tag{4.82}$$

(4.82) shows that the method adds a penalty term $\alpha||f||^2$ to the minimization of $||r - Tf||^2$. α is the regularization parameter which is chosen to be positive and converging to zero. It can be estimated by minimizing a criterion such as the leave-one-out Cross Validation function below (see Centorrino et al., 2017):

$$CV_n(\alpha) = \sum_{i=1}^{n}\left[(\hat{T}\hat{f}^\alpha_{(-i)})(w_i) - \hat{r}(w_i)\right]^2. \tag{4.83}$$

$\hat{T}\hat{f}^\alpha_{(-i)}$ is the leave-one-out estimator which is derived by eliminating the i-th observation from the sample before estimating Tf. The regularization parameter serves to control the convergence of the eigenvalues of T to zero through eq.(4.81).

Landweber - Fridman regularization

The Landweber - Fridman method belongs to the family of truncated iterative regularization methods. It avoids the inversion of the $\hat{T}^*\hat{T}$ matrix in (4.78) by using an iterative approximation process. According to the method, (4.78) is multiplied by a quantity c such that $c||T^*T|| < 1$, yielding:

$$c\hat{T}^*\hat{r} = c\hat{T}^*\hat{T}\hat{f}. \tag{4.84}$$

c ensures the convergence of the iterative process. The SVD states that the largest eigenvalue of T^*T is 1. Therefore, any c smaller than 1 will guarantee convergence. Adding $\hat{f}$ in both sides of (4.84) provides the following recursive solution of the method:

$$\hat{f}_{k+1} = \hat{f}_k + c\hat{T}^* \left(\hat{r} - \hat{T}\hat{f}_k \right), \ \ \forall k = 0, 1, \ldots \tag{4.85}$$

The solution in (4.85) can be rewritten as:

$$\hat{f}^{1/\alpha} = c \sum_{k=0}^{1/\alpha - 1} \left(I - c\hat{T}^*\hat{T} \right)^k \hat{T}^*\hat{r}. \tag{4.86}$$

In (4.86), $1/\alpha$ is the total number of iterations needed and α represents the regularization parameter. The number of iterations can be specified minimizing the leave-one-out cross validation criterion described before (see Centorrino, 2015). Florens and Racine (2012) mention that the smoothing parameters for T and T^* can be updated at every iteration in (4.85). Therefore, after deriving $\hat{f}_0 = c\hat{T}^*\hat{r}$ from a first estimation of r, T and T^*, the bandwidths of T and T^* can be updated using $\hat{f}_0$. $\hat{f}_1$ is computed from (4.85). These steps are repeated until a stopping rule, such as the minimization criterion presented in Florens and Racine (2012):

$$SSR(k) = k \sum_{i=1}^{n} \left[(\hat{T}\hat{f}_k)(w_i) - \hat{r}(w_i) \right]^2, \ \ k = 1, 2, \ldots \tag{4.87}$$

The iteration procedure stops when the above criterion, which describes a locally convex function, starts to increase.

Appendix A: Assumptions in local constant nonparametric regression

To establish the asymptotic properties of $\hat{f}$, the following assumptions are made (see Hall et al., 2007).

Assumption 4 *We assume the following:*

1. $(Y_i, \mathbf{Z}_i),\ i = 1, \dots, n$ *are i.i.d.*

2. $u_i,\ i = 1, \dots, n$ *has finite moments.*

3. f*,* ϕ *and* $\sigma^2(\bar{\mathbf{z}}) = E(u_i^2 | \bar{\mathbf{Z}}_i = \bar{\mathbf{z}})$ *have two continuous derivatives.*

4. M *is continuous, nonnegative, with compact support.*

5. *The distribution* ϕ *of* $\mathbf{Z}$ *is bounded away from zero; there is* $m > 0$ *such that* $|\phi(\mathbf{z})| > m,\ \forall \mathbf{z}$*.*

Assumption 5 *Define:*

$$H = \prod_{s=1}^{\bar{c}} h_s \prod_{s=\bar{c}+1}^{c} \min(h_s, 1).$$

For $0 < \epsilon < 1/(c+4)$ *and for a constant* $\gamma > 0$*, assume:*

1. $n^{\epsilon - 1} \leq H \leq n^{-\epsilon}$

2. $n^{-\gamma} < h_s < n^{\gamma},\ \forall s = 1, \dots, c$

3. *The kernel* $k(h_s, u, v) = k\left(\frac{u-v}{h_s}\right)$ *is a symmetric, compactly supported, Holder-continuous distribution, with* $k(\rho) < k(0),\ \forall \rho > 0$*.*

Assumption 6 *Define* $\bar{\varphi}(\bar{\mathbf{z}}) = E[\hat{f}(\mathbf{z})\hat{\phi}(\mathbf{z})]/E[\hat{\phi}(\mathbf{z})]$*. Assume that the function*

$$\int [\bar{\varphi}(\bar{\mathbf{z}}) - \bar{f}(\bar{\mathbf{z}})]^2 \bar{M}(\bar{\mathbf{z}}) \bar{\phi}(\bar{\mathbf{z}}) d\bar{\mathbf{z}}$$

vanishes if and only if all smoothing parameters $h_1, \dots, h_{\bar{c}}$ *and* $\lambda_1, \dots, \lambda_{\bar{d}}$ *vanish.* $\bar{M}$ *is defined as* $\bar{M}(\bar{\mathbf{z}}) = \int \tilde{f}(\tilde{\mathbf{z}}) M(\mathbf{z}) d\tilde{\mathbf{z}}$*.*

Moreover, denote by $\alpha_1^0, \dots, \alpha_{\bar{c}}^0, \beta_1^0, \dots, \beta_{\bar{d}}^0$ the values of $\alpha_1, \dots, \alpha_{\bar{c}}, \beta_1, \dots, \beta_{\bar{d}}$ in (4.43) that minimize χ^r. We assume the following:

Assumption 7 *Every* $\alpha_s^0,\ s = 1, \dots, \bar{c}$ *is positive and every* $\beta_s^0,\ s = 1, \dots, \bar{d}$ *is nonnegative. Every one of them is finite and uniquely defined.*

Appendix B: Assumptions in local linear nonparametric regression

To determine the asymptotic normality of the local linear estimator, we assume that the following three requirements hold. More information on these assumptions are found in Li and Racine (2004).

Assumption 8 *Every $\alpha_s^0,\ s = 1, \dots, c$ and every $\beta_s^0,\ s = 1, \dots, d$ is uniquely defined and finite.*

Assumption 8 and equations (4.58) imply that the smoothing parameters cannot be infinity. Then, (4.56) reveals that this requirement rules out the case where $f_{ss}(\mathbf{z}) = 0,\ \forall \mathbf{z}$ and $\forall s = 1, \dots, c$. Therefore, f cannot be a linear function in any of its components.

Assumption 9 *It is assumed that:*

1. *$(\hat{h}_1, \dots, \hat{h}_c, \hat{\lambda}_1, \dots, \hat{\lambda}_d) \in [0, \eta]^{c+d}$ lies in a shrinking set. $\eta = \eta(n)$ is a positive sequence that approaches zero slower than the inverse of any polynomial in n. Further, we assume that $nh_1 \dots h_c \geq t_n$, where $t_n \to \infty$ as $n \to \infty$.*

2. *the kernel function $k : \mathbb{R} \to \mathbb{R}$ described in eq.(4.32) is a bounded symmetric density function. It is m times differentiable, where $m > \max\{2 + 4/c, 1 + c/2\}$ is a positive integer. Moreover, it holds that $\int k(x) x^4 dx < \infty$ and that $\int |k_\varrho(x) x^\varrho| dx < \infty$ for all $\varrho = 1, \dots, m$.*

3. *$\phi(\mathbf{z})$ is bounded below by a positive constant on the support of $\mathbf{Z}$.*

The first point of assumption 9 implies that as $n \to \infty$, $h_s \to 0,$ for $s = 1, \dots, c$ and $nh_1 \dots h_c \to \infty$. Moreover, it entails that as $n \to \infty$, $\lambda_s \to 0,\ s = 1, \dots, d$. Therefore, the assumption does not allow smoothing out irrelevant discrete regressors.

Assumption 10 *We assume the following:*

1. *Consider the family $\mathscr{F}$ of functions from $\mathbb{R}^c$ to $\mathbb{R}$, which are two times differentiable and bounded by functions that have finite 4th moment. Then, for every $\boldsymbol{D}$ of the support, all functions $\sigma^2(\cdot, \boldsymbol{D})$, $g(\cdot, \boldsymbol{D})$ and $f(\cdot, \boldsymbol{D})$ belong to $\mathscr{F}$.*

2. *$(Y_i, \mathbf{Z}_i),\ i = 1, \dots, n$ are independent and identically distributed. Also, $u_i,\ i = 1, \dots, n$ as defined in eq.(4.1) have finite 4th moment.*

3. *The quantity*

$$\int \left\{ \frac{\kappa_2}{2} \sum_{s=1}^{c} f_{ss}(\mathbf{z}) \alpha_s^2 + \sum_{s=1}^{d} X_s'(\mathbf{z}) \beta_s \right\}^2 \phi(\mathbf{z}) d\mathbf{z} + \frac{\Omega'}{n \alpha_1 \dots \alpha_c}$$

   is uniquely minimized at $(\alpha_1^0, \dots, \alpha_c^0, \beta_1^0, \dots, \beta_d^0)$. All α_s^0 and β_s^0 are finite.

Bibliography

Acemoglu, D. (2008). *Introduction to modern economic growth*. Princeton University Press.

Acemoglu, D. (2015). Localised and biased technologies: Atkinson and Stiglitz's New view, induced innovations, and directed technological change. *Economic Journal*, 125(583):443–463.

Aitchison, J. and Aitken, C. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420.

Angrist, J. (1991). Instrumental variables estimation of average treatment effects in econometrics and epidemiology.

Angrist, J. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. Technical report, National Bureau of Economic Research.

Antle, J. M. and Capalbo, S. M. (1988). An introduction to recent developments in production theory and productivity measurement. *Agricultural productivity: Measurement and explanation*, pages 17–95.

Antonioli, D., Gioldasis, G., and Musolesi, A. (2018). Estimating a non-neutral production function: a heterogeneous treatment effect approach. SEEDS Working Papers 0618, SEEDS, Sustainability Environmental Economics and Dynamics Studies.

Atkinson, A. and Stiglitz, J. (1969). A new view of technological change. *The Economic Journal*, 79(315):573–578.

Bai, J. and Ng, S. (2004). A panic attack on unit roots and cointegration. *Econometrica*, 72(4):1127–1177.

Baltagi, B. H., Bresson, G., Griffin, J. M., and Pirotte, A. (2003). Homogeneous, heterogeneous or shrinkage estimators? some empirical evidence from french regional gasoline consumption. *Empirical Economics*, 28(4):795–811.

Barrio-Castro, T., López-Bazo, E., and Serrano-Domingo, G. (2002). New evidence on international r&d spillovers, human capital and productivity in the oecd. *Economics Letters*, 77(1):41–45.

Basu, S. and Fernald, J. G. (1997). Returns to Scale in U.S. Production: Estimates and Implications. *Journal of Political Economy*, 105(2):249–283.

Blackorby, C., Knox Lovell, C. A., and Thursby, M. C. (1976). Extended Hicks Neutral Technical Change. *The Economic Journal*, 86(344):845–852.

Bound, J., Jaeger, D., and Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Bowman, A. W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford.

Brys, G., Hubert, M., and Struyf, A. (2003). A comparison of some new measures of skewness. In *Developments in robust statistics*, pages 98–113. Springer.

Canay, I. A., Santos, A., and Shaikh, A. M. (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6):2535–2559.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751.

Castellacci, F. (2008). Technological paradigms, regimes and trajectories: Manufacturing and service industries in a new taxonomy of sectoral patterns of innovation. *Research Policy*, 37(6-7):978–994.

Centorrino, S. (2015). Data-Driven selection of the Regularization Parameter in Additive Nonparametric Instrumental Regressions. *Under Revision*.

Centorrino, S., Feve, F., and Florens, J.-P. (2017). Additive nonparametric instrumental regressions: A guide to implementation. *Journal of Econometric Methods*, 6(1).

Cerulli, G. (2014). Ivtreatreg: A command for fitting binary treatment models with heterogeneous response to treatment and unobservable selection. *Stata Journal*, 14(3):453–480.

Chambers, R. G. (1988). *Applied production analysis: a dual approach*. Cambridge University Press.

Chu, C.-K. and Marron, J. S. (1991). Choosing a kernel regression estimator. *Statistical Science*, pages 404–419.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

Coe, D. T. and Helpman, E. (1995). International r&d spillovers. *European economic review*, 39(5):859–887.

Coe, D. T., Helpman, E., and Hoffmaister, A. W. (2009). International r&d spillovers and institutions. *European Economic Review*, 53(7):723–741.

Crépon, B., Duguet, E., and Mairessec, J. (1998). Research, Innovation And Productivi[Ty: An Econometric Analysis At The Firm Level. *Economics of Innovation and New Technology*, 7(2):115–158.

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

David, P. A. (1975). *Technical choice innovation and economic growth: essays on American and British experience in the nineteenth century*. Cambridge University Press.

Delgado, M. S., McCloud, N., and Kumbhakar, S. C. (2014). A generalized empirical model of corruption, foreign direct investment, and growth. *Journal of Macroeconomics*, 42:298–316.

D'Haultfoeuille, X. (2011). On the Completeness Condition in Nonparametric Instrumental Problems. *Econometric Theory*, 27(03):460–471.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania 19103.

Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems*, volume 375. Springer Science & Business Media.

Ertur, C. and Musolesi, A. (2017). Weak and strong cross-sectional dependence: A panel data analysis of international technology diffusion. *Journal of Applied Econometrics*, 32(3):477–503.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, pages 196–216.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, pages 2008–2036.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

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

Fridman, V. (1956). A method of successive approximations for Fredholm integral equations of the first kind. *Uspeskhi Math. Nauk.*, 11:233–234.

Griliches, Z. (1998). Patent Statistics as Economic Indicators: A Survey. In Griliches, Z., editor, *R&D and Productivity: The Econometric Evidence*, pages 287–343. University of Chicago Press, Chicago.

Hadamard, J. (2014). *Lectures on Cauchy's problem in linear partial differential equations*. Courier Corporation.

Hall, B. H. (2011). Innovation and productivity. *Nordic Economic Policy Review*, 2:37.

Hall, B. H., Lotti, F., and Mairesse, J. (2008). Employment, innovation, and productivity: evidence from italian microdata. *Industrial and Corporate Change*, 17(4):813–839.

Hall, B. H., Lotti, F., and Mairesse, J. (2009). Innovation and productivity in smes: empirical evidence for italy. *Small Business Economics*, 33(1):13–33.

Hall, P., Li, Q., and Racine, J. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics*, 84(4):784–789.

Hall, P. and Marron, J. S. (1991). Local minima in cross-validation functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 245–252.

Hall, P. and Racine, J. (2015). Infinite order cross-validated local polynomial regression. *Journal of Econometrics*, 185:510–525.

Hall, P., Racine, J., and Q., L. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the %American Statistical Association*, 99(468):1015–1026.

Hansen, B. E. (2014). Nonparametric sieve regression: Least squares, averaging least squares, and cross-validation. In Racine, J. S., Su, L., and Ullah, A., editors, *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, pages 215–248. Oxford University Press.

Hansen, C., Hausman, J., and Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry. *Statistical Science*, pages 120–129.

He, Z. and Opsomer, J. D. (2015). Local polynomial regression with an ordinal covariate. *Journal of Nonparametric Statistics*, 27(4):516–531.

Henderson, D. J. and Parmeter, C. F. (2015). *Applied nonparametric econometrics*. Cambridge University Press.

Hicks, J. (1932). *The theory of wages*. London: Macmillan.

Hicks, J. (1963). *The theory of wages*. Springer.

Horowitz, J. (2011). Applied Nonparametric Instrumental Variables Estimation. *Econometrica*, 79(2):347–394.

Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, 52(12):5186–5201.

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 (Statistical Methodology)*, 60(2):271–293.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

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

Jones, M. and Kappenman, R. (1992). On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, pages 337–349.

Jones, M., Marron, J. S., and Sheather, S. J. (1992). *Progress in data-based bandwidth selection for kernel density estimation*. Department of Statistics [University of North Carolina at Chapel Hill].

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407.

Jones, R. W. (1965). The structure of simple general equilibrium models. *Journal of Political Economy*, 73(6):557–572.

Kandylas, V., Upham, S., and Ungar, L. H. (2010). Analyzing knowledge communities using foreground and background clusters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):7.

Kao, C., Chiang, M.-H., and Chen, B. (1999). International r&d spillovers: an application of estimation and inference in panel cointegration. *Oxford Bulletin of Economics and statistics*, 61(S1):691–709.

Kapetanios, G., Pesaran, M. H., and Yamagata, T. (2011). Panels with non-stationary multifactor error structures. *Journal of Econometrics*, 160(2):326–348.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Kelejian, H. H. (1971). Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association*, 66(334):373–374.

Kennedy, C. and Thirlwall, A. (1977). Extended hicks neutral technical change-a comment. *Economic Journal*, 87(348).

Kennedy, C. and Thirlwall, A. P. (1972). Surveys in applied economics: technical progress. *The Economic Journal*, 82(325):11–72.

Kiefer, N. M. and Racine, J. S. (2009). The smooth colonel meets the reverend. *Journal of Nonparametric Statistics*, 21(5):521–533.

Kiefer, N. M. and Racine, J. S. (2017). The smooth colonel and the reverend find common ground. *Econometric Reviews*, 36(1-3):241–256.

Kress, R. (1999a). *Linear integral equations*, volume 82. Springer.

Kress, R. (1999b). Linear integral equations, volume 82 of applied mathematical sciences.

Landweber, L. (1951). An iterative formula for Fredholm integral equations of the first kind. *American Journal of Mathematics*, 73:615–624.

Lapan, H. and Bardhan, P. (1973). Localized technical progress and transfer of technology and economic development. *Journal of Economic Theory*, 6(6):585–595.

Lee, G. (2006). The effectiveness of international knowledge spillover channels. *European Economic Review*, 50(8):2075–2088.

Li, D., Simar, L., and Zelenyuk, V. (2016). Generalized nonparametric smoothing with mixed discrete and continuous data. *Computational Statistics & Data Analysis*, 100:424–444.

Li, Q. and Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *journal of multivariate analysis*, 86(2):266–292.

Li, Q. and Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14(2):485–512.

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

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436.

Lichtenberg, F. R. and van Pottelsberghe de la Potterie, B. (1998). International r&d spillovers: a comment. *European Economic Review*, 42(8):1483–1491.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Annals of Statistics*, pages 415–438.

Lööf, H., Mairesse, J., and Mohnen, P. (2016). CDM 20 Years After CDM 20 Years After.

Ma, S., Racine, J. S., and Yang, L. (2015). Spline regression in the presence of categorical predictors. *Journal of Applied Econometrics*, 30:703–717.

Maasoumi, E., Racine, J. S., and Stengos, T. (2007). Growth and convergence: A profile of distribution dynamics and mobility. *Journal of Econometrics*, 136:483–508.

Maddala, G. S., Trost, R. P., Li, H., and Joutz, F. (1997). Estimation of short-run and long-run elasticities of energy demand from panel data using shrinkage estimators. *Journal of Business & Economic Statistics*, 15(1):90–100.

Mairesse, J. and Mohnen, P. (2010). Using innovation surveys for econometric analysis. In *Handbook of the Economics of Innovation*, volume 2, pages 1129–1155. Elsevier.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, pages 712–736.

Masry, E. (1996). Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications*, 65(1):81–101.

Mohnen, P. and Hall, B. H. (2013). Innovation and Productivity: An Update. *Eurasian Business Review*, 3(1):47–65.

Montiel Olea, J. L. and Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3):358–369.

Morimoto, Y. (1974). Neutral technical progress and the separability of the production function. *The Economic Studies Quarterly (Tokyo. 1950)*, 25(3):66–69.

Musolesi, A. and Huiban, J.-P. (2010). Innovation and productivity in knowledge intensive business services. *Journal of Productivity Analysis*, 34(1):63–81.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Newey, W. (2013). Nonparametric instrumental variables estimation. *American Economic Review*, 103(3):550–556.

Newey, W. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Pakes, P. and Griliches, Z. (1984). Patents and R&D at the Firm Level: A First Look. In Griliches, Z., editor, *R&D, Patents, and Productivity*, pages 55–71. Chicago University Press, Chicago.

Parisi, M. L., Schiantarelli, F., and Sembenelli, A. (2006). Productivity, innovation and R&D: Micro evidence for Italy. *European Economic Review*, 50(8):2037–2061.

Parmeter, C. and Racine, J. (2018). Nonparametric estimation and inference for panel data models.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Pavitt, K. (1984). Sectoral patterns of technical change: towards a taxonomy and a theory. *Research policy*, 13(6):343–373.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130.

Racine, J. and Parmeter, C. (2014). Data-driven model evaluation: a test for revealed performance. In Racine, J. S., Su, L., and Ullah, A., editors, *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, pages 308–345. Oxford University Press.

Reiss, P. T. and Todd Ogden, R. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):505–523.

Rodriguez-Poo, J. M. and Soberon, A. (2017). Nonparametric and semiparametric panel data models: Recent developments. *Journal of Economic Surveys*, 31(4):923–960.

Rosenblatt, M. et al. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.

Salter, W. E. G. (1966). *Productivity and technical change*. Cambridge University Press Cambridge.

Sheather, S. J. (2004). Density estimation. *Statistical Science*, 19(4):588–597.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. Springer Science & Business Media.

Solow, R. M. (1957). Technical change and the aggregate production function. *The review of Economics and Statistics*, pages 312–320.

Steedman, I. (1985). On the 'Impossibility' of Hicks-Neutral Technical Change. *The Economic Journal*, 95(379):746–758.

Stewart, F. (1978). *Technology and Underdevelopment*. Springer.

Stone, C. J. (1977). Consistent nonparametric regression. *The annals of statistics*, pages 595–620.

Su, L. and Jin, S. (2012). Sieve estimation of panel data models with cross section dependence. *Journal of Econometrics*, 169(1):34–47.

Tikhonov, A. (1943). On the stability of inverse problems. *Toklady Akademii Nauk SSSR*, 39(5):195–198.

UNICREDIT (2008). Decima indagine sulle imprese manifatturiere italiane. Technical report, UNICREDIT.

Uzawa, H. and Watanabe, T. (1961). A note on the classification of technical inventions. *The Economic Studies Quarterly (Tokyo. 1950)*, 12(1):68–72.

Verbeek, M. (2008). *A guide to modern econometrics*. John Wiley & Sons.

Wand, M. and Jones, M. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036.

Wood, S. N. (2012). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1):221–228.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Wooldridge, J. M. (1995). Score diagnostics for linear models estimated by two stage least squares. *Advances in econometrics and quantitative economics: Essays in honor of Professor CR Rao*, pages 66–87.

Wooldridge, J. M. (2003). Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics letters*, 79(2):185–191.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Xia, Y. and Li, W. (2002). Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *Journal of multivariate analysis*, 83(2):265–287.

Zhao, Y. (2012). *R and data mining: Examples and case studies*. Academic Press.