UNIVERSIDADE DA CORUÑA

Methodological contributions in semiparametric regression models for functional data

Doctoral Thesis

Author:

Silvia Novo Díaz

Supervisors:

Germán Aneiros Pérez

Philippe Vieu

2021

UNIVERSIDADE DA CORUÑA

Methodological contributions in semiparametric regression models for functional data

Author:

Silvia Novo Díaz

Supervisors:

Germán Aneiros Pérez

Philippe Vieu

Doctoral Program:

Statistics and Operational Research

2021

UNIVERSIDADE DA CORUÑA

The undersigned certify that they are the advisors of the Doctoral Thesis entitled "Methodological contributions in semiparametric regression models for functional data", developed by Silvia Novo Díaz at the University of A Coruña (Department of Mathematics), as part of the interuniversity PhD program (UDC, USC and UVigo) of Statistics and Operational Research, and hereby give their consent to the author to proceed with the thesis presentation and the subsequent defense.

Los abajo firmantes hacen constar que son los directores de la Tesis Doctoral titulada "Methodological contributions in semiparametric regression models for functional data", realizada por Silvia Novo Díaz en la Universidade da Coruña (Departamento de Matemáticas) en el marco del programa interuniversitario (UDC, USC y UVigo) de doctorado en Estadística e Investigación Operativa, dando su consentimiento para que la autora proceda a su presentación y posterior defensa.

Os abaixo asinantes fan constar que son os directores da Tese de Doutoramento titulada "Methodological contributions in semiparametric regression models for functional data", desenvolta por Silvia Novo Díaz na Universidade da Coruña (Departamento de Matemáticas) no marco do programa interuniversitario (UDC, USC e UVigo) de doutoramento en Estatística e Investigación de Operacións, dando o seu consentimento para que a autora proceda á súa presentación e posterior defensa.

A Coruña, September 20th, 2021.

Advisors:

Dr. Germán Aneiros Pérez

Dr. Philippe Vieu

PhD student:

Silvia Novo Díaz

Á miña familia.
Aos que están, Alejandro, María, José Luis, Aurora, Noelia e Julia,
e á memoria dos que xa non están.

Agradecementos persoais

Gustaríame expresar o meu máis sincero agradecemento aos meus directores de tese, Germán e Philippe, por compartir comigo os seus coñecementos e polo seu apoio constante ao longo destes anos. A Germán, pola atención e a axuda que me brindou incondicionalmente, incluso en fin de semana, festivo ou vacacións; por ser para min un exemplo de bo facer. *A Philippe, por su apoyo en la distancia, por poner esa gota de humor que les quita hierro a las dificultades y por acogerme tan bien en Toulouse. Este agradecimiento lo extiendo a su familia.*

Quixera agradecerlles tamén aos membros do tribunal de predefensa, Pilar García Soidán, Wenceslao González Manteiga e Juan Vilar Fernández, os seus comentarios e aportacións, os cales contribuíron a aumentar a calidade desta memoria.

Como toda aventura ten un inicio, quixera dar as grazas ás persoas que me motivaron a dar o primeiro paso. Grazas aos profesores do Máster en Técnicas Estatísticas por despertar en min esa inquietude pola investigación, en especial a Alberto, César, Rosa e Juan Carlos, quen ademais de coñecemento académico me brindaron orientación e consellos. Grazas tamén a Ricardo, por aquelas conversas telefónicas que me levaron a iniciar o doutoramento na Coruña e por seguir pendente de min e dos demais doutorandos ao longo deste tempo.

Ademais, sinto a necesidade de dar grazas infinitas a aqueles que sempre se embarcan comigo en todo o que decido: á miña madriña, aos meus pais, á miña avoa e á miña irmá, por axudarme sempre en todo o que puideron e ser unha fonte permanente de apoio, cariño e comprensión; ao meu padriño e á miña avoa Toña, porque lles tería gustado verme chegar ata aquí; á miña parella, Alejandro, por atreverse a acompañarme e ser quen de sacarme un sorriso sempre, incluso nos momentos baixos.

Gustaríame darlles as grazas tamén aos meus amigos, en especial a Lucía, polas

visitas esperadas e inesperadas e polas súas longas chamadas; a Tere, por eses cafés de desconexión e esas conversas vía mensaxes de audio; a Julia por facer un oco para visitarnos a Lucía e a min no verán; a Manu, pola acollida en Toulouse; aos meus compañeiros e amigos do Laboratorio 2.1 e do CITIC que fixeron máis amenas as xornadas de traballo e a estadía na Coruña, en especial a Eva, Jonathan e Luis, a mellor mesa, a Inés e Andrea, as miñas compañeiras de congresos, e a Isa, compañeira de comidas.

Finalmente, grazas a todas esas persoas que fun atopando no camiño e que contribuíron a que hoxe estea aquí.

A Coruña, 20 de setembro de 2021.
Silvia Novo Díaz

Institutional acknowledgements

This research has been partially supported by the Spanish Ministerio de Economía y Competitividad (MINECO) under Grants MTM2014-52876-R and MTM2017-82724-R, by the Spanish Ministerio de Ciencia e Innovación (MICINN) under Grant PID2020-113578RB-I00, by the Xunta de Galicia through Centro Singular de Investigación de Galicia accreditation under Grants ED431G/01 2016-2019 and ED431G 2019/01 and through the Grupos de Referencia Competitiva under Grants ED431C 2016-015 and ED431C2020-014 and in part by the European Union (European Regional Development Fund-ERDF).

The author particularly thanks the contracts financed by the research group MODES (from May 1, 2017 to August 31, 2017) and by the CITIC (from September 1, 2017 to May 30, 2018) and the PhD contract financed by the Xunta de Galicia and the European Union (European Social Fund-ESF), the reference of which is ED481A-2018/191 (from May 31, 2018). Some results of this thesis have been obtained during a stay of the author at the Université Paul Sabatier, Toulouse (from March 13, 2019 to June 12, 2019), financed by the Xunta de Galicia, with reference ED481A-2018/191.

Abstract

This doctoral thesis is dedicated to functional regression for scalar response. In particular, we focus on functional semiparametric models, which combine the practical advantages of parametric and nonparametric approaches, surpassing both methodologies. Accordingly, several semiparametric models involving a functional single-index component were studied from a theoretical and practical perspective. First, for the functional single-index model (FSIM) and the semi-functional partial linear single-index model (SFPLSIM), we provide uniform consistency results (over all parameters involved) for kernel- and k-Nearest-Neighbours-based statistics related to the estimation of the semiparametric component. Second, for the sparse semi-functional partial linear single-index model (SSFPLSIM), we develop a variable selection procedure in the linear component based on penalized least squares (PLS). The good behaviour of this method is theoretically assured (rates of convergence of the estimators are obtained, as well as asymptotic behaviour of the variable selection procedure). Third, the SSFPLSIM is adapted to the case in which covariates with linear effect come from the discretization of a curve. For this new model, the multi-functional partial linear single-index model (MFPLSIM), the variable selection problem was also studied. Consequently, two new algorithms were proposed (providing theoretical results that ensure their good performance) to solve the inefficiency of the PLS method when it is directly applied to the MFPLSIM. For all the models and procedures mentioned above, theoretical results are accompanied by both simulation studies and real data applications which illustrate the good performance of the proposed methodology in practice.

Resumo

Esta tese está adicada ao estudo da regresión funcional con variable resposta escalar. En particular, centrámonos en modelos funcionais semi-paramétricos, os cales combinan as vantaxes prácticas dos enfoques paramétrico e non-paramétrico, superando a ambas metodoloxías. Desta maneira, estudáronse, tanto dende o punto de vista teórico como dende a perspectiva práctica, varios modelos semi-paramétricos que involucran unha compoñente funcional *single-index*. En primeiro lugar, para o *functional single-index model* (FSIM) e para o *semi-functional partial linear single-index model* (SFPLSIM) establecemos resultados de consistencia uniforme (sobre todos os parámetros involucrados) para os estatísticos de tipo núcleo e tipo k-veciños-máis-próximos relacionados coa estimación da compoñente semi-paramétrica do modelo. En segundo lugar, para o *sparse semi-functional partial linear single-index model* (SSFPLSIM) desenvolvemos un procedemento de selección de variables na compoñente linear baseado en mínimos cadrados penalizados (PLS, iniciais de *penalized least squares*). O bo comportamento deste método asegurouse dende o punto de vista teórico (obtendo taxas de converxencia dos estimadores, así como o comportamento asintótico do procedemento de selección de variables). En terceiro lugar, o SSFPLSIM adaptouse ao escenario no cal as covariables con efecto linear proveñen da discretización dunha curva. Para este novo modelo, o *multi-functional partial linear single-index model* (MFPLSIM), estudouse tamén o problema da selección de variables e propuxéronse dous novos algoritmos (dos que aseguramos teoricamente o seu bo comportamento) para resolver a ineficacia do método PLS cando se aplica directamente ao MFPLSIM. Para todos os modelos e procedementos citados, os resultados teóricos acompañáronse de estudos de simulación e aplicacións a datos reais que ilustran o bo comportamento na práctica da metodoloxía presentada.

Resumen

Esta tesis está dedicada al estudio de la regresión funcional con variable respuesta escalar. En particular, nos centramos en modelos funcionales semi-paramétricos, los cuales combinan las ventajas prácticas de los enfoques paramétrico y no-paramétrico, superando a ambas metodologías. De esta forma, se estudiaron, tanto desde el punto de vista teórico como desde la perspectiva práctica, varios modelos semi-paramétricos que involucran una componente funcional *single-index*. En primer lugar, para el *functional single-index model* (FSIM) y para el *semi-functional partial linear single-index model* (SFPLSIM) establecemos resultados de consistencia uniforme (sobre todos los parámetros involucrados) para los estadísticos de tipo núcleo y de tipo k-vecinos-más-próximos relacionados con la estimación de la componente semi-paramétrica del modelo. En segundo lugar, para el *sparse semi-functional partial linear single-index model* (SSFPLSIM) desarrollamos un procedimiento de selección de variables en la componente linear basado en mínimos cuadrados penalizados (PLS, iniciales de *penalized least squares*). El buen comportamiento de este método se ha asegurado desde el punto de vista teórico (obteniendo tasas de convergencia de los estimadores, así como el comportamiento asintótico del procedimiento de selección de variables). En tercer lugar, el SSFPLSIM se ha adaptado al escenario en el cual las covariables con efecto linear provienen de la discretización de una curva. Para este nuevo modelo, el *multi-functional partial linear single-index model* (MFPLSIM), se ha estudiado también el problema de selección de variables y se propusieron dos nuevos algoritmos (de los que aseguramos teóricamente su buen comportamiento) para resolver la ineficiencia del método PLS cuando se aplica directamente al MFPLSIM. Para todos los modelos y procedimientos citados, los resultados teóricos se acompañaron de estudios de simulación y aplicaciones a datos reales que ilustran el buen comportamiento en la práctica de la metodología presentada.

Introduction

Nowadays, Functional Data Analysis (FDA) is one of the main disciplines of Statistics. The emergence of functional variables in applications made it necessary to adapt the traditional methodology for finite-dimensional data to these infinite-dimensional structures, as well as the development of new statistical tools. The point is that the direct use of traditional techniques would force us to work with the discretized observations of functional variables. However, this way of proceeding would have at least three important disadvantages: the existence of strong correlations between the resulting variables, the waste of the functional origin or the dimension of the problem (the ratio between sample size and number of variables).

Precisely, dimensionality was one of the first concerns in the FDA literature. Researchers realized that transforming the functional data sample into elements of finite-dimensional spaces allows a simpler statistical treatment and easier practical interpretation. These facts led to the development of dimension reduction techniques, such as functional principal component analysis (see Dauxois et al. [30], Silverman [103], Boente and Fraiman [16] or Li and Hsing [73]), partial least squares (see Preda and Saporta [91], Krämer et al. [70], Delaigle and Hall [32] or Aguilera et al. [2] for the regression context, Preda et al. [92] for the supervised classification setting and Reiss and Ogden [99] or Febrero-Bande et al. [43] for a comparison between functional principal component and partial least squares approaches) or variable selection in the regression framework (for the extension of ideas in the multivariate context, as Tibshirani [104] or Fan and Lv [40], see Aneiros and Vieu [4] or Aneiros and Vieu [5]).

The mentioned tools have been used in the functional regression setting (mainly in linear modelling). Nevertheless, recent surveys highlighted the need to go further

and develop flexible dimension reduction models for functional regression (see Cuevas [29], Goia and Vieu [55], Vieu [108] or Aneiros et al. [8]). For this purpose, semiparametric ideas seem to be the suitable candidates. Semiparametric modelling combines the flexibility of the nonparametric approach with the advantages of involving parameters in the estimation (see Goia and Vieu [54]): on the one hand, interpretability in practical applications; on the other hand, less sensitivity to dimension effects. However, functional semiparametric regression is still a very undeveloped field.

For these reasons, this dissertation deals with several semiparametric models that involve one or more functional covariates, focusing on the estimation task. The exposition will be organized as explained below:

- Chapter 1 provides an introduction to the statistical framework in which this dissertation is located. The semiparametric regression models that we are going to analyse are presented, together with other regression models related to them. These models will be used in applications to compare results.

- Chapter 2 develops a new automatic and location-adaptive procedure for estimating regression in the functional single-index model (FSIM). This procedure is based on k-Nearest-Neighbours (kNN) ideas. The asymptotic study includes results of uniform consistency over all the parameters involved in the estimation by means of the kNN-based statistic. In addition, we establish analogous asymptotic results for the Nadaraya-Watson kernel-based statistic, which are used as preliminary tools. The results obtained generalize to the case of unknown functional index those provided by Kara-Zaitri et al. [66, 67] for the functional nonparametric model. One of the main characteristics of the convergence rates obtained is that they are similar to those achieved in the one-dimensional setting. This feature gives evidence of the dimension reduction property of the studied methodology. An important consequence of these asymptotics is that they provide theoretical validation to automatic data-driven selectors of the involved parameters, making both procedures (kernel- and kNN-based one) directly usable in practice. The local feature of the kNN approach ensures higher predictive power compared to usual kernel estimates. This fact was illustrated in a simulation study and in an application to a chemometric dataset (Tecator's data). The investigations developed in this chapter

are published in Novo, Aneiros, and Vieu [87].

- Chapter 3 extends uniform consistency results obtained in Chapter 2 to a more complex model, which combines partial linear ideas with a functional single-index component, the semi-functional partial linear single-index model (SFPLSIM). Asymptotics were accompanied by simulated experiments which highlight the advantages of the kNN procedure over alternative techniques. In addition, the real data application based on Tecator's data shows how semi-parametric modelling outperforms alternative modelling ideas. The results obtained in this chapter are published in Novo, Aneiros, and Vieu [89].

- Chapter 4 aims to address dimensionality reduction in the regression context when the predictors are a mixture of functional variable and high-dimensional vector. A flexible model is proposed, combining both sparse linear ideas and semiparametric modelling, the sparse semi-functional partial linear single-index model (SSFPLSIM). A procedure for selecting relevant variables in the linear component of the model is presented. This procedure is based on penalized least squares (PLS). A wide variety of asymptotic results is provided: this includes rates of convergence of the estimators, as well as the asymptotic behaviour of the variable selection procedure. The rates of convergence obtained for the estimator of coefficients in the linear component are the same than those provided by Aneiros et al. [7] in a less complex setting (and the same reached by Fan and Lv [40] in the linear regression context). Furthermore, we showed that the proposed variable selection procedure satisfies the oracle property (see Fan and Li [38]) and that the functional single-index component is estimated with the same rate as if the functional variable were unidimensional (supporting the dimension reduction property of the proposed methodology). Practical issues are analysed through finite sample simulated experiments, while an application to Tecator's data illustrates the usefulness of our methodology. The investigations developed in this chapter are published in Novo, Aneiros, and Vieu [88].

- In Chapter 5, a new sparse semiparametric model is proposed, which incorporates the influence of two functional random variables in a scalar response in

a flexible and interpretable way. One of the functional covariates is included through a single-index structure, and the other one linearly, but through the high-dimensional vector formed by its discretized observations. That is, this model is an adaptation of the SSFPLSIM to the case of real covariates in the linear component with functional origin. However, the direct application of the methodology presented in Chapter 4 is unfeasible: on the one hand, a lot of computational time is needed to carry out variable selection, even for moderate values of the discretization size; on the other hand, the variable selection procedure may be negatively affected by the strong correlations between the covariates with linear effect. Accordingly, two new algorithms are presented to select relevant variables in the linear component and to estimate the MFPLSIM. Both procedures take advantage of the functional origin of the linear covariates. The first method is a fast algorithm which provides results in reasonable time, even for very large values of the discretization size. The second algorithm is a refined procedure, which adds a second step to the fast algorithm, allowing to complete and specify the set of relevant variables selected by the fast method (the second algorithm is an adaptation of the variable selection method presented in Aneiros and Vieu [4]). Since the second algorithm proceed in two stages, it requires dividing the sample into two parts. Some asymptotic results will theoretically support both methods. Finite sample experiments will show the scope of application of both algorithms: the first method provides a solution (without loss in predictive power) to the huge computational time required by standard variable selection methods to estimate the MFPLSIM, and since it does not need the division of the sample, it provides better results under small sample size than the second algorithm; the second method completes the set of relevant linear covariates provided by the first, improving its predictive efficiency in the case of enough sample size. A real data application will show the great applicability of the presented methodology, due to its high predictive power, the interpretability of the outputs and the low computational cost. The investigations contained in this chapter are part of the paper Novo, Aneiros, and Vieu [90], which was submitted for publication.

- In Chapter 6 a brief summary of conclusions is presented, together with some

investigation ideas to be developed in the future.

Contents

Chapter 1

Towards functional semiparametric regression

1.1 Introduction

Nowadays, technological advances in collecting and storage data make more and more frequent having observations of variables which are measured over a continuum[1](at a time interval, over a surface...). As a consequence, measurements in form of curves, images or even more complex structures are obtained, instead of scalars or multivariate vectors. Then, in many applied sciences (as medicine, environmetrics, chemometrics, biometrics, econometrics...) the study of real phenomenons produces observations of *functional variables*, that is, *functional data*.

To better understand what functional data are, let us start introducing two datasets in the field of chemometrics containing functional variables. In chemometrics for analysing and/or detecting some components of a chemical mixture, a common procedure is to observe spectrometric data. This kind of data is obtained by measuring light absorbance of the mixture at several different wavelengths (which will produce functional data). In this way, lengthy, expensive (and sometimes dangerous) chemical experiments can be avoided just by analysing the spectrometric data.

The first chemometric dataset that we are going to present is the well-known

[1]Or which can be assumed to be measured over a continuum (for instance, values are obtained at many discrete time points).

Tecator's data (for more details on the description, see Ferraty and Vieu [47]). Given 215 finely chopped pieces of meat, Tecator's data contain their corresponding near-infrared absorbance spectra observed on 100 equally spaced wavelengths in the range 850–1050 nm. Tecator's data are available at http://lib.stat.cmu.edu/datasets/tecator.

Figure 1.1: Original chemometric data: absorbance versus wavelength.

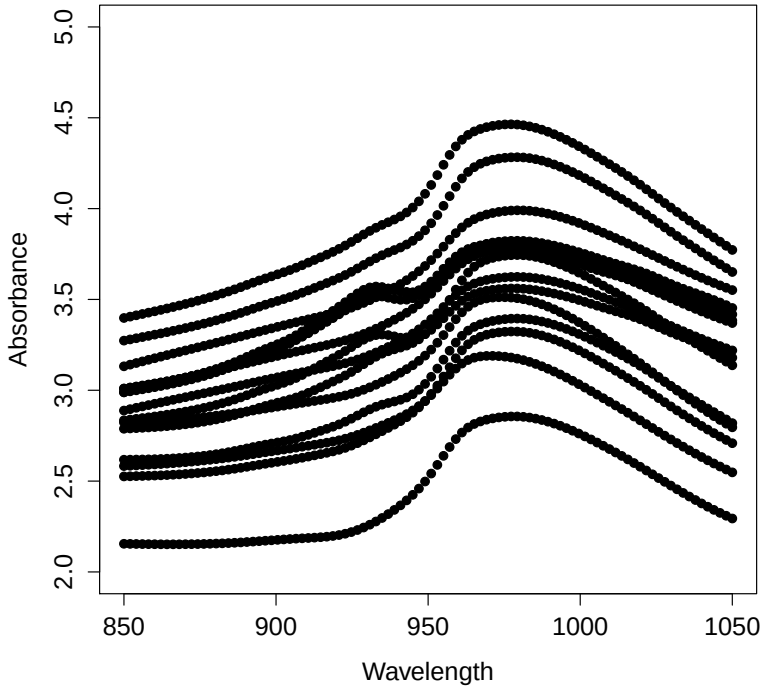
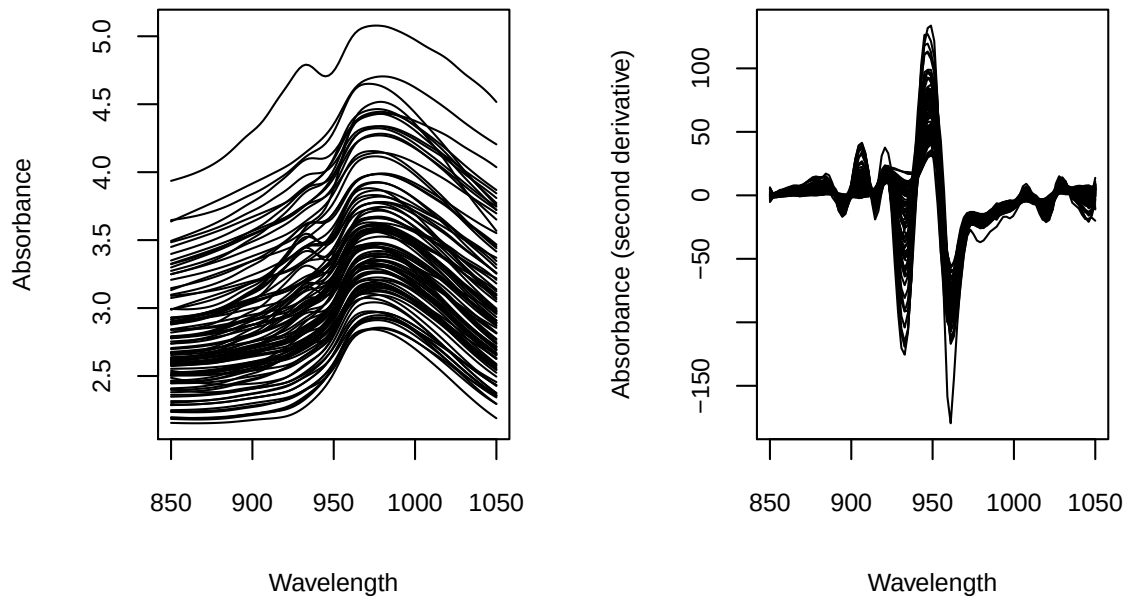


Figure 1.1 plots absorbance versus wavelength for 15 randomly selected pieces of meat. As can be observed, each unit clearly appears as a discretized curve. Because of the fineness of the grid, we can consider each of them as a continuous curve. Figure 1.2 displays samples of both the absorbance curves and their second derivatives.

Tecator's dataset also contains measurements of the fat percentage of each piece of meat. Obtaining this scalar variable requires more expensive and longer chemical experiments, so spectrometric data is used to predict its value in a new piece of meat. In statistical literature, the problem of predicting the fat content using spec-

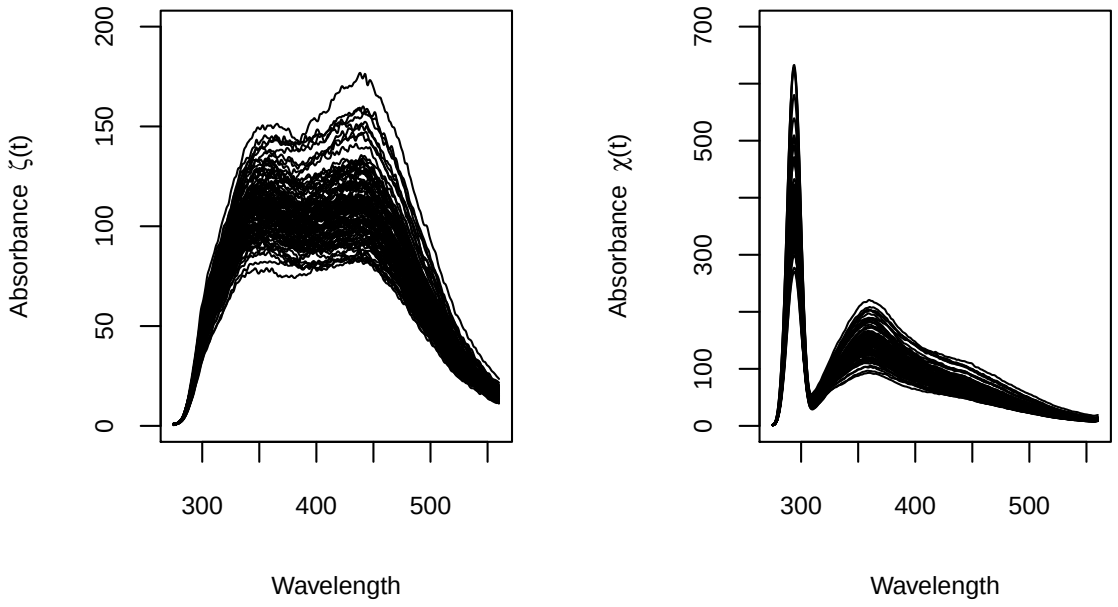
trometric curves was widely studied and it will be also of our interest throughout this dissertation.

Figure 1.2: Sample of 100 absorbance curves (left panel) together with their second derivatives (right panel).



The second chemometric dataset we are going to present is known as Sugar data. At a sugar plant in Scandinavia, 268 samples were obtained by sampling sugar every 8 hours for 3 months. For each sample, the absorbance spectra observed in 571 equally spaced wavelengths in the range 275–560 nm (measured in 0.5 nm intervals) was obtained. But in this case, measurements at two different excitation wavelengths were recorded, generating two functional variables: absorbance spectra at excitation wavelengths 240 nm (the first functional variable, which will be denoted as ζ) and at excitation wavelengths 290 nm (the second functional variable, which will be named $\mathcal{X}$). Sugar dataset is available at `http://www.models.kvl.dk/Sugar_Process`. Samples of both curves can be seen in Figure 1.3.

Figure 1.3: Left panel: Absorbance curves at excitation wavelengths 240 nm (ζ). Right panel: Absorbance curves at excitation wavelengths 290 nm ($\mathcal{X}$).



As part of the experiment, the ash content of sugar, Y, was also determined for each sample. The practical question here is whether one can predict the value of ash content for a new sample simply by using the two functional variables derived from the spectrometric analysis. In this dissertation, we will discuss this issue throughout Chapter 5.

From a statistical point of view, a variable $\mathcal{X}$ is called a functional variable if it takes values in an infinite-dimensional space (the functional space). Then, a *functional dataset* is composed of observations of n functional variables $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ identically distributed as $\mathcal{X}$. In this case, "data atoms" are random functions and datasets contain samples of these random functions. Functional variables have an important distinctive feature: they are infinite-dimensional, in contrast to usual data types found in Statistics. Therefore, statistical methods used in non-functional (finite-dimensional) setting fail when we work with functional data and new specific

statistical methodology had to be developed.

The term *Functional Data Analysis* (FDA) was coined by Ramsay [94] and Ramsay and Dalzell [95] to refer to those statistical tools designed for dealing with functional data. But the origins of FDA are older than the name given to the area, and can be dated to the mid-20th century with the works of Karhunen [68] or Grenander [56] (see Section 2 in Müller [84] for an overview in the early history of FDA). However, scientific production in FDA was sporadic at first. The popularization FDA come with the end of the nineties as functional data started to be common in practical applications and as monographs reviewing a selection of topics in FDA appear (for instance, Bosq [17], Ramsay and Silverman [96], Ramsay and Silverman [97] or Ferraty and Vieu [47]). In the last two decades, FDA became one of the main topics in Statistics and there is an extensive literature in FDA covering multiple areas (see Wang et al. [112] for an overview): principal component analysis, clustering and classification for functional data, correlation and functional regression... but there are still many methodological challenges for analysing functional data (see, for instance, Aneiros et al. [9] for an overview of methodological issues in FDA).

Precisely, regression became one of the trending topics in FDA. Regression is a tool that is commonly used with two main objectives: on the one hand, to model the dependence between a variable of interest (the response variable) and other variables (the explanatory variables or covariates) which often are easier to obtain or to measure; on the other hand, using the proposed model to predict the value of the response variable for new values of the covariates. Regression problems have been widely studied for real or multivariate variables and, as functional data appear in applications, researchers became increasingly interested in relating functional variables to other variables of interest (functional or not). As a consequence, there is an extensive literature on functional regression modelling (see Greven and Scheipl [57] for a general presentation), for both functional response and/or functional covariates. Regarding the case of scalar response with functional covariates, this literature focused mainly on *parametric models*[2] (see Chapter 11 of Hsing and Eubank [63]) or on *nonparametric models* (popularized by Ferraty and Vieu [47]; see Geenens [53],

[2]Let $\mathcal{X}$ be a random variable valued in some infinite-dimensional space $\mathcal{H}$ and let γ be a mapping defined on $\mathcal{H}$ and depending on the distribution of $\mathcal{X}$. A model for the estimation of γ consists in

Ling and Vieu [76] for recent surveys), but *semiparametric regression* is still a very underdeveloped field in FDA (see, however, Goia and Vieu [54] for a review). As will be discussed throughout this dissertation, semiparametric framework is a good middle point between parametric and nonparamentric methodologies, outperforming both of them in many senses: it allows flexibility (unlike parametric models) and interpretability and dimension reduction (unlike nonparametric models).

In order to briefly present the framework in which this dissertation is placed, this chapter is dedicated to making an introduction to the functional semiparametric regression models that we are going to study. For this, we will also describe other models proposed in literature, which are a pillar for the studied ones and with which we will compare them. The presentation will start from the simplest models to finally deal with the more complex structures that we will analyse in this thesis. In Section 1.2 we are going to introduce models with scalar response and only one functional covariate. In Section 1.3, we will present models which combine in an additive way scalar covariates with linear effect and a functional covariate with non-linear effect. Finally, in Section 1.4 we will make an introduction to sparse regression, focused on the models that we are going to study and in those with which we will compare them.

1.2 Univariate functional models for scalar response

In this section we are going to make a brief review of models with scalar response variable and a single covariate, which has functional nature.

1.2.1 The functional linear model

The natural extension of the traditional simple linear model is the *functional linear model* (FLM) proposed in Cardot et al. [24]. The FLM is a parametric model defined

introducing some constraint of the form

$$\gamma \in \mathcal{C}.$$

The model is called a *functional parametric model* for the estimation of γ if $\mathcal{C}$ is indexed by a finite number of elements of $\mathcal{H}$. Otherwise, the model is called a *functional nonparametric model*. For more details on the definition see Ferraty and Vieu [47].

by the relationship

$$Y = \gamma_0 + \int_{\mathcal{I}} \gamma(t)\mathcal{X}(t)dt + \varepsilon, \tag{1.1}$$

where Y is a scalar random variable, while $\mathcal{X}(t)$ with $t \in \mathcal{I}$ is a functional random covariate valued in $L^2(\mathcal{I})$; $\gamma(\cdot)$ is an unknown coefficient-function (square integrable) defined on $\mathcal{I}$ and γ_0 is an unknown real parameter. Finally, ε denotes de random error verifying $\mathbb{E}(\varepsilon|\mathcal{X}) = 0$.

Estimation of model (1.1) has been extensively studied in literature: Cardot et al. [24] use functional principal component analysis to estimate $\gamma(\cdot)$, Cardot et al. [25] propose procedures based on B-spline basis, Ramsay and Silverman [97] estimate model (1.1) using Fourier basis functions, Preda and Saporta [91] carry out estimation using partial least squares procedure...

The interpretability of the estimation of $\gamma(\cdot)$ is one of the main advantages of the FLM. In fact, this feature becomes even more important in practical applications of FDA because of infinite-dimensionality of the data. However, the FLM assumes a linear relation between the response and the functional covariate, a hypothesis which is rarely verified in practice and which could be very restrictive in many contexts.

1.2.2 The functional nonparametric model

An alternative to the FLM is the *functional nonparametric model* (FNM), proposed in Ferraty and Vieu [46]. The FNM is given by the expression

$$Y = m(\mathcal{X}) + \varepsilon. \tag{1.2}$$

In model (1.2) the relation between the scalar response, Y, and the functional random covariate, $\mathcal{X}$, is modelled by an unknown non-linear operator $m(\cdot) : \mathcal{H} \to \mathbb{R}$. Finally, ε is the random error verifying $\mathbb{E}(\varepsilon|\mathcal{X}) = 0$.

Compared to the FLM, the main advantage of the FNM is its flexibility: there is no assumption about the form of the operator $m(\cdot)$. This fact gives the model great applicability in practice with good predictive power. However, the lack of parameters makes it difficult to interpret estimations of the FLM.

1.2.2.1 Estimators

Since the estimation of model (1.2) is related to the methodology that we are going to present in this dissertation, we will do a more detailed explanation about the estimation procedure. To estimate $m(\cdot)$ following a nonparametric approach, the functional covariate $\mathcal{X}$ is assumed to be valued in a semi-metric space $\mathcal{H}$ (in order to quantify proximity between functional elements). Let us denote by $d(\cdot,\cdot)$ a semi-metric[3] in $\mathcal{H}$.

Ferraty and Vieu [47] propose the functional extension of the classical Nadaraya-Watson *kernel estimator* (see Nadaraya [86] and Watson [113]) for estimating the FNM (1.2). That is, given a sample $\{(Y_i,\mathcal{X}_i)\}_{i=1}^n$ of n pairs independent and identically distributed (i.i.d.) to $(Y,\mathcal{X})$, which verify the FNM (1.2):

$$Y_i = m(\mathcal{X}_i) + \varepsilon_i \quad (i = 1,\ldots,n),$$

the kernel estimator of $m(\cdot)$ is given by the expression

$$\widehat{m}(\chi) = \frac{\sum_{i=1}^n Y_i K\left(h^{-1} d(\mathcal{X}_i,\chi)\right)}{\sum_{i=1}^n K\left(h^{-1} d(\mathcal{X}_i,\chi)\right)},\ \forall \chi \in \mathcal{H}, \tag{1.3}$$

where $h \in \mathbb{R}^+$ is the bandwidth, d is the semi-metric and K is the real valued kernel. Some observations should be made about expression (1.3):

- Note that the main difference between expression (1.3) and the Nadaraya-Watson estimator in the finite-dimensional case is the presence of the semi-metric for measuring proximity between functional elements. Since $\mathcal{H}$ is an infinite-dimensional space, the equivalence between norms fails (in contrast to what happens in the finite-dimensional Euclidean space). Therefore, in the functional case we must pay special attention to the choice of the semi-metric.

- Since $\forall \chi_1,\chi_2 \in \mathcal{H}$, the value of $d(\chi_1,\chi_2)$ is always a non-negative quantity,

[3] d is a semi-metric on some space $\mathcal{H}$, if verifies:

1. $\forall \chi \in \mathcal{H}, \quad d(\chi,\chi) = 0.$
2. $\forall \chi_1,\chi_2,\chi_3 \in \mathcal{H}, \quad d(\chi_1,\chi_2) \leq d(\chi_1,\chi_3) + d(\chi_3,\chi_2).$

For details about semi-metrics, see Chapter 3 in Ferraty and Vieu [47].

K must have non-negative support. This leads to the use of asymmetrical functions for kernel as in the multivariate case (unlike the univariate case).

Estimator (1.3) was widely studied in literature: rates of convergence of the estimator (1.3) can be seen in Ferraty and Vieu [46] or in Ferraty and Vieu [47], and for recent results about (1.3) see, for instance, Kara-Zaitri et al. [67].

An alternative to the kernel estimator is the use of the *k-Nearest-Neighbours* (kNN) *estimator*. The kNN procedures are based on the estimation in each element of the considered space using only the k sample observations that are closest to this element. The kNN estimator of $m(\cdot)$, proposed in Burba et al. [22], can be seen as an extension of the kernel estimator and is given by the expression

$$\widehat{m}^*(\chi) = \frac{\sum_{i=1}^n Y_i K\left(H_{k,\chi}^{-1} d(\mathcal{X}_i, \chi)\right)}{\sum_{i=1}^n K\left(H_{k,\chi}^{-1} d(\mathcal{X}_i, \chi)\right)}, \ \forall \chi \in \mathcal{H}, \tag{1.4}$$

where $k \in \mathbb{Z}^+$ is a smoothing factor and K is an asymmetrical kernel. In addition, we have denoted

$$H_{k,\chi} = \min\left\{ h \in \mathbb{R}^+ \text{ such that } \sum_{i=1}^n 1_{B(\chi,h)}(\mathcal{X}_i) = k \right\}, \tag{1.5}$$

with

$$B(\chi, h) = \{ z \in \mathcal{H} : d(\chi, z) \leq h \}. \tag{1.6}$$

Note that, unlike the kernel case, in the kNN estimator (1.4) the smoothing parameter $H_{k,\chi}$ (1.5) depends on χ and on k. For that reason, kNN ideas have been used in early nonparametric one-dimensional literature to build *location-adaptive*[4] smoothers (see e.g. Collomb [28] or Devroye et al. [33]), and they have recently been extended to nonparametric FDA (see, for instance, Biau et al. [15] and Kara-Zaitri et al. [66] for recent results, and Section 2.2 in Ling and Vieu [76] for a survey).

The kNN estimator is more appealing than the kernel one for two reasons. On the one hand, it involves a local smoothing factor $H_{k,\chi}$ making it possible to capture

[4]Note that, in nonparametric statistics, an estimator is said to be "location-adaptive" when the smoothing parameter depends on the element in which one wishes to estimate, χ; in the particular case of nonparametric regression estimation by means of the kNN estimator, the corresponding smoothing parameter is a bandwidth depending on the fixed value k as well as on χ.

local features of the data (while the smoothing factor h of the kernel statistic does not depend on χ). On the other hand, this local smoothing factor depends only on a discrete parameter k taking values in the finite set $\{1, 2, \ldots, n\}$. This fact makes it much easier to select k in practice than the bandwidth h appearing in kernel methods (which takes values in a continuous interval). Nevertheless, the price to pay for such flexibility of the procedure is that the theoretical properties are much more difficult to analyse (because of the randomness of the smoothing factor $H_{k,\chi,\theta}$). More precisely, neither of the two terms in the ratio (1.4) can be written as a sum of independent and identically distributed variables (as can be written those that appear in (1.3)); therefore their analysis will require much more sophisticated tools than standard limit theorems for i.i.d. sequences.

1.2.3 The functional single-index model

A nice middle-point between the FLM and FNM is the *functional single-index model* (FSIM) proposed in Ferraty et al. [48]. The FSIM is a semiparametric model given by the expression

$$Y = r\left(\langle \theta_0, \mathcal{X} \rangle\right) + \varepsilon,$$

where Y denotes a scalar response $\mathcal{X}$ is a functional explanatory random variable valued in a separable Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$, ε is a random error verifying $\mathbb{E}\left(\varepsilon | \mathcal{X}\right) = 0$, $\theta_0 \in \mathcal{H}$ is the functional index and $r(\cdot)$ is the unknown link function. As usual, it is assumed that an only index of $\mathcal{X}$ is sufficient to summarize all the information carried in $\mathcal{X}$ to predict Y. In this way, the functional index θ_0 appears as a filter allowing the extraction of the part of $\mathcal{X}$ explaining the response Y (see Ait-Saïdi et al. [3]).

The FSIM is the first model of study in this dissertation. In fact, Chapter 2 is devoted to contributions on its estimation (both from theoretical and practical point of view). The FSIM can be seen as an extension of the FLM (see Hsing and Eubank [63] for discussion), as well as a special case of the FNM (see Ferraty and Vieu [47]). In fact, the FSIM is an appealing trade-off between these two approaches. On the one hand, it is much more flexible, and hence more reliable in practice than the parametric model (1.1). On the other hand, it presents much less sensitivity

to dimensionality effects than the nonparametric model (1.2) since it involves the estimation of the one-dimensional function $r(\cdot)$ instead of the nonlinear infinite-dimensional operator $m(\cdot)$. These facts allow us to say that the FSIM is a nice competitor for models (1.1) and (1.2).

1.3 Semi-functional partial linear regression for scalar response

In practical applications it is usual to find more than one covariate. In particular, it is common the situation in which, in addition to a functional explanatory variable, there are several scalar variables related to the response. For instance, in Tecator's dataset there are also measurements of the percentage of protein and the percentage of moisture for each piece of meat, which can also help to predict the fat content. In these situations partial linear ideas could be an interesting approach. In this section, we will present some regression models which combine partial linear ideas with nonparametric or semiparametric modelling.

1.3.1 The semi-functional partial linear model

The first model that we are going to mention is the *semi-functional partial linear model* (SFPLM) proposed in Aneiros-Pérez and Vieu [11]. The SFPLM is given by the expression

$$Y = X_1\beta_{01} + \cdots + X_p\beta_{0p} + m\left(\mathcal{X}\right) + \varepsilon, \tag{1.7}$$

where Y denotes the scalar response, $(X_1, \ldots, X_p)^\top$ is a vector of real random covariates while $\mathcal{X}$ is a explanatory variable of functional nature valued in a semi-metric space; $(\beta_{01}, \ldots, \beta_{0p})^\top$ is the vector of unknown real coefficients and $m(\cdot)$ denotes an unknown real-valued operator; ε denotes the random error verifying $\mathbb{E}(\varepsilon|X_1, \ldots, X_p, \mathcal{X}) = 0$.

Estimation of the SFPLM was widely studied in literature (see Aneiros-Pérez and Vieu [11], Aneiros-Pérez and Vieu [13] or Shang [101] for estimation based on kernel procedures and Ling et al. [77] for estimation based on kNN methods) as well as applications and extensions (see Aneiros-Pérez and Vieu [12] for applications of this

model to time-series prediction or Lian [74] for the extension to the case in which the linear variable is also of functional nature). This literature showed that the SFPLM provides both interesting asymptotics and good practical behaviour, since it combines the interpretability of the effect of the linear variables with the flexibility of the effect of the functional variable. However, as the functional variable enters in the model nonparametrically, this component of the model has the usual disadvantages of lack of interpretability and sensitivity to dimensionality effects of the FNM.

1.3.2 The semi-functional partial linear single-index model

An alternative to the SFPLM is the *semi-functional partial linear single-index model* (SFPLSIM), firstly presented in Wang et al. [110]. The SFPLSIM is given by the relationship

$$Y = X_1\beta_{01} + \cdots + X_p\beta_{0p} + r\left(\langle\theta_0, \mathcal{X}\rangle\right) + \varepsilon,$$

where X_j $(j = 1, \ldots, p)$ and Y are real random variables, while $\mathcal{X}$ is a functional random variable valued in a separable Hilbert space $\mathcal{H}$ with inner product denoted by $\langle\cdot,\cdot\rangle$. ε denotes a random error verifying $\mathbb{E}\left(\varepsilon|X_1, \ldots, X_p, \mathcal{X}\right) = 0$. The vector $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^\top \in \mathbb{R}^p$, the functional direction $\theta_0 \in \mathcal{H}$ and the link real-valued function $r(\cdot)$ are supposed unknown. As in the FSIM, it is assumed that an only index of $\mathcal{X}$ is sufficient to summarize all the information carried in $\mathcal{X}$ to predict Y, and in this case, also to predict $X_1, \ldots, X_p$.

The SFPLSIM is the second model that we are going to analyse in this dissertation: Chapter 3 is devoted to contributions related to the estimation task, both from practical and theoretical perspective. In the literature, the model presented in this section was mainly studied in the case in which covariates both in the linear and the semiparametric component are of finite-multidimensional (not functional) nature. That is the case of the partially linear single-index model (PLSIM) introduced in Carroll et al. [26]. The SFPLSIM arises from the need to build a model that takes care of both the functional $\mathcal{X}$ using functional single-index ideas (see, for instance, Ait-Saïdi et al. [3], Chen et al. [27] or Ma [81]) and of the multivariate covariate using partial linear ideas (see e.g. Aneiros-Pérez and Vieu [11] or Feng and Xue [45]). That combination of ideas provides a flexible model with a great advantage in

practice in comparison with the SFPLM: all covariates (functional or not) enter in the model involving an interpretable parameter. Furthermore, the SFPLSIM inherits the characteristic dimension reduction property of the FSIM.

1.4 Sparse regression for scalar response

Another frequent situation in practical applications is the case in which there are a very large number of observed real covariates, p_n, but only a few of them, s_n, have a real effect on the response variable. From a statistical point of view, that is a typical *sparse regression* problem. In this situation, the estimation of the model involves a previous/simultaneous task: using a variable selection method in order to discard the non-influential variables.

In this section we are going to present three sparse models: the first one belongs to the non-functional framework, but it helps to present ideas which will be used later in this dissertation; the second model and the third one also involve a functional covariate and can be seen as an extension of the SFPLM and the SFPLSIM, respectively.

1.4.1 The sparse linear model

Sparse regression problem was firstly studied in the multivariate/high-dimensional context. The *sparse linear model* (SLM) is given by the expression

$$Y = \beta_{00} + X_1\beta_{01} + \cdots + X_{p_n}\beta_{0p_n} + \varepsilon, \tag{1.8}$$

where Y is the real response and $X_1, \ldots, X_{p_n}$ are real random covariates; in addition, $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0p_n})^\top$ is the vector of unknown real parameters. As usual, ε is the random error verifying $\mathbb{E}(\varepsilon|X_1, \ldots, X_{p_n}) = 0$. Moreover, only s_n from the p_n covariates have an influence on the response (that is, only s_n covariates are associated with $\beta_{0j} \neq 0$).

The production in variable selection procedures for model (1.8) started with naive ideas such as stepwise regression (backward (Efroymson [36]), forward (Weisberg [114]) or both), forward-stagewise regression or best subset regression (Furnival and

Wilson [52]). However, these methods are computationally intensive, unstable (see Breiman [20] or Fan and Li [38]) and it is hard to derive sampling properties. They are "discrete procedures" (variables are either selected or discarded), so they often exhibit high variance, and therefore, in some cases they do not reduce the prediction error of the full model. For that, other techniques appeared, like *shrinkage methods* (also known as *regularization*, *penalty-based* or *penalized methods*). Shrinkage procedures are more continuous, and do not suffer as much from high variability (see Hastie et al. [62]). Most of these procedures attempt to select variables automatically and simultaneously (a notorious exception is bridge regression for L_δ norms with $\delta > 1$; see Frank and Friedman [51]). These methods are based on adding a penalization term in the estimation task which generates a sparse solution, in the sense that some estimated coefficients are zero. Penalized methods are highly developed. Having a sample $\{(X_{i1}, \ldots, X_{ip_n}, Y_i)\}_{i=1}^n$ of n vectors i.i.d to $(X_1, \ldots, X_{p_n}, Y)$, the penalized estimator of the vector of unknown parameters, is the solution of the optimization problem

$$\widehat{\boldsymbol{\beta}}_0 = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p_n}} \left(\mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}(\boldsymbol{\beta})\right), \tag{1.9}$$

where $\mathcal{L}(\cdot)$ is a real-valued function which depends on the model and on its estimation procedure; if the estimation is made through penalized least squares, then $\mathcal{L}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip_n})^\top$ $(i = 1, \ldots, n)$. $\mathcal{P}(\cdot)$ is a *penalty function* and depends on a parameter $\lambda > 0$ which controls the amount of penalizations, and then, the sparseness of the resultant vector.

The penalty function employed has a big influence to the properties of the derived estimator (see Fan and Li [38]) and in the literature there are several proposals for this penalization term. Among penalty functions, the majority of them are based on norms. Probably the most famous shrinkage method based on norms was proposed in Tibshirani [104], where L_1 penalty was used $\mathcal{P}(\boldsymbol{\beta}) = \lambda\sum_{j=1}^{p_n}|\beta_j|$. He gave the name *least absolute shrinkage and selection operator* (LASSO) method to the combination of this penalty with the least squares procedure. However, some objections emerged about this penalty. On the one hand, LASSO estimators do not satisfy *oracle properties* (see Fan and Li [38]). On the other hand, Meinshausen and Bühlmann [83] showed that in LASSO the optimal λ for predic-

tion gives inconsistent variable selection results. For that, other penalties were studied. Zou [119] proposed *adaptive* LASSO, where the penalty term has the form $\mathcal{P}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p_n} w_j |\beta_j|$ and w_j with $j = 1, \ldots, p_n$ are known weights. They showed that if the weights are data-dependent and cleverly chosen, then the adaptive LASSO estimators can have the oracle properties. Another famous proposal is the elastic-net penalty (see Zou and Hastie [120]) which is a compromise between L_1 and L_2 penalties: $\mathcal{P}(\boldsymbol{\beta}) = \lambda_2 \sum_{j=1}^{p_n} |\beta_j| + \lambda_1 \sum_{j=1}^{p_n} \beta_j^2$. In a general way, Huang et al. [64] studied *bridge penalties* $\mathcal{P}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p_n} |\beta_j|^\delta$ (related with the L_δ norm) and showed that they verify the oracle property for $0 < \delta < 1$. In addition, a robust approach was studied in Wang et al. [111], where instead of least squares estimation, they used least absolute deviation (LAD) with $\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} |Y_i - \sum_{j=1}^{p_n} \beta_j X_{ij}|$ combined with L_1 penalty (LAD-LASSO).

Probably the main competitor of penalties based on norms is the proposal in Fan [37] later studied in Fan and Li [38]: the *smoothly clipped absolute deviation penalty* (SCAD) defined, for $a > 2$, as

$$\mathcal{P}(\boldsymbol{\beta}) = \sum_{j=1}^{p_n} \mathcal{P}_\lambda(\beta_j) \quad \text{where} \quad \mathcal{P}_\lambda(u) = \begin{cases} \lambda\,|u| & |u| < \lambda, \\ \frac{(a^2-1)\lambda^2 - (|u|-a\lambda)^2}{2(a-1)} & \lambda \leq |u| < a\lambda, \\ \frac{(a+1)\lambda^2}{2} & |u| \geq a\lambda \end{cases} \tag{1.10}$$

(Fan and Li [38] suggested to take $a = 3.7$). SCAD penalty improves properties of L_1 penalty, satisfying the oracle property. For that, it was often used in works related to *generalized linear models* (GLM), in which were assumed that Y_i is a real variable verifying $\mathbb{E}(Y_i|\boldsymbol{X}_i) = g^{-1}(\eta_i)$ with $\eta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ $(i = 1, \ldots, n)$ and where $g(\cdot)$ is a known injective continuous link function. Fan and Li [38] studied GLM and proposed obtaining a penalized log-likelihood estimator using SCAD. That is, the estimator derived from (1.9) when $\mathcal{L}(\boldsymbol{\beta})$ denotes the conditional log-likelihood of Y_i and $\mathcal{P}(\boldsymbol{\beta})$ is the SCAD (1.10).

Although we have focused the exposition in shrinkage methods, other different procedures have been proposed in literature to select relevant variables. In the context of linear modelling, Efron et al. [35] proposed *least angle regression* (LARS) algorithm, a refined version of the forward stagewise procedure that uses a simple

mathematical formula to accelerate the computations. This method is computationally efficient and it has LASSO (LARS-LASSO) and forward stagewise methods as variants. Other different ideas are the *Dantzig selector* proposed in Candès and Tao [23], based on linear programming, or the *sure independence screening* procedure proposed in Fan and Lv [39], based on correlations. And the enumeration of methods could go on.

In general, methodology dealing with estimation-variable selection in model (1.8) can be classified by differencing two situations:

- The context of bounded p_n (that is, $p_n = p$, since the number of covariates not depends on the sample size; standard regression). The mentioned works of Tibshirani [104], Zou [119] or Fan and Li [38] belong to this context.

- The context of divergent p_n (that is, $p_n \to \infty$ as $n \to \infty$; high-dimensional regression). Results in this framework came latter. Here we can cite the mentioned work of Huang et al. [64], or Huang et al. [65] who studied the behaviour of the adaptive LASSO estimator under divergent p_n. We also should mention Fan and Peng [41] and Fan and Lv [40] who studied variable selection in the GLM (of which model (1.8) is a particular case) via penalized log-likelihood with SCAD penalty in this new context, or the mentioned work of Candès and Tao [23].

From a theoretical perspective, in the case of bounded p_n, authors obtain the same rate of convergence for the estimator of the linear coefficients in (1.8) as in standard linear regression ($n^{-1/2}$) (see Fan and Li [38]); while for divergent p_n, under general conditions, various authors obtained the rate $\sqrt{s_n/n}$ (see Fan and Lv [40]).

The introduction of more complex models, containing even functional objects, led to adapt some of the mentioned procedures to perform variable selection in that new framework (see Aneiros et al. [10] for a review in variable selection in functional models). In particular, since the situation of very big p_n started to be common in practice, variable selection methodology adapted to diverging p_n became of great interest, and sparse models are usually studied in this general context.

1.4.2 The sparse semi-functional partial linear model

In model (1.7), let us consider the following two changes: firstly, instead of a fix number of covariates with linear effect (p), assume that we have a high number of real explanatory variables (p_n) and this number of covariates increases with n ($p_n \to \infty$ as $n \to \infty$); secondly, assume that only s_n from the set of p_n real explanatory variables influence the response variable. In this situation we deal with a *sparse semi-functional partial linear model* (SSFPLM):

$$Y = X_1\beta_{01} + \cdots + X_p\beta_{0p_n} + m\left(\mathcal{X}\right) + \varepsilon, \tag{1.11}$$

where, as before, Y is the scalar response, $X_1 \ldots, X_{p_n}$ are real random covariates, $\mathcal{X}$ is the functional random covariate valued in a semi-metric space, $(\beta_{01}, \ldots, \beta_{0p_n})^\top \in \mathbb{R}^{p_n}$ is the vector of unknown parameters, $m(\cdot)$ the non-linear unknown link operator and ε is the random error verifying $\mathbb{E}(\varepsilon|X_1, \ldots, X_{p_n}, \mathcal{X}) = 0$.

Model (1.11) was proposed in Aneiros et al. [7] and it is a combination (in an additive way) of a sparse high-dimensional multivariate predictor with a functional nonparametric one. Aneiros et al. [7] proposed carrying out the variable selection-estimation of (1.11) by means of the penalized least squares method with SCAD penalty. In addition, they showed the existence of a $\sqrt{s_n/n}$-consistent estimator for the vector of linear parameters as well as an oracle property for the variable selection procedure. They also obtained the rate of convergence for the nonparametric estimator of the non-linear functional component.

1.4.3 The sparse semi-functional partial linear single-index model

An alternative of the SSFPLM is the *sparse semi-functional partial linear single-index model* (SSFPLSIM), which is an extension of the SFPLSIM to the case of divergent p_n and only s_n influential variables. The SSFPLSIM is defined by the relationship

$$Y = X_1\beta_{01} + \cdots + X_{p_n}\beta_{0p_n} + r\left(\langle\theta_0, \mathcal{X}\rangle\right) + \varepsilon,$$

where Y denotes a scalar response, $X_1, \ldots, X_{p_n}$ are random covariates taking values in $\mathbb{R}$ and $\mathcal{X}$ is a functional random covariate valued in a separable Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$. In this equation, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^\top \in \mathbb{R}^{p_n}$, $\theta_0 \in \mathcal{H}$ and $r(\cdot)$ are a vector of unknown real parameters, an unknown functional direction and an unknown smooth real-valued function, respectively. Finally, ε is the random error, which verifies $\mathbb{E}(\varepsilon | X_1, \ldots, X_{p_n}, \mathcal{X}) = 0$. As usual, it assumed that an only index of $\mathcal{X}$ is sufficient to summarize all the information carried in $\mathcal{X}$ to predict Y and to predict $X_1, \ldots, X_{p_n}$.

Precisely, the SSFPLSIM proposed in Novo et al. [88] is the third model that we are going to study in this dissertation. Chapter 4 is devoted to contributions on the SSFPLSIM both from theoretical and practical point of view. Note that the SSFPLSIM incorporates the situation of high number of covariates (that is, when $p_n \to \infty$ as $n \to \infty$) and the sparse feature with s_n (*sparseness parameter*) that will be supposed to be much smaller than p_n (see technical assumptions later in Chapter 4). Then, the aim of Chapter 4 is to construct a procedure to select the relevant covariates and, simultaneously, to estimate their corresponding effects, β_{0j}. From a theoretical point of view, the challenge is double: i) obtain consistency of the model selection procedure; ii) get the same rate of convergence for the estimator of $\boldsymbol{\beta}_0$ in the SSFPLSIM as those obtained in the standard literature for the SLM or for the SSFPLM, that is, $O_p(\sqrt{s_n/n})$ (see Sections 1.4.1 and 1.4.2). Once the linear part of the model is dealt with, the functional single-index component $r(\langle \theta_0, . \rangle)$ will be estimated with univariate nonparametric rate of convergence (see Section 4.3).

1.4.4 Sparse regression involving scalar variables with functional origin

In this section we are going to present a modification of the sparse models previously presented (the SLM, the SSFPLM and the SSFPLSIM). The point is that in some situations we have a scalar variable of interest, Y, and we want to know which points of the grid in which is observed a functional variable, namely $\zeta(t)$, are the most influential (*impact points*) on this scalar variable. In other words, we want to select the relevant variables from the set of discretized observations of ζ. The problem is that standard variable selection methods, coming from an adaptation of

the multivariate methodology, can provide inadequate results. On the one hand, these procedures are affected by the strong dependence between variables, which in this case is directly derived from their functional origin. On the other hand, the large number of observations makes it difficult to obtain results in a reasonable amount of time. Therefore, specific methodology has to be developed in these cases.

In Aneiros and Vieu [4], a new method is presented, the *partitioning variable selection* (PVS) procedure, for selecting impact points in the modification of the sparse linear model given by the expression

$$Y = \beta_{00} + \sum_{j=1}^{p_n} \beta_{0j}\zeta(t_j) + \varepsilon, \tag{1.12}$$

where ζ is a random curve defined on some interval $[a, b]$ and is observed in the points $a \leq t_1 < \cdots < t_{p_n} \leq b$ and ε denotes the random error. In addition, $(\beta_{00}, \beta_{01}, \ldots, \beta_{0p_n})^\top$ is a vector of unknown real coefficients. The main idea of the PVS method is to create a two-stage algorithm for selecting relevant variables, taking advantage of the fact that the covariates with linear effect come from a discretization of a curve. In this case, variables that are close in the discretization will contain very similar information about the response.

In Aneiros and Vieu [5], the PVS procedure has been extended to the multi-functional version of the SSFPLM given by the expression

$$Y = \sum_{j=1}^{p_n} \beta_{0j}\zeta(t_j) + m(\mathcal{X}) + \varepsilon, \tag{1.13}$$

where $\mathcal{X}$ denotes a random variable valued on some semi-metric space, $m(\cdot)$ is an unknown non-linear operator and notations used in expression (1.12) are maintained. However, practical requirements of controlling dimensionality and associating interpretable parameters to both functional objects lead us to propose a new model.

This new model, the so-called *multi-functional partial linear single-index model* (MFPLSIM), will be studied in Chapter 5 and is an adaptation of the SSFPLSIM to the case in which covariates with linear effect come from the discretization of a functional variable. In other words, this model assumes that ζ acts only through its p_n discretized points while $\mathcal{X}$ acts in a continuous semiparametric way. That is, the

MFPLSIM is defined by the following relationship:

$$Y = \sum_{j=1}^{p_n} \beta_{0j}\zeta(t_j) + r\left(\langle\theta_0, \mathcal{X}\rangle\right) + \varepsilon,$$

where:

- as before, Y is a real random response and $\mathcal{X}$ denotes a random element belonging to some separable Hilbert space $\mathcal{H}$ with inner product denoted by $\langle\cdot,\cdot\rangle$. The second functional predictor ζ is supposed to be a random curve defined on some interval $[a, b]$ which is observed at the points $a \leq t_1 < \cdots < t_{p_n} \leq b$. In addition, an only index of $\mathcal{X}$ is sufficient to summarize all the information carried in $\mathcal{X}$ to predict Y and $\zeta(t_j)$ $(j = 1, \ldots, p_n)$.

- $(\beta_{01}, \ldots, \beta_{0p_n})^\top$ is a vector of unknown real coefficients and $r(\cdot)$ denotes a smooth unknown link function. In addition, θ_0 is an unknown functional direction in $\mathcal{H}$.

- ε denotes the random error.

In the MFPLSIM (as in models (1.12) and (1.13)), we assume that only a few scalar variables from the set $\{\zeta(t_1), \ldots, \zeta(t_{p_n})\}$ are going to form part of the model.

Our wish in Chapter 5 is to study the MFPLSIM and its associated variable selection problem. At this stage it is worth being pointed that this cannot be done as direct application of methodologies used in the multivariate framework. This is because the variables $\zeta(t_j)$ come from a continuous variable, adding the two following major methodological difficulties in the estimation and the variable selection task. On one hand, the continuous nature of ζ causes strong correlation between them: when t_j is close from t_k then the two corresponding variables $\zeta(t_j)$ and $\zeta(t_k)$ roughly contain the same information about the response Y; then the PLS method presented in Chapter 4 can provide inaccurate results. On the other hand, in many applications p_n is often a very large number; therefore, we deal with a very high-dimensional problem. This has to be added together with the estimation of the direction θ_0, which is usually computationally expensive. This drawback means that neither the PLS procedure and nor even the adaptation of the PVS methodology can be sufficient in

some contexts. These additional difficulties make it crucial to develop specific tools for selecting relevant variables and estimating the MFPLSIM in reasonable feasible computational time. This is what will be presented in Chapter 5.

Chapter 2

Contributions on the functional single-index model

2.1 Introduction

As discussed in Chapter 1, one of the key issues in regression analysis is to build methods combining flexibility and interpretability of the derived estimations. Moreover, these procedures should not be too sensitive to dimensionality effects.

These purposes have been the starting point for many advances around semiparametric modelling, firstly, in multivariate regression analysis (see Härdle et al. [61]), and then, in the functional data framework. In that way, Ferraty et al. [48] and Ait-Saïdi et al. [3] studied the FSIM, briefly presented in Section 1.2.3. Specifically, the FSIM can be written as

$$Y = r\left(\langle \theta_0, \mathcal{X} \rangle\right) + \varepsilon, \tag{2.1}$$

where Y denotes a scalar response $\mathcal{X}$ is a functional explanatory random variable valued in a separable Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$, ε is a random error verifying $\mathbb{E}\left(\varepsilon | \mathcal{X}\right) = 0$, $\theta_0 \in \mathcal{H}$ is the functional index and $r(\cdot)$ is the unknown link function. Regarding the estimation of (2.1), conditions ensuring identifiability in

FSIM have been stated. We assume that

$$\langle \theta_0, \theta_0 \rangle = 1, \tag{2.2}$$

and that for some arbitrary t_0 in the domain of θ_0

$$\theta_0(t_0) > 0 \tag{2.3}$$

(see e.g. Ait-Saïdi et al. [3]).

Ferraty et al. [48] focused on the case of known θ_0 and obtained the pointwise rate of convergence of a kernel estimator of $r(\langle \theta_0, \chi \rangle)$, where $\chi \in \mathcal{H}$. The case of unknown θ_0 was covered in Ait-Saïdi et al. [3], where both the consistency and optimality of a cross-validation-based estimator of θ_0 were proved. In addition, the FSIM (2.1) was extended in different directions, see Bouraine et al. [18], Chen et al. [27], Ferraty et al. [50], Ma [81] and Wang et al. [110], among others.

This chapter presents a comprehensive study of the functional semiparametric model FSIM (2.1). Section 2.2 develops a new automatic and location-adaptive procedure for estimating regression in the FSIM based on kNN ideas. Section 2.3 states general asymptotic results for the kNN procedure, with the main interest of being uniform over all the parameters of the model. As discussed in Section 2.3.4, results for random data-driven choices of these parameters can be derived from this uniformity feature, making our procedure directly applicable in practice. Although our main goal is to study kNN procedures, we also get similar results for the standard kernel approach throughout Section 2.3. The main feature of the rates of convergence obtained is that they are similar to those achieved in one-dimensional problems, which shows the dimensionality reduction property of the method. Suggestions to address some practical issues related to the proposed methodology are shown in Section 2.4. These suggestions are supported in Section 2.5 by means of a simulation study which also compares the performance of the kNN- and kernel-based procedures. Section 2.6 illustrates, through some benchmark real curves dataset, how the kNN approach outperforms standard procedures. It also shows that the semiparametric feature of the FSIM has not only nice predictive performance, but it also provides easily interpretable and representable outputs. Finally, the proofs of the main results

are presented in Section 2.7.

2.2 The statistics

Let $\{(\mathcal{X}_i, Y_i)\}_{i=1}^n$ be a sample of n pairs i.i.d. as $(\mathcal{X}, Y)$, which verifies the FSIM (2.1); that is,

$$Y_i = r\left(\langle \theta_0, \mathcal{X}_i \rangle\right) + \varepsilon_i \quad (i = 1, \ldots, n).$$

For any $\theta \in \mathcal{H}$, we consider the operator

$$r_\theta(\cdot) : \mathcal{H} \longrightarrow \mathbb{R}$$

defined as

$$r_\theta(\chi) = r\left(\langle \theta, \chi \rangle\right), \quad \forall \chi \in \mathcal{H}, \tag{2.4}$$

and we define the *projection semi-metric* as follows

$$d_\theta\left(\chi_1, \chi_2\right) = \left|\langle \theta, \chi_1 - \chi_2 \rangle\right|, \quad \text{for } \chi_1, \chi_2 \in \mathcal{H}. \tag{2.5}$$

For each direction θ, we construct the kNN (Nadaraya-Watson type) *statistic*[1] as

$$\widehat{r}^*_{k,\theta}(\chi) = \sum_{i=1}^n w^*_{n,k,\theta}(\chi, \mathcal{X}_i) Y_i, \quad \forall \chi \in \mathcal{H}, \tag{2.6}$$

where we have denoted

$$w^*_{n,k,\theta}(\chi, \mathcal{X}_i) = \frac{K\left(H^{-1}_{k,\chi,\theta} d_\theta\left(\mathcal{X}_i, \chi\right)\right)}{\sum_{i=1}^n K\left(H^{-1}_{k,\chi,\theta} d_\theta\left(\mathcal{X}_i, \chi\right)\right)}, \tag{2.7}$$

being $k \in \mathbb{Z}^+$ a smoothing factor ($k = k_n$ depends on n) and K a kernel. In addition, we have denoted

$$H_{k,\chi,\theta} = \min\left\{ h \in \mathbb{R}^+ \text{ such that } \sum_{i=1}^n 1_{B_\theta(\chi,h)}(\mathcal{X}_i) = k \right\} \tag{2.8}$$

[1]Note that we use the term *statistic* instead of *estimator* since expression (2.6) (and likewise expression (2.10)) depends on the unknown parameter θ.

with

$$B_\theta(\chi, h) = \{z \in \mathcal{H} : d_\theta(\chi, z) \leq h\}. \tag{2.9}$$

The kNN statistic $\widehat{r}^*_{k,\theta}$ can be seen as an extension of the usual kernel statistic

$$\widehat{r}_{h,\theta}(\chi) = \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) Y_i \quad \forall \chi \in \mathcal{H}, \tag{2.10}$$

where

$$w_{n,h,\theta}(\chi, \mathcal{X}_i) = \frac{K\left(h^{-1} d_\theta(\mathcal{X}_i, \chi)\right)}{\sum_{i=1}^{n} K\left(h^{-1} d_\theta(\mathcal{X}_i, \chi)\right)}, \tag{2.11}$$

with $h \in \mathbb{R}^+$ being the bandwidth ($h = h_n$ depends on n). Note that the main difference between the statistics defined in expressions (2.6) and (2.10) and nonparametric estimators defined for the FNM (1.4) and (1.3), respectively, is the semi-metric: in the case of the FSIM we are forced to use the projection semi-metric (2.5), which involves the unknown parameter θ.

The kNN statistic presents some advantages in practice compared to the kernel one as formulated in Section 1.2.2 and as will be seen in practical applications in this chapter and in Chapter 3. Basically, the kNN statistic allows adaptation to heterogeneous designs (since $H_{k,\chi,\theta}$ varies with χ) and the smoothing parameter k is easier to select. However, its theoretical properties are more difficult to analyse. These features of kNN estimates have been widely highlighted in one-dimensional problems (see Györfi et al. [58] for a general discussion), but very little progress has been made in the framework of functional regression. The existing literature on kNN functional regression mainly focuses on nonparametric modelling (see Biau et al. [15], Kudraszow and Vieu [71], Müller and Dippon [85], Kara-Zaitri et al. [66] and Ling et al. [79] for the most recent advances, and see Ling and Vieu [76] for an exhaustive survey) or partial linear modelling (see Ling et al. [77]), but to our knowledge this chapter states the first advances in functional semiparametric regression.

In Section 2.3 we provide a complete study of the kNN procedure in the semiparametric model (2.1). The main idea is to establish asymptotic results in a uniform sense over all the parameters of the method (that is, over the direction θ and over the smoothing factor k). For that, we will follow the uniform in bandwidth ideas widely developed in the non-functional setting (see e.g. Dony and Einmahl [34]) and

recently adapted to the functional setting (see Kara-Zaitri et al. [67]). However, we will have to include suitable technical changes to adapt such ideas to both the kNN procedures and the infinite-dimensional parameter θ. Note that, although our main purpose is to study the kNN procedure, we also get a full asymptotic analysis of the standard kernel statistic (2.10) as a by-product, extending earlier results of Ferraty et al. [48], Ait-Saïdi et al. [3], Chen et al. [27], Ferraty et al. [50].

2.3 Asymptotic theory

2.3.1 Presentation and general notation

Section 2.3.2 begins by presenting the uniform in bandwidth (UIB) and uniform in the number of neighbours (UINN) consistency of the statistics $\widehat{r}_{h,\theta}(\chi)$ (2.10) and $\widehat{r}^*_{k,\theta}(\chi)$ (2.6), respectively, when θ is fixed. Then, Section 2.3.3 extends these asymptotics by also providing uniform consistency over the functional parameter θ.

Let us first introduce some terms and notation:

- Throughout this dissertation, χ denotes a fixed element in $\mathcal{H}$ while θ is some direction in $\mathcal{H}$.

- Furthermore, note that in the infinite-dimensional space $\mathcal{H}$, a universal reference measure is not available (in contrast to finite-dimensional spaces where Lebesgue measure is taken as a reference). For that, the density function of the variable $\mathcal{X}$ might not exist. One way to overcome this problem is to use small-ball probability considerations (see Kara-Zaitri et al. [67]). For that, let us define

$$\phi_{\chi,\theta}(h) = \mathbb{P}\left(d_\theta(\mathcal{X}, \chi) \leq h\right).$$

  The function $\phi_{\chi,\theta}(\cdot)$ controls the concentration of the data in the functional space and it is usually known in literature as *small-ball probability function*. This function has a direct impact on the variance of the estimators derived from (2.6) and (2.10). For more details on the topological structure of the functional space and its links to the small-ball probability function see Ferraty and Vieu [47].

- In addition, to present the asymptotic results we will need to define the following class of functions

$$\mathcal{K}_\theta = \left\{ \cdot \longrightarrow K\left(h^{-1} d_\theta(\chi, \cdot)\right),\ h > 0 \right\}, \tag{2.12}$$

for each $\theta \in \Theta_n$, where $\Theta_n \subset \mathcal{H}$ is the set of directions of interest. The class (2.12) is contained in the class

$$\mathcal{K}_{\Theta_n} = \cup_{\theta \in \Theta_n} \mathcal{K}_\theta = \left\{ \cdot \longrightarrow K\left(h^{-1} d_\theta(\chi, \cdot)\right),\ h > 0,\ \theta \in \Theta_n \right\} \tag{2.13}$$

(note that both $\mathcal{K}_\theta$ and $\mathcal{K}_{\Theta_n}$ are classes of functions that should satisfy conditions (2.23) and (2.35), respectively; in addition, Assumption (2.25) allows the size of Θ_n to grow to infinite as n does).

- Moreover, let $\mathcal{Q}$ be a probability measure on the space $(\mathcal{H}, \mathcal{A})$. Then, $||\cdot||_{\mathcal{Q},2}$ means the norm $L_2(\mathcal{Q})$ defined on certain space $S = \{f : \mathcal{H} \longrightarrow \mathbb{R}\}$, while $d_{\mathcal{Q},2}(\cdot,\cdot)$ is the metric associated to the norm $L_2(\mathcal{Q})$; that is, for $f, g \in S$,

$$||f||_{\mathcal{Q},2} = \left(\int_{\mathcal{H}} |f(t)|^2 \, d\mathcal{Q}(t) \right)^{\frac{1}{2}}$$

and

$$d_{\mathcal{Q},2}(f,g) = ||f-g||_{\mathcal{Q},2} = \left(\int_{\mathcal{H}} |f(t)-g(t)|^2 \, d\mathcal{Q}(t) \right)^{\frac{1}{2}}.$$

- Finally, given a metric space $(\mathcal{K}, d)$, $\mathcal{N}(\epsilon, \mathcal{K}, d)$ denotes the minimal number of open balls (in the topological space given by d) with radius ϵ which are needed to cover $\mathcal{K}$. The quantity $\log(\mathcal{N}(\epsilon, \mathcal{K}, d))$ is called *Kolmogorov's ϵ-entropy* of the set $\mathcal{K}$. This term was introduced by Kolmogorov and Tikhomirov [69] and represents a measure of the complexity of the set: high entropy means that a lot of information is needed to describe an element with accuracy ϵ (see Ferraty et al. [49]).

2.3.2 The case of θ_0 known

Let us assume that the true direction, θ_0, in the FSIM (2.1) is known. In order to state the UIB and the UINN almost-complete convergence of the estimators $\widehat{r}_{h,\theta_0}(\chi)$ and $\widehat{r}^*_{k,\theta_0}(\chi)$, some of the following assumptions will be used.

2.3.2.1 Assumptions for UIB and UINN consistency

About the small-ball probability. Let us assume that:

- For all $h > 0$,
$$\phi_{\chi,\theta_0}(h) > 0. \tag{2.14}$$
That is, we assume that if we fix $\chi \in \mathcal{H}$, the probability that the variable takes values in the ball of centre χ and radius h created with the projection semi-metric (2.5) is greater than zero.

- There exist a constant $0 < C_1$ and sequences $\{a_n\}, \{b_n\} \subset \mathbb{R}^+$ $(a_n \leq b_n)$ such that, for $h \in [a_n, b_n]$ with n large enough,
$$C_1 \leq \frac{\phi_{\chi,\theta_0}(h/2)}{\phi_{\chi,\theta_0}(h)}. \tag{2.15}$$
With this assumption we ensure that when we halve the radius of the projection semi-metric ball, the probability of finding sample elements does not drop drastically to zero.

- The sequences $\{a_n\}$ and $\{b_n\}$ verify:
$$a_n \to 0,\ b_n \to 0 \text{ and } \frac{\log n}{n \min\{a_n, \phi_{\chi,\theta_0}(a_n)\}} \to 0. \tag{2.16}$$
These hypotheses ensure that the convergence rate in Proposition 2.2 (a) tends to zero.

- There exist sequences $\{\rho_n\} \subset (0,1)$, $\{k_{1,n}\} \subset \mathbb{Z}^+$ and $\{k_{2,n}\} \subset \mathbb{Z}^+$ $(k_{1,n} \leq k_{2,n} \leq n)$ such that
$$\phi^{-1}_{\chi,\theta_0}\left(\frac{k_{2,n}}{\rho_n n}\right) \to 0, \tag{2.17}$$

$$\min\left\{\frac{1-\rho_n}{4}\frac{k_{1,n}}{\ln n}, \frac{(1-\rho_n)^2}{4\rho_n}\frac{k_{1,n}}{\ln n}\right\} > 2 \tag{2.18}$$

and

$$\frac{\log n}{n\min\left\{\phi^{-1}_{\chi,\theta_0}(\rho_n k_{1,n}/n), \rho_n k_{1,n}/n\right\}} \to 0 \tag{2.19}$$

Hypotheses (2.17) and (2.19) ensure that convergence rate in Proposition 2.2 (b) tends to zero.

About the model. We assume that:

- There exist constants $\beta_0 > 0$ and $C_3 > 0$, such that:

$$\forall \chi_1, \chi_2 \in N_{\chi,\theta_0}, \quad |r_{\theta_0}(\chi_1) - r_{\theta_0}(\chi_2)| \leq C_3 d_{\theta_0}(\chi_1, \chi_2)^{\beta_0}, \tag{2.20}$$

  where N_{χ,θ_0} denotes a fixed neighbourhood of χ in the topological space induced by the semi-metric $d_{\theta_0}(\cdot,\cdot)$. That is, it is assumed that the objective regression is Hölder continuous, which will have a direct impact on the bias of the estimators.

- There exist constants $m \geq 2$ and $C_4 > 0$, such that:

$$\mathbb{E}\left(|Y|^m|\mathcal{X}\right) < C_4 < \infty, \ a.s. \tag{2.21}$$

About the kernel. We assume that:

- For all $u \in (0, 1/2)$, there exist constants $0 < C_5 \leq C_6 < \infty$, such that:

$$C_5 \leq K(u) \leq C_6 \tag{2.22}$$

  and for all $u \notin (0, 1/2)$, $K(u) = 0$. This condition is satisfied by the usual discontinuous (asymmetrical) kernels.

- The class of functions $\mathcal{K}_{\theta_0}$ (see (2.12)) is a *pointwise measurable class*[2]

[2]A class of functions $\mathcal{K}$ is said to be *pointwise measurable* if there exists a countable subclass $\mathcal{K}_0$, such that for any function $f \in \mathcal{K}$, there exists a sequence of functions $\{f_m\}$ in $\mathcal{K}_0$ such that: $|f_m(z) - f(z)| = o(1)$ (see Kara-Zaitri et al. [67]).

such that

$$\sup_{\mathcal{Q}} \int_0^1 \sqrt{1+\log \mathcal{N}\left(\epsilon||F_{\theta_0}||_{\mathcal{Q},2}, \mathcal{K}_{\theta_0}, d_{\mathcal{Q},2}\right)} d\epsilon < \infty, \tag{2.23}$$

where F_{θ_0} is the minimal *envelope function*[3] of the set $\mathcal{K}_{\theta_0}$ and the supremum is taken over all probability measures $\mathcal{Q}$ on the measurable space $(\mathcal{H}, \mathcal{A})$ with $||F_{\theta_0}||^2_{\mathcal{Q},2} < \infty$. Note that this condition is a uniform integral entropy condition used to characterise the Donsker-class of functions (see van der Vaart and Wellner [106]); it allows to derive a uniform limit distribution and is useful for evaluating moments of empirical processes (see Kara-Zaitri et al. [66]).

Remark 2.1 *Assumptions (2.14), (2.15) and (2.20)-(2.22) are standard ones in the setting of functional nonparametric regression models (see Ferraty and Vieu [47]), while assumptions (2.16) and (2.23) are usual to obtain UIB consistency in such setting (see Kara-Zaitri et al. [67]). In fact, assumptions (2.16) and (2.23) adapt those used in Kara-Zaitri et al. [67] to the case where the nonparametric regression function is $r_{\theta_0}(\cdot)$ and the semi-metric to use in the kernel estimator is $d_{\theta_0}(\cdot,\cdot)$. Focusing now on the UINN consistency, assumptions (2.17)-(2.19) adapt (in the same way as in the previous case of UIB consistency) and correct those used in Kara-Zaitri et al. [66]. Specifically, in Kara-Zaitri et al. [66], they forgot to include the parameter α in their expression (17). As a consequence, Assumption (H4) in Kara-Zaitri et al. [66] should be modified in the way of our assumptions (2.17)-(2.19), where the notation ρ_n was considered instead of α; in addition, α should be introduced in the rates of convergence corresponding to their Theorem 3.1 in the same way as ρ_n in our Theorem 2.2(b). The justification for these changes in both the assumptions and the rates of convergence in Kara-Zaitri et al. [66] can be seen in the proof of our Theorem 2.5(b). Finally, in the particular case of assumptions (2.15) and (2.22), it is worth noting that they are even weaker than the corresponding ones in Kara-Zaitri et al. [66, 67].*

[3] An *envelope function* F for a class of functions $\mathcal{K}$ is any measurable function such that: $\sup_{f\in\mathcal{K}} |f(z)| \leq F(z)$ (see Kara-Zaitri et al. [67]).

2.3.2.2 Result

Our first result states the UIB and UINN consistency of the estimators $\widehat{r}_{h,\theta_0}(\chi)$ and $\widehat{r}^*_{k,\theta_0}(\chi)$, respectively, of $r_{\theta_0}(\chi)$. The type of convergence considered is *almost complete* convergence (a.co.)[4].

Proposition 2.2 *Let us assume that conditions (2.14), (2.15) and (2.20)-(2.23) hold.*

(a) If Assumption (2.16) also holds, then we have that

$$\sup_{a_n\leq h\leq b_n}|\widehat{r}_{h,\theta_0}(\chi)-r_{\theta_0}(\chi)|=O\left(b_n^{\beta_0}\right)+O_{a.co.}\left(\sqrt{\frac{\log n}{n\phi_{\chi,\theta_0}(a_n)}}\right).$$

(b) If assumptions (2.17)-(2.19) also hold, then we have that

$$\sup_{k_{1,n}\leq k\leq k_{2,n}}|\widehat{r}^*_{k,\theta_0}(\chi)-r_{\theta_0}(\chi)|=O\left(\phi^{-1}_{\chi,\theta_0}\left(\frac{k_{2,n}}{\rho_n n}\right)^{\beta_0}\right)+O_{a.co.}\left(\sqrt{\frac{\log n}{\rho_n k_{1,n}}}\right).$$

Remark 2.3 *Proposition 2.2(a) extends Theorem 3.1 in Ferraty et al. [48] to the case where h varies in an interval $[a_n, b_n]$ (Ferraty et al. [48] focused on the case $a_n = b_n = h$). This fact represents a very important improvement because, as will be shown in Section 2.3.4, one of the applications of Proposition 2.2(a) is the validation of data-driven bandwidth selectors from an asymptotic point of view. In the same way, Proposition 2.2(b) will be used in Section 2.3.4 to validate data-driven selectors for the number of neighbours.*

[4]For sequences of real random variables and positive real numbers, $\{Z_n\}$ and $\{u_n\}$, respectively, it says that $Z_n = O_{a.co.}(u_n)$ if, and only if

$$\exists\eta_0>0\quad\sum_{n=1}^{\infty}\mathbb{P}(|Z_n|>\eta_0 u_n)<\infty. \tag{2.24}$$

This kind of convergence implies almost-sure convergence (and then, it also implies convergence in probability). For more details, see Appendix in Ferraty and Vieu [47].

2.3.3 The case of θ_0 unknown

In practice, the direction θ_0 is usually unknown, so it must be estimated. The results that will be presented in this section, the uniform in both bandwidth and direction (UIBD) and in both number of neighbours and direction (UINND) consistency of $\widehat{r}_{h,\theta}(\chi)$ and $\widehat{r}^*_{k,\theta}(\chi)$, respectively, play a main role to study the asymptotic behaviour of $\widehat{r}_{\widehat{h}}(\langle\widehat{\theta},\chi\rangle) := \widehat{r}_{\widehat{h},\widehat{\theta}}(\chi)$ and $\widehat{r}^*_{\widehat{k}}(\langle\widehat{\theta},\chi\rangle) := \widehat{r}^*_{\widehat{k},\widehat{\theta}}(\chi)$, where $\widehat{h}$ and $\widehat{k}$ denote some appropriate selectors for h and k, respectively, while $\widehat{\theta}$ is a suitable estimator of θ_0. First, we list some additional assumptions that we will use to establish such results.

2.3.3.1 Additional assumptions for UIBD and UINND consistency

About the space of directions. We assume that

$$\text{card}(\Theta_n) = n^\alpha \quad \text{with} \quad \alpha > 0, \tag{2.25}$$

which means that the number of directions contained in Θ_n depends on the sample size and converges to infinity at an algebraic rate. In addition,

$$\forall\ \theta \in \Theta_n,\ \langle\theta - \theta_0, \theta - \theta_0\rangle^{1/2} \leq C_7 b_n. \tag{2.26}$$

That is, the elements of Θ_n are relatively close to the target direction θ_0.

About the functional explanatory variable. We assume that the functional covariate is bounded in the sense that

$$\langle\mathcal{X},\mathcal{X}\rangle^{1/2} \leq C_8 \tag{2.27}$$

(remember that $\langle\cdot,\cdot\rangle$ denotes the inner product associated with $\mathcal{H}$).

About the small-ball probability. We assume that:

- There exist constants $0 < C_9 \leq C_{10} < \infty$ and a function $f : \mathbb{R} \longrightarrow (0,\infty)$ such that

$$\forall\theta \in \Theta_n,\ C_9 f(h) \leq \phi_{\chi,\theta}(h) \leq C_{10} f(h). \tag{2.28}$$

(Actually, it might be the case that $f(\cdot) = f_\chi(\cdot)$. In the sake of brevity, we have not added the subscript.) In this way, it is assumed that there exist common lower and upper bounds for the small-ball probability functions associated to each $\theta \in \Theta_n$ (that is, uniform bounds on $\theta \in \Theta_n$).

- There exist constants $0 < C_{11} \leq C_{12} < \infty$ and sequences $\{a_n\}, \{b_n\} \subset \mathbb{R}^+$ $(a_n \leq b_n)$ such that, for $h \in [a_n, b_n]$ with n large enough,

$$C_{11} \leq \frac{f(h/2)}{f(h)} \leq C_{12}. \tag{2.29}$$

  Note that this hypothesis is an adaptation of Assumption (2.15) to the case of unknown θ_0.

- The sequences $\{a_n\}$ and $\{b_n\}$ verifies:

$$a_n \to 0, b_n \to 0 \text{ and } \frac{\log n}{n \min\{a_n, f(a_n)\}} \to 0. \tag{2.30}$$

  These assumptions generalize hypotheses (2.16) and ensure that convergence rate in Theorem 2.5 (a) tends to zero.

- There exist sequences $\{\rho_n\} \subset (0,1)$, $\{k_{1,n}\} \subset \mathbb{Z}^+$, $\{k_{2,n}\} \subset \mathbb{Z}^+$ $(k_{1,n} \leq k_{2,n} \leq n)$ and constants $0 < \lambda \leq \delta < \infty$ such that

$$\lambda f^{-1}\left(\frac{\rho_n k_{1,n}}{n}\right) \leq \phi^{-1}_{\chi,\theta}\left(\frac{\rho_n k_{1,n}}{n}\right) \text{ and } \phi^{-1}_{\chi,\theta}\left(\frac{k_{2,n}}{\rho_n n}\right) \leq \delta f^{-1}\left(\frac{k_{2,n}}{\rho_n n}\right), \tag{2.31}$$

  (we also need to assume lower and upper bounds for the inverse of the small-ball probability function (for each $\theta \in \Theta_n$ and at those particular points) involving $f^{-1}(\cdot)$, since both functions will be used in proofs for the kNN statistic)

$$f^{-1}\left(\frac{k_{2,n}}{\rho_n n}\right) \to 0, \tag{2.32}$$

$$\min\left\{\frac{1-\rho_n}{4}\frac{k_{1,n}}{\ln n}, \frac{(1-\rho_n)^2}{4\rho_n}\frac{k_{1,n}}{\ln n}\right\} > \alpha + 2 \tag{2.33}$$

and

$$\frac{\log n}{n \min\left\{\lambda f^{-1}(\rho_n k_{1,n}/n), f\left(\lambda f^{-1}(\rho_n k_{1,n}/n)\right)\right\}} \to 0. \tag{2.34}$$

Note that assumptions (2.32), (2.33) and (2.34) are adaptations of assumptions (2.17), (2.18) and (2.19), respectively, to the case of unknown θ_0. In addition, conditions (2.32) and (2.34) ensure that the convergence rate in Theorem 2.5 (b) tends to zero.

About the kernel. The class of functions $\mathcal{K}_{\Theta_n}$ (see (2.13)) is a pointwise measurable class such that

$$\sup_{\mathcal{Q}} \int_0^1 \sqrt{1 + \log \mathcal{N}\left(\epsilon \|F_{\Theta_n}\|_{\mathcal{Q},2}, \mathcal{K}_{\Theta_n}, d_{\mathcal{Q},2}\right)} d\epsilon < \infty, \tag{2.35}$$

where F_{Θ_n} is the minimal envelope function of the set $\mathcal{K}_{\Theta_n}$ and the supremum is taken over all probability measures $\mathcal{Q}$ on the measurable space $(\mathcal{H}, \mathcal{A})$ with $||F_{\Theta_n}||^2_{\mathcal{Q},2} < \infty$. Note that in case of unknown θ_0, we need to impose the uniform integral entropy condition to the class $\mathcal{K}_{\Theta_n}$ (2.13) instead of only to the class $\mathcal{K}_{\theta_0}$ as in (2.23).

Remark 2.4 *Assumption (2.25) imposes that the set of directions Θ_n contains a finite number of directions, but allows it to grow to infinity as the sample size increases. In addition, taking into account the kind of results we want to establish (UIBD and UINND consistency; see Theorem 2.5), it is necessary to impose some condition to control the bias caused by the use of values $\theta \in \Theta_n$ different from the true value θ_0 in the studied statistics. In particular, such condition should allow to link the behaviour of $d_\theta(\cdot,\cdot)$ and $d_{\theta_0}(\cdot,\cdot)$ (for details, see the proof of Lemma 2.17). In this chapter this is done by means of Assumption (2.26). Note that, on the one hand, Assumption (2.26) implies that the larger n is, the closer Θ_n and θ_0 are; this is needed to obtain uniform consistency results on Θ_n. On the other hand, the order b_n in Assumption (2.26) is a technical condition (the minimal one when our proof is used) that allows to obtain the same rates of convergence as in the case of $\Theta_n = \{\theta_0\}$ (see Proposition 2.2). The interested reader can find similar conditions to our Assumption (2.26) in Härdle et al. [60] and Xia and Li [115] (multivariate setting), and Ma [81] (functional setting; see also Ferraty et al. [50] for a different version of*

Assumption (2.26)). Assumption (2.27), which imposes that the explanatory variable is bounded, is not very restrictive in practice and is introduced to make the proofs easier. The role of Assumption (2.28) is to ensure uniform results among all possible directions; that assumption generalizes Assumption (4) in Ait-Saïdi et al. [3] and was also used in Wang et al. [110]. Assumption (2.29) is weaker than the usual condition

$$0 < \lim_{h\to 0} \frac{f(sh)}{f(h)} = \tau(s) < \infty, \quad \forall s \in (0,1)$$

(considered, for instance, in Kara-Zaitri et al. [66, 67]). The technical assumptions (2.30) and (2.31)-(2.34) adapt those considered in Kara-Zaitri et al. [66, 67] (in the context of functional nonparametric regression), respectively, to the setting of the FSIM (2.1) (remember that, as noted in the last paragraph in Section 2.3.2.1, Assumption (H4) in Kara-Zaitri et al. [66] should be modified in the way of our assumptions(2.17)-(2.19)). Note that assumptions (2.28)-(2.34) (the ones related to the small-ball probability), although technical, are not very restrictive. For instance, Wang et al. [110] showed that, under suitable conditions, $\phi_{\chi,\theta}(h) \approx C_{\chi,\theta} h$. Therefore, one can consider $f(h) = h$. Then, for such functions $\phi_{\chi,\theta}(\cdot)$ and $f(\cdot)$, assumptions (2.28), (2.29) and (2.31) are trivially verified while assumptions (2.30), (2.32) and (2.34) are satisfied under the conditions

$$\frac{\log n}{n a_n} \to 0, \quad \frac{k_{2,n}}{\rho_n n} \to 0 \quad \textit{and} \quad \frac{\log n}{\rho_n k_{1,n}} \to 0,$$

respectively. In addition, for Assumption (2.33) to be verified it is sufficient that the condition

$$(1-\rho_n)^2 > 4(\alpha+2)\frac{\ln n}{k_{1,n}}$$

holds (note that none of those three conditions are restrictive and they allow that $\rho_n \to 1$). Finally, Assumption (2.35) is a natural extension of Assumption (2.23) to the current case where card$(\Theta_n) > 1$.

2.3.3.2 Main results

Theorem 2.5 below states the UIBD and UINND consistency of $\widehat{r}_{h,\theta}(\chi)$ and $\widehat{r}^*_{k,\theta}(\chi)$, respectively, under general assumptions while, to fix the ideas, Corollary 2.6 shows

how the rates of convergence behave in some simple case. In particular, as will be seen throughout Remark 2.7, these rates of convergence are similar to the optimal ones for one-dimension problems. This fact evidences that the main goal of constructing procedures insensitive to the dimensionality of the problem has been achieved.

Theorem 2.5 *Let us assume that conditions (2.20)-(2.22), (2.25)-(2.29) and (2.35) hold.*

(a) If Assumption (2.30) also holds, then we have that

$$\sup_{\theta\in\Theta_n}\sup_{a_n\leq h\leq b_n}|\widehat{r}_{h,\theta}(\chi)-r_{\theta_0}(\chi)|=O\left(b_n^{\beta_0}\right)+O_{a.co.}\left(\sqrt{\frac{\log n}{nf(a_n)}}\right).$$

(b) If assumptions (2.31)-(2.34) also hold, then we have that

$$\begin{aligned}\sup_{\theta\in\Theta_n}\sup_{k_{1,n}\leq k\leq k_{2,n}}\left|\widehat{r}^*_{k,\theta}(\chi)-r_{\theta_0}(\chi)\right| &= O\left(f^{-1}\left(\frac{k_{2,n}}{\rho_n n}\right)^{\beta_0}\right)\\ &+ O_{a.co.}\left(\sqrt{\frac{\log n}{nf\left(\lambda f^{-1}(\rho_n k_{1,n}/n)\right)}}\right).\end{aligned}$$

Corollary 2.6 *Let us assume that conditions (2.20)-(2.22), (2.25)-(2.27) and (2.35) hold. If in addition assumptions (2.28) and (2.31) hold with $f(h)=h$ and $\rho_n=\rho$ (where $0<\rho<1$ is a constant), and $k_{2,n}/n\to 0$ and $\log n/k_{1,n}\to 0$, then we have that*

$$\sup_{\theta\in\Theta_n}\sup_{k_{1,n}\leq k\leq k_{2,n}}\left|\widehat{r}^*_{k,\theta}(\chi)-r_{\theta_0}(\chi)\right|=O\left(\left(\frac{k_{2,n}}{n}\right)^{\beta_0}\right)+O_{a.co.}\left(\sqrt{\frac{\log n}{k_{1,n}}}\right).$$

Remark 2.7 *Theorem 2.5 extends Proposition 2.2 to the case where θ_0 is unknown. As can be noted in Theorem 2.5(b), the parameters λ and ρ_n (defined in assumptions (2.31)-(2.34)) affect to the rate of convergence of the kNN estimators. Actually, that is a consequence of having formulated our assumptions about $f(\cdot)$ in a fairly general way. Corollary 2.6 shows that, under the weak condition $f(h)=h$ (see the comments on assumptions in Section 2.3.3.1), these effects disappear. Focusing now on the specific case $f(h)=h$, let us take*

- $h_o \sim (\log n/n)^{1/(2\beta_0+1)}, \quad k_o \sim (n^{2\beta_0}\log n)^{1/(2\beta_0+1)},$
- $a_n = h_o - c_n, \quad b_n = h_o + c_n,$ *with* $c_n = c(\log n/n)^{1/(2\beta_0+1)} \quad 0 < c < 1,$
- $k_{1,n} = k_o - d_n, \quad k_{2,n} = k_o + d_n$ *with* $d_n = c(n^{2\beta_0}\log n)^{1/(2\beta_0+1)} \quad 0 < c < 1.$

Then, it can be seen from Theorem 2.5(a) and Corollary 2.6 that both estimates reach the rate of convergence

$$\left(\frac{\log n}{n}\right)^{\beta_0/(2\beta_0+1)},$$

which is the well-known optimal rate for nonparametric one-dimensional problems. This attests to the dimensionality reduction property of our model and estimates.

2.3.4 Data-driven parameters selection

An application of Theorem 2.5(a) (Theorem 2.5(b)) is related to the theoretical validation of both data-driven selectors for the bandwidth h (for the number of neighbours k) and estimators for the direction θ_0. Next result, which is a corollary of Theorem 2.5, focuses on data-driven selectors based on cross-validation ideas (similar results can be derived for other usual selectors).

Let us denote

$$CV(h,\theta) = n^{-1}\sum_{j=1}^{n}\left(Y_j - \widehat{r}_{h,\theta}^{(-j)}(\mathcal{X}_j)\right)^2 \text{ and } CV^*(k,\theta) = n^{-1}\sum_{j=1}^{n}\left(Y_j - \widehat{r}_{k,\theta}^{*(-j)}(\mathcal{X}_j)\right)^2,$$

where, as usual, $\widehat{r}_{h,\theta}^{(-j)}(\cdot)$ and $\widehat{r}_{k,\theta}^{*(-j)}(\cdot)$ are the leave-one-out versions of $\widehat{r}_{h,\theta}(\cdot)$ and $\widehat{r}_{k,\theta}^{*}(\cdot)$, respectively. Then, one can consider the kernel-based estimator of θ_0

$$\widehat{\theta}_h = \arg\min_{\theta\in\Theta_n} CV(h,\theta)$$

(for asymptotic properties of $\widehat{\theta}_h$, see Ait-Saïdi et al. [3]) and the kNN-based estimator

$$\widehat{\theta}_k^* = \arg\min_{\theta\in\Theta_n} CV^*(k,\theta).$$

Following the same ideas, it seems natural to construct the data-driven selectors $\widehat{h}$

and $\widehat{k}$ as

$$\widehat{h} = \arg \min_{a_n \leq h \leq b_n} CV\left(h, \widehat{\theta}_h\right) \text{ and } \widehat{k} = \arg \min_{k_{1,n} \leq k \leq k_{2,n}} CV^*\left(k, \widehat{\theta}_k\right),$$

respectively. In that way, we have two automatic estimators of θ_0: one based on kernel estimation, $\widehat{\theta}_{\widehat{h}}$, and another based on kNN ideas, $\widehat{\theta}^*_{\widehat{k}}$.

Corollary 2.8 *(a) Under assumptions of Theorem 2.5(a), we have that*

$$\left|\widehat{r}_{\widehat{h},\widehat{\theta}_{\widehat{h}}}(\chi) - r_{\theta_0}(\chi)\right| = O\left(b_n^{\beta_0}\right) + O_{a.co.}\left(\sqrt{\frac{\log n}{n f(a_n)}}\right).$$

(b) Under assumptions of Corollary 2.6, we have that

$$\left|\widehat{r}^*_{\widehat{k},\widehat{\theta}^*_{\widehat{k}}}(\chi) - r_{\theta_0}(\chi)\right| = O\left(\left(\frac{k_{2,n}}{n}\right)^{\beta_0}\right) + O_{a.co.}\left(\sqrt{\frac{\log n}{k_{1,n}}}\right).$$

Remark 2.9 *Corollary 2.8 validates the use of cross-validation ideas to construct both estimators of the direction θ_0 and data-driven selectors for the parameters h and k (in other words, it justifies adaptive estimation based on cross-validation ideas in the FSIM (2.1)). To the best of our knowledge, this is the first result in the literature on kernel or kNN adaptive estimation in the FSIM (2.1). Actually, in the case of kNN adaptive estimation, there are no such kind of results even in the multivariate single-index model.*

2.4 Practical issues

In the previous Section 2.3.4, it was given theoretical validation of the estimators of the nonparametric link, $r(\cdot)$, based on both CV-kernel and CV-kNN ideas, $\widehat{r}_{\widehat{h},\widehat{\theta}_{\widehat{h}}}(\chi)$ and $\widehat{r}^*_{\widehat{k},\widehat{\theta}^*_{\widehat{k}}}(\chi)$, respectively. Therefore, in practice the only additional issues that must be addressed are how to construct Θ_n, a_n, b_n, $k_{1,n}$ and $k_{2,n}$. That is the aim of this section.

The set of functional directions, Θ_n. We propose to construct Θ_n in a similar way as in Ait-Saïdi et al. [3]. Specifically:

(i) Each direction $\theta \in \Theta_n$ is obtained from a d_n-dimensional space generated by B-spline basis functions, $\{e_1(\cdot), \ldots, e_{d_n}(\cdot)\}$. Therefore, we focus on directions

$$\theta(\cdot) = \sum_{j=1}^{d_n} \alpha_j e_j(\cdot) \text{ where } (\alpha_1, \ldots, \alpha_{d_n}) \in \mathcal{V}. \tag{2.36}$$

(ii) The set of vectors of coefficients in (2.36), $\mathcal{V}$, is obtained by means of the following procedure:

Step 1 For each $(\beta_1, \ldots, \beta_{d_n}) \in \mathcal{C}^{d_n}$, where $\mathcal{C} = \{c_1, \ldots, c_J\} \subset \mathbb{R}^J$ denotes a set of J "seed-coefficients", construct the initial functional direction

$$\theta_{init}(\cdot) = \sum_{j=1}^{d_n} \beta_j e_j(\cdot).$$

Step 2 For each θ_{init} in Step 1 that verifies the condition $\theta_{init}(t_0) > 0$, where t_0 denotes a fixed value in the domain of $\theta_{init}(\cdot)$, compute $\langle \theta_{init}, \theta_{init} \rangle$ and construct

$$(\alpha_1, \ldots, \alpha_{d_n}) = \frac{(\beta_1, \ldots, \beta_{d_n})}{\langle \theta_{init}, \theta_{init} \rangle^{1/2}}.$$

Step 3 Construct $\mathcal{V}$ as the set of vectors $(\alpha_1, \ldots, \alpha_{d_n})$ obtained in Step 2.

Therefore, the final set of eligible functional directions is

$$\Theta_n = \left\{ \theta(\cdot) = \sum_{j=1}^{d_n} \alpha_j e_j(\cdot);\ (\alpha_1, \ldots, \alpha_{d_n}) \in \mathcal{V} \right\}.$$

Remark 2.10 *As usual, in item (i) above we consider splines of order $l \geq 1$ (degree $l-1$) and m_n regularly spaced interior knots (so, $d_n = l + m_n$); note that, from the Jackson type theorem in de Boor [31] (page 149), if θ_0 is sufficiently smooth, then it will be well approximated by some function in the d_n-dimensional space generated by B-spline basis. In addition, by construction (see Step 2), each $\theta \in \Theta_n$ verifies the constraints $\langle \theta, \theta \rangle = 1$ and $\theta(t_0) > 0$; so the identifiability of the FSIM (2.41) is guaranteed (for details, see Proposition*

1 in Ait-Saïdi et al. [3]). Of course, the larger J in Step 1 is, the higher the size of Θ_n is (in fact, the number of initial functional directions in Step 1 is J^{d_n}). At this moment, it should be noted that our approach requires intensive computation due to the optimization on both θ and h or k. Therefore, we need to seek for a trade-off between the size of Θ_n and the performance of the estimators. In that way, Ait-Saïdi et al. [3] suggested to consider $l = 3$ and $\mathcal{C} = \{-1, 0, 1\}$.

***The set of values for** h:* $[a_n, b_n]$. In practice, for selecting some parameter, for instance h, via the minimization of some criterion function (e.g. the CV function), it is usual to minimize over a "wide" set, so that any reasonable set of values for h (for instance the set $[a_n, b_n]$ verifying the technical conditions assumed in the theoretical study) should be included in such wide set. The question of automatic selection of the interval $[a_n, b_n]$ is still unsolved in one-dimensional nonparametric statistics, and becomes of minor importance because the criterion function is usually quite flat around its minimum. Earlier references in one-dimensional setting go back to Härdle and Marron [59] and Marron [82], and the usual recommendation is to choose an interval such that the corresponding bandwidths allow to use up to 95% of the sample. As we will see later along Section 2.5, this recommendation will remain efficient in the functional framework.

***The set of values for** k:* $\{k_{1,n}, k_{1,n} + 1, \ldots, k_{2,n}\}$. The reasoning pointed just before for global kernel estimates is still valid for estimates using local bandwidths (see Vieu [107] for earlier advances); therefore, the same recommendation can be made for the choice of the set $\{k_{1,n}, k_{1,n} + 1, \ldots, k_{2,n}\}$.

2.5 Simulation study

The aim of this section is twofold. On the one hand, to support the suggestions given in sections 2.3.4 and 2.4 related to practical issues inherent to our procedures: selection of the bandwidth (h) and the number of neighbours (k), as well as of the intervals $[a_n, b_n]$ and $[k_{1,n}, k_{2,n}]$. On the other hand, to show the better performance

of the kNN-based estimators versus the kernel-based estimators when heterogeneous designs are considered.

2.5.1 The design

For different values of n, observations i.i.d. $\{(\mathcal{X}_i, Y_i)\}_{i=1}^{n+25}$ were generated from the FSIM

$$Y = r\left(\langle \theta_0, \mathcal{X} \rangle\right) + \varepsilon, \tag{2.37}$$

where:

- The functional covariate was generated from the expression

$$\mathcal{X}(t) = a\cos(2\pi t) + b\sin(4\pi t) + 2c(t-0.25)(t-0.5) \quad (t \in [0,1]). \tag{2.38}$$

  The same mixture distribution was considered for the random variables a, b and c in (2.38): $U(5,10)$ with probability 0.5, and $U(20,20.5)$ with probability 0.5, while each curve $\mathcal{X}_i$ was discretized in 100 equispaced points ($0 = t_1 < t_2 < \cdots < t_{100} = 1$).

- We considered as link function $r(u) = u^3$ and as inner product $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$.

- θ_0 was selected at random in Θ_n (more details will be given at the end of this section).

- Finally, ε is a centred Gaussian random variable with variance equal to 0.025 times the empirical variance of $r\left(\langle \theta_0, \mathcal{X} \rangle\right)$ (i.e. signal-to-noise= 2.5%).

Figure 2.1 shows a sample of 50 curves (left panel) and the corresponding scatter plot of $\{(\langle \theta_0, \mathcal{X}_i \rangle, Y_i)\}_{i=1}^{50}$ (right panel). Clearly one can see two subsamples of curves, being the variability in one of them much greater than in the other. This fact gives rise to two clusters in the sample of projections, $\{\langle \theta_0, \mathcal{X}_i \rangle\}_{i=1}^{50}$; so, taking into account their location-adaptive property, one expects that the kNN-based estimators take advantage on the kernel-based ones.

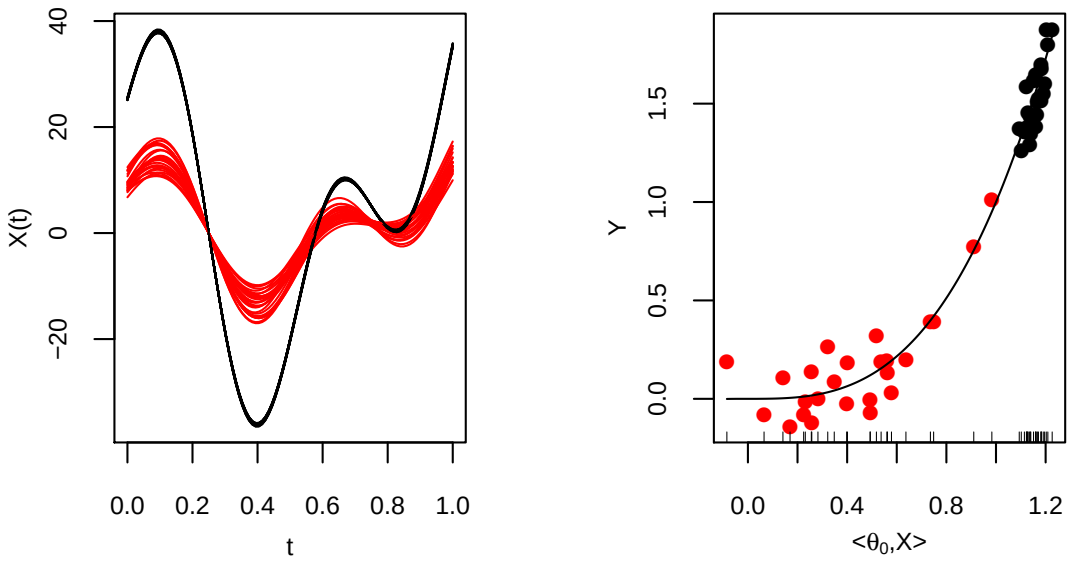
The sample $\mathcal{D}_n = \{(\mathcal{X}_i, Y_i)\}_{i=1}^{n+25}$ was split into two samples: a training sample, $\mathcal{D}_{n,train} = \{(\mathcal{X}_i, Y_i)\}_{i=1}^{n}$, and a testing sample, $\mathcal{D}_{n,test} = \{(\mathcal{X}_i, Y_i)\}_{i=n+1}^{n+25}$. The tuning

parameters ($\widehat{h}$ and $\widehat{k}$) and the estimates of θ_0 ($\widehat{\theta}$ and $\widehat{\theta}^*$) were constructed from the training sample by means of the cross-validation procedure proposed in Section 2.3.4. The sets of functional directions (Θ_n), values for h ($[a_n, b_n]$) and values for k ($\{k_{1,n}, k_{1,n}+1, \ldots, k_{2,n}\}$) were constructed as recommended in Section 2.4. The value for t_0 related to Θ_n (see Step 2 in Section 2.4) was fixed to $t_0 = 0.5$, while the considered order of the basis functions and number of interior knots were $l = 3$ and $m_n = 3$, respectively. As mentioned above, θ_0 was selected at random in Θ_n; once the values of t_0, l and m_n were established, we can indicate what are the coefficients of θ_0 in expression (2.36):

$$(1.201061, 1.201061, 1.201061, 1.201061, 0, 0) \tag{2.39}$$

(see Step 3 in Section 2.4).

Figure 2.1: Sample of 50 curves $\mathcal{X}$ (left panel) together with the corresponding scatter plot of $\{(\langle \theta_0, \mathcal{X} \rangle, Y)\}$ (right panel).



Then, the testing sample was used to measure the quality of the corresponding predictions (i.e., the performance of our procedures) through the *mean squared error of prediction* (MSEP):

$$\mathrm{MSEP}_n = \frac{1}{n_{test}} \sum_{i=n+1}^{n+n_{test}} \left(Y_i - \widehat{Y}_i\right)^2, \tag{2.40}$$

where, in this case $n_{test} = card(\mathcal{I}_{n,test}) = 25$ (since $\mathcal{I}_{n,test} = \{n+1, \ldots, n+25\}$) and $\widehat{Y}_i$ denotes a predicted value for Y_i.

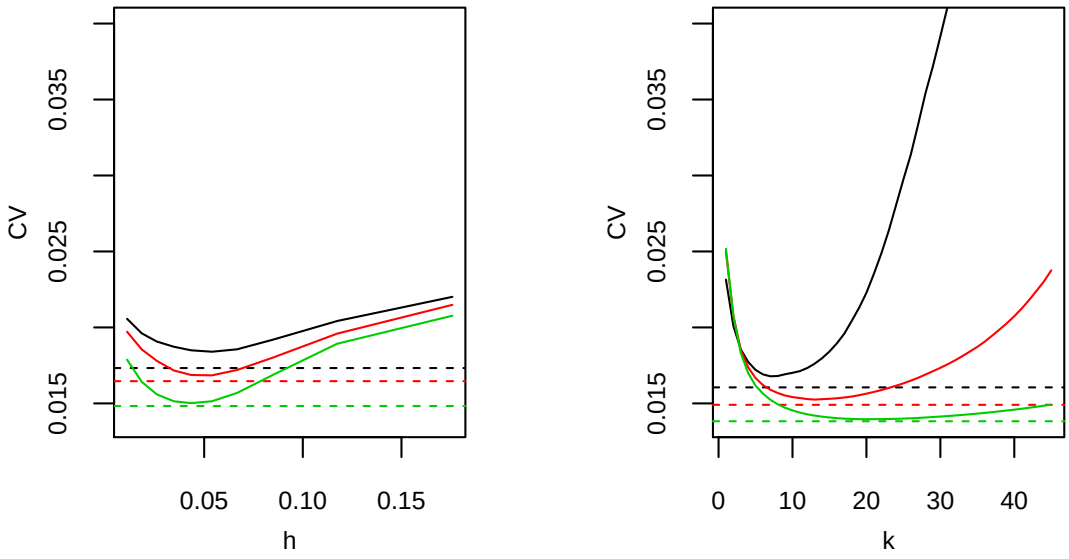
2.5.2 Results

For each sample size considered ($n = 50, 100, 200$), $M = 100$ replicates were generated. In order to support the suggestions given in Section 2.4 to construct $a_n, b_n, k_{1,n}$ and $k_{2,n}$, Figure 2.2 displays the average of the cross-validation functions obtained from both the kernel-based estimator (left panel) and the kNN-based estimator (right panel) when different values for the bandwidth h and the number of neighbours k are considered, respectively.

An interesting practical consequence of what is shown in Figure 2.2 is that the optimal value for h or k will not change as long as reasonable intervals are chosen.

Figure 2.3 shows the average of the MSEP functions obtained from both the kernel-based estimator (left panel) and kNN-based estimator (right panel) when different values for the bandwidth h and the number of neighbours k are considered, respectively. The corresponding values when h and k are obtained from the cross-validation method are reported in Table 2.1.

Figure 2.2: Average of the cross-validation functions obtained from both the kernel-based estimators (left panel) and the kNN-based ones as function of the bandwidth (h) and the number of neighbours (k), respectively. The dashed lines show the average of the cross-validation functions when optimal values for h (left panel) and k (right panel) are considered. From top to bottom, the pairs (solid curve, dashed line) correspond to $n = 50, 100, 200$.



The main conclusions from Figure 2.3 and Table 2.1 are that, for each considered sample size:

(i) The estimators are very sensitive to the values of their tuning parameters.

(ii) The recommendation given in Section 2.4 to construct $a_n, b_n, k_{1,n}$ and $k_{2,n}$ is appropriate (in the sense indicated in such section).

(iii) The cross-validation selectors are competitive ones.

(iv) The performance of the kNN-based estimator is better than the provided by the kernel-based one.

Figure 2.3: Average of the MSEP functions obtained from both the kernel-based estimators (left panel) and the kNN-based ones as function of the bandwidth (h) and the number of neighbours (k), respectively. The dashed lines show the average of the MSEP functions when values for h (left panel) and k (right panel) obtained from the cross-validation method are considered. From top to bottom, the pairs (solid curve, dashed line) correspond to $n = 50, 100, 200$.

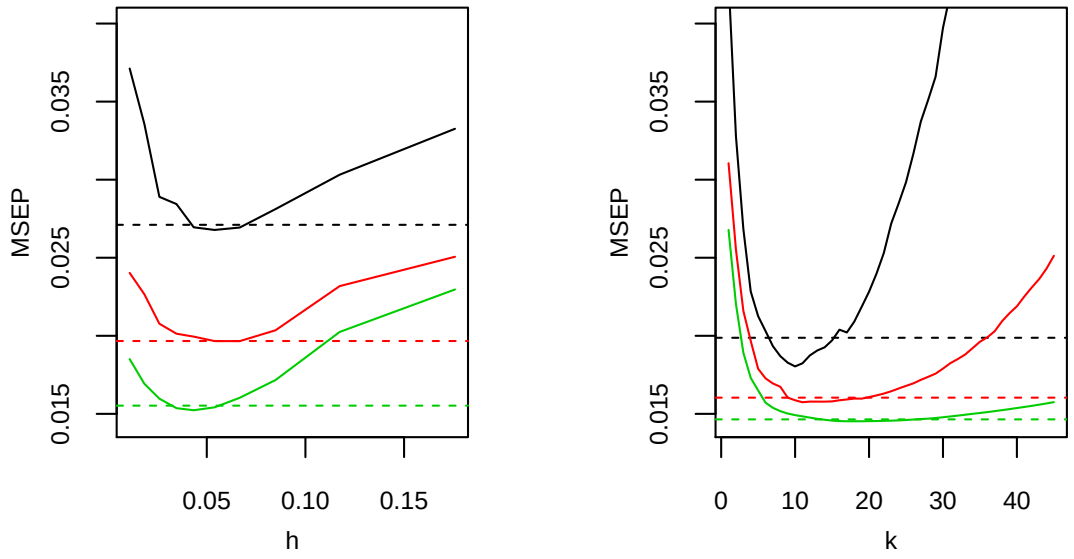


Table 2.1: Average of the MSEPs obtained when the CV selectors for h and k are used.

$n = 50$		$n = 100$		$n = 200$	
kernel	kNN	kernel	kNN	kernel	kNN
0.0271	0.0199	0.0197	0.0160	0.0155	0.0146

2.6 Application to real data

This section is devoted to illustrate, using a real dataset, the usefulness of the FSIM (2.1), as well as to compare the performance of the proposed adaptive kernel- and kNN-based estimators, $\widehat{r}_{\widehat{h},\widehat{\theta}_{\widehat{h}}}(\cdot)$ and $\widehat{r}^{*}_{\widehat{k},\widehat{\theta}^{*}_{\widehat{k}}}(\cdot)$, respectively, when the sample size increases (for details on those estimators, see Section 2.3.4).

2.6.1 The data

We will analyse the previously mentioned Tecator's data (see Section 1.1), a benchmark dataset in the setting of nonparametric functional modelling (see, for instance, Burba et al. [22], Chen et al. [27] and Aneiros and Vieu [6] for functional nonparametric pure, multiple index and sparse additive nonparametric regressions, respectively). Specifically, Tecator's data contain measurements of the fat contents $(Y_i,\ i = 1,\ldots,215)$ and near-infrared absorbance spectra $(\mathcal{X}_i,\ i = 1,\ldots,215)$ observed on 100 equally wavelengths in the range $850 - 1050$ nm (see Figure 1.2 for representation of samples of the absorbance curves and their second derivatives) of 215 finely chopped pieces of meat.

As usual when one deals with Tecator's dataset, the second derivatives of the absorbance curves $(\mathcal{X}^{(2)})$ will play the role of functional covariate. So, we focus on the FSIM

$$Y = r\left(\left\langle \theta_0, \mathcal{X}^{(2)} \right\rangle\right) + \varepsilon. \tag{2.41}$$

We are interested in the performance of our procedures for different sample sizes n. Then, for each $n = 50, 100, 160$, we will consider subsamples

$$\mathcal{D}_n = \left\{ (\mathcal{X}_i^{(2)}, Y_i),\ i \in \mathcal{I}_n \right\},\ \text{where we have denoted}\ \ \mathcal{I}_n = \{1, 2, \ldots, n + 55\}.$$

Each subsample $\mathcal{D}_n$ was split at random into two samples: a training sample,

$$\mathcal{D}_{n,train} = \{(\mathcal{X}_i^{(2)}, Y_i),\ i \in \mathcal{I}_{n,train}\},$$

and a testing sample,

$$\mathcal{D}_{n,test} = \{(\mathcal{X}_i^{(2)}, Y_i),\ i \in \mathcal{I}_{n,test}\},$$

where $n_{train} = \text{card}(\mathcal{I}_{n,train}) = n$, $n_{test} = \text{card}(\mathcal{I}_{n,test}) = 55$, $\mathcal{I}_{n,train} \cup \mathcal{I}_{n,test} = \mathcal{I}_n$ and $\mathcal{I}_{n,train} \cap \mathcal{I}_{n,test} = \emptyset$.

In the estimation procedures, the parameters h, k, a_n , b_n, $k_{1,n}$ and $k_{2,n}$ were constructed from the training sample in the same way as in the simulation study (see Section 2.4 or Section 2.5.1). Several sets of functional directions (Θ_n), depending on the tuning parameter m_n (number of interior knots), also were constructed as recommended in Section 2.4. Values considered for m_n were $2, 3, 4, 5, 6$ (note that the corresponding cardinals of Θ_n were 108, 243, 1053, 2187 and 9477, respectively), and the used value was selected by means of cross-validation ideas. The value for t_0 related to Θ_n (see Step 2 in Section 2.4) was fixed to $t_0 = (850 + 1050)/2)$. The Epanechnikov kernel was used in the nonparametric estimates $\widehat{r}(\cdot)$ and $\widehat{r}^*(\cdot)$.

The testing sample was used to measure the quality of the corresponding predictions through the MSEP (see (2.40)).

2.6.2 Results

2.6.2.1 Performance of the procedures for different sample sizes n

In order to show the performance of the proposed procedures when the sample size increases, twenty partitions $(\mathcal{D}^{(j)}_{n,train}, \mathcal{D}^{(j)}_{n,test})$ of $\mathcal{D}_n$ were generated at random ($n = 50, 100, 160$; $j = 1, \ldots, 20$). Then, the corresponding prediction errors, $\text{MSEP}^{(j)}_n$, were computed. Table 2.2 reports the average of such MSEPs.

Table 2.2: Average of the MSEPs obtained when the CV selectors for h, k and $nknots$ are used.

$n = 50$		$n = 100$		$n = 160$	
kernel	kNN	kernel	kNN	kernel	kNN
11.66	10.97	5.80	5.72	4.66	3.88

A main suggestion from Table 2.2 is that, for each considered sample size, the performance of the kNN-based estimator is slightly better than the corresponding to the kernel-based one. In addition, the performance of each estimator improves as the sample size increases.

2.6.2.2 Benchmark partition: Adaptive estimation in action

From now on, we focus on $\mathcal{D}_{160}$ (i.e., all the Tecator's dataset) and the partition given by $\mathcal{I}_{160,train} = \{1, 2, \ldots, 160\}$ and $\mathcal{I}_{160,test} = \{161, 162, \ldots, 215\}$. Note that this partition can be considered as a benchmark one in the sense that it is the usually considered in papers analysing the Tecator's dataset (see, for instance, Aneiros and Vieu [6], Burba et al. [22] and Ferraty et al. [50], among others).

In a first attempt, we focus on the proposed kernel- and kNN-based estimates $\widehat{r}(\cdot)$ and $\widehat{r}^*(\cdot)$, respectively. In both cases, the same value for m_n ($\widehat{m_n}_{CV} = 4$) was selected, while the optimal bandwidth and number of neighbours where $\widehat{h}_{CV} = 15.80106$ and $\widehat{k}_{CV} = 9$, respectively. In addition, the same estimate for θ_0 ($\widehat{\theta} = \widehat{\theta}_{\widehat{h}_{CV}} = \widehat{\theta}^*_{\widehat{k}_{CV}}$) was obtained.

Figure 2.4: Left panel: Estimate of the functional direction θ_0. Right panel: estimates of the regression $r(\cdot)$ by means of the kNN-based (solid line) and kernel-based (dashed line) estimates.

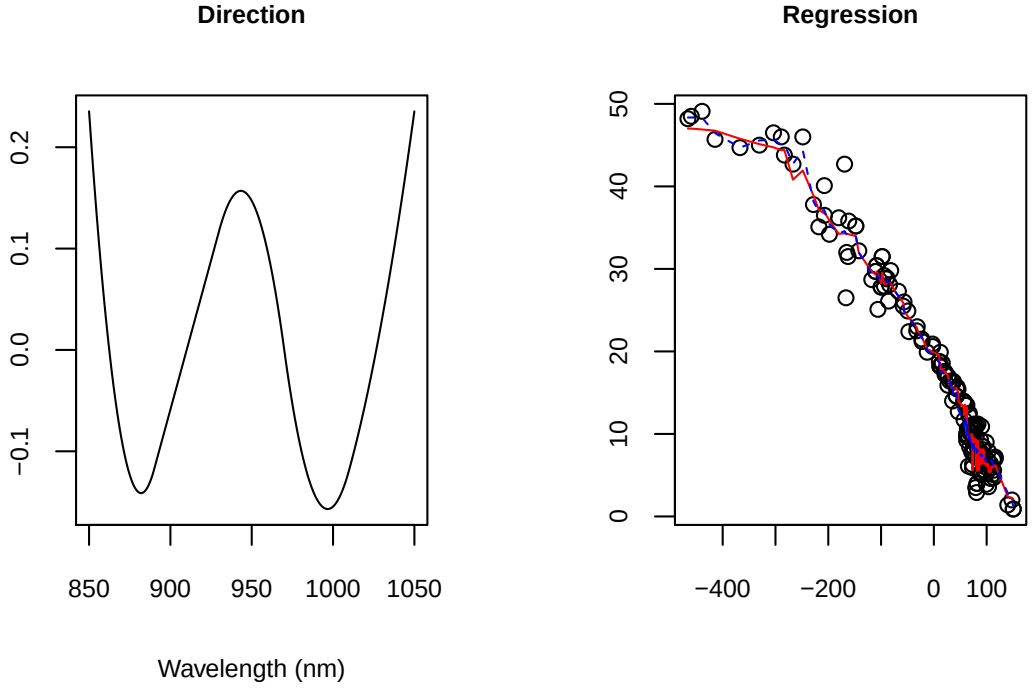


Figure 2.4 displays both the estimate for θ_0 and the estimates of the regression $r(\cdot)$ (i.e., $\widehat{r}_{\widehat{h}_{CV}}(\cdot)$ and $\widehat{r}^*_{\widehat{k}_{CV}}(\cdot)$). On the one hand, the graphic of $\widehat{\theta}$ suggests that the two bumps around wavelengths 880 and 1000, as well as the peak around wavelength 940, could be important indicators of the fat content (note that this suggestion is compatible with the findings in Aneiros and Vieu [6]). We would like to emphasize that one of the advantages of the FSIM against functional models dealing with whole curves instead of projected ones is the possibility of interpretation; as noted in the previous sentence, nice and easy interpretation is obtained in our application. On the other hand, the two estimates of the regression suggest nonlinear relationship between the fat content and the absorbance spectra (in fact, the p-value of the Ramsey's RESET test for linearity is 0.000; for details, see Ramsey [98]). Finally, it is worth highlighting the different behaviour of the considered estimates $\widehat{r}_{\widehat{h}_{CV}}(\cdot)$ and $\widehat{r}^*_{\widehat{k}_{CV}}(\cdot)$: in general, the kernel-based estimate is smoother than the kNN-based one. This fact is a consequence of two reasons: (i) the heterogeneity in the values of the covariates $\left\langle \widehat{\theta}, \mathcal{X}_i^{(2)} \right\rangle$, and (ii) the bandwidth ($\widehat{h}_{CV}$) used in $\widehat{r}_{\widehat{h}_{CV}}(\cdot)$ is global (it does not depend on χ) while the one used in $\widehat{r}^*_{\widehat{k}_{CV}}(\cdot)$ ($H_{\widehat{k}_{CV},\chi,\widehat{\theta}}$) is local (it depends on χ). Actually, the local-adaptive bandwidth is a main appealing feature of kNN estimators in different settings (not only in the FSIM); in fact, as it will be shown in the rest of this section, such feature plays a major role in achieving accurate predictions.

Table 2.3: Values of the MSEPs from some functional models.

	Model	Method	MSEP
FLM	$Y = \gamma_0 + \int_{850}^{1050} \mathcal{X}^{(2)}(t)\gamma(t)dt + \varepsilon$	FPC	7.17
FNM	$Y = m(\mathcal{X}^{(2)}) + \varepsilon$	kernel	4.06
		kNN	1.79
FSIM	$Y = r\left(\left\langle \theta_0, \mathcal{X}^{(2)} \right\rangle\right) + \varepsilon$	kernel	3.49
		kNN	2.69

Table 2.3 reports the values of the MSEPs obtained from the FSIM (2.41) when

it is estimated using both the kernel- and kNN-based adaptive estimators $\widehat{r}_{\widehat{h}_{CV}}(\cdot)$ and $\widehat{r}^{*}_{\widehat{k}_{CV}}(\cdot)$, respectively. The corresponding values obtained from the FNM (1.2) (using kernel-and kNN-based estimators) and the FLM (1.1) (estimating by means of functional principal components regression (FPC); see e.g. Aguilera et al. [1] for partial least squares regression, including an application to Tecator's data, and Febrero-Bande et al. [43] for a comparative study between these two dimensionality reduction techniques) are also included in the table.

In our real data application, two main conclusions can be drawn from Table 2.3: (i) the relationship between the fat content and the absorbance curve is nonlinear, and (ii) the FSIM estimated by means of the proposed kNN estimator achieves better predictive power than when it is estimated through the proposed kernel one. Nevertheless, the smallest value of the MSEP is obtained when the kNN estimator is applied to the FNM.

In a second attempt, we implement a full nonparametric boosting step in the estimated FSIM. Specifically, we consider the following FNM to regress the residuals ($\widehat{\varepsilon}_i$) from the FSIM on the first derivative ($\mathcal{X}_i^{(1)}$) of the absorbance curves (the order of the derivative was selected using cross-validation ideas):

$$\widehat{\varepsilon}_i = m\left(\mathcal{X}_i^{(1)}\right) + e_i, \tag{2.42}$$

where e_i denotes the corresponding random error. Then, if $\widehat{m}(\cdot)$ denotes the nonparametric estimator of $m(\cdot)$ in (2.42), a new prediction for Y_j in the test sample can be constructed as

$$\widehat{Y}_j = \widehat{r}\left(\left\langle \widehat{\theta}, \mathcal{X}_j^{(2)} \right\rangle\right) + \widehat{m}\left(\mathcal{X}_j^{(1)}\right) \ \ (j = 161, \ldots, 215).$$

Table 2.4 reports the values of the MSEP corresponding to such predictions when both functions $r(\cdot)$ and $m(\cdot)$ are estimated by means of either kernel-based or kNN-based estimators. Several conclusions can be drawn from Table 2.4. On the one hand, it shows (again) the convenience of using kNN estimates instead of kernel ones. On the other hand, it supports the idea of considering a boosting procedure to take, from the whole curve, information not captured by the functional index.

Table 2.4: Values of the MSEP when a full nonparametric boosting is applied on the residuals of the FSIM.

	Model	Method	MSEP
FSIM & FNM	$Y = r\left(\left\langle \theta_0, \mathcal{X}^{(2)} \right\rangle\right) + m\left(\mathcal{X}^{(1)}\right) + \varepsilon$	kernel	1.74
		kNN	1.53

2.6.3 Conclusions

This real data analysis illustrates both the interest of the semiparametric approach and the efficiency of the kNN estimation procedure. On the one hand, because of its location-adaptive feature, the kNN approach exceeds the performances of usual global smoothers, such as kernel ones, while the cross-validation procedure makes this estimate of fully automatic use. On the other hand, the semiparametric feature of the FSIM approach has the double advantage of combining interpretability of the outputs (see Figure 2.4) together with low prediction errors (see Tables 2.3 and 2.4).

2.7 Appendix Chapter 2: Proofs

From now on, C denotes a generic positive constant which may take different values from one formula to another.

Before presenting the proofs of our Proposition 2.2, Theorem 2.5, Corollary 2.6 and Corollary 2.8, we first enunciate some known auxiliary results that play a main role in our proofs. In such results, $Z_1, Z_2, \ldots, Z_n$ are i.i.d. variables taking values in a measurable space $(\mathcal{Z}, \mathcal{A})$ and $\mathcal{K}$ is a pointwise measurable class of functions $\{g : \mathcal{Z} \longrightarrow \mathbb{R}\}$ with envelope function F. In addition, we denote

$$\alpha_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(g(Z_i) - \mathbb{E}(g(Z_i))\right), \ ||\alpha_n(g)||_{\mathcal{K}} = \sup_{g \in \mathcal{K}} |\alpha_n(g)|, \ ||\cdot||_p = \sqrt[p]{\mathbb{E}(\cdot)^p}$$

and

$$J(1,\mathcal{K}) = \sup_{\mathcal{Q}} \int_0^1 \sqrt{1 + \log \mathcal{N}\left(\epsilon ||F||_{\mathcal{Q},2}, \mathcal{K}, d_{\mathcal{Q},2}\right)} d\epsilon,$$

where the supremum is taken over all the probability measures $\mathcal{Q}$ on the measure space $(\mathcal{Z}, \mathcal{A})$ with $||F||_{\mathcal{Q},2} < \infty$ (for additional notation, see Section 2.3).

2.7.1 Some auxiliary results

Lemma 2.11 (Theorem 2.14.1 in van der Vaart and Wellner [106], p. 239) *We have that:*

$$\left|\left| \, ||\alpha_n(g)||_{\mathcal{K}} \right|\right|_p \leq C J(1,\mathcal{K}) \, ||F||_{p \vee 2} \, ,$$

where $s \vee t$ is the spermium of s and t.

Lemma 2.12 (Theorem 3.1 in Dony and Einmahl [34], p. 314) *If the class $\mathcal{K}$ is such that $\mathbb{E} \, ||\alpha_n(g)||_{\mathcal{K}} \leq C \, ||F||_2$, then, for any $A \in \mathcal{A}$, we have:*

$$\mathbb{E} \, ||\alpha_n(g 1_A)||_{\mathcal{K}} \leq 2C \, ||F 1_A||_2 \, .$$

Lemma 2.13 (Bernstein type inequality in Dony and Einmahl [34], p. 321) *Assume that the variables $Z_1, Z_2, \ldots, Z_n$ satisfy for some $H > 0$,*

$$\mathbb{E}\left(F^p(Z)\right) \leq \frac{p!}{2} \sigma^2 H^{p-2},$$

where $\sigma^2 \geq \mathbb{E}(F^2(Z))$. Then, by denoting $\beta_n = \mathbb{E}(||\sqrt{n}\alpha_n(g)||_{\mathcal{K}})$ we have for any $t > 0$:

$$\mathbb{P}\left\{ \max_{1 \leq k \leq n} \left|\left| \sqrt{k}\alpha_k(g) \right|\right|_{\mathcal{K}} \geq \beta_n + t \right\} \leq \exp\left(- \frac{t^2}{2n\sigma^2 + 2tH} \right).$$

Lemma 2.14 (Lema 6.1 in Kara-Zaitri et al. [66], p. 186) *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with $\mathbb{P}(X_i) = p$ for all $i = 1, \ldots, n$. Set $U = X_1 + \cdots + X_n$ and $\mu = pn$. Then, for any $w > 0$, we have:*

$$\mathbb{P}\left(U \geq (1+w)\mu\right) \leq \exp\{-\mu \min\{w, w^2\}/4\},$$

and if $w \in (0,1)$*, we have*

$$\mathbb{P}\left(U \leq (1-w)\mu\right) \leq \exp\{-\mu w^2/2\}.$$

2.7.2 Proof of Proposition 2.2

Results in Proposition 2.2(a) and Proposition 2.2(b) are direct consequence of Theorems 3.1 in Kara-Zaitri et al. [66, 67], respectively. On the one hand, one should note that, when θ_0 is known, $\widehat{r}_{h,\theta_0}(\cdot)$ and $\widehat{r}^*_{k,\theta_0}(\cdot)$ are kernel- and kNN-type estimators, respectively, based on the semi-metric $d_{\theta_0}(\cdot,\cdot)$, of the nonparametric regression operator, $r_{\theta_0}(\cdot)$, between the scalar variable Y and the functional covariate $\mathcal{X}$. On the one hand, in the case of Proposition 2.2(b), we must take into account the correction relative to the rate of convergence in Theorem 3.1 in Kara-Zaitri et al. [66] indicated in the Remark 2.1.

Actually, our assumptions (2.15) and (2.22) are slightly different (weaker) of assumptions (6) and (10) in Kara-Zaitri et al. [66] and assumptions H1 and H3 in Kara-Zaitri et al. [67]; to show that their Theorems 3.1 hold using our assumptions instead of the corresponding ones in Kara-Zaitri et al. [66, 67], it is sufficient to prove Corollary 3.3 in Kara-Zaitri et al. [67] following the proof of our Corollary 2.16 (see below). ■

2.7.3 Proof of Theorem 2.5 (a)

We will follow the scheme used in Kara-Zaitri et al. [67], who focused on the UIB consistency of the kernel estimator of the nonparametric regression (see (1.3)) in the FNM (1.2). Although our theorem differs with respect to that of Kara-Zaitri et al. [67] in both the model and the type of consistency to prove (we focus on the FSIM (2.1) instead of the FNM (1.2) and our aim is the UIBD consistency instead of the UIB one), their scheme of proof can be followed once the assumptions are adapted in a suitable way.

Taking into account that

$$\widehat{r}_{h,\theta}(\chi) = \frac{\widehat{g}_{h,\theta}(\chi)}{\widehat{F}_{h,\theta}(\chi)},$$

where we have denoted

$$\widehat{g}_{h,\theta}(\chi) = \frac{1}{n\phi_{\chi,\theta}(h)} \sum_{i=1}^{n} K\left(h^{-1} d_\theta\left(\chi, \mathcal{X}_i\right)\right) Y_i$$

and

$$\widehat{F}_{h,\theta}(\chi) = \frac{1}{n\phi_{\chi,\theta}(h)} \sum_{i=1}^{n} K\left(h^{-1} d_\theta\left(\chi, \mathcal{X}_i\right)\right),$$

we can write

$$\widehat{r}_{h,\theta}(\chi) - r_{\theta_0}(\chi) = \widehat{B}_{h,\theta}(\chi) + \frac{\widehat{R}_{h,\theta}(\chi)}{\widehat{F}_{h,\theta}(\chi)} + \frac{\widehat{Q}_{h,\theta}(\chi)}{\widehat{F}_{h,\theta}(\chi)},$$

where

$$\widehat{B}_{h,\theta}(\chi) = \frac{\mathbb{E}\left(\widehat{g}_{h,\theta}(\chi)\right)}{\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)} - r_{\theta_0}(\chi), \;\; \widehat{R}_{h,\theta}(\chi) = -\widehat{B}_{h,\theta}(\chi)\left(\widehat{F}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right)$$

and

$$\widehat{Q}_{h,\theta}(\chi) = \left(\widehat{g}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{g}_{h,\theta}(\chi)\right)\right) - r_{\theta_0}(\chi)\left(\widehat{F}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right).$$

Thus, the proof of our Theorem 2.5(a) is completed once we prove the following four results.

Lemma 2.15 *Under assumptions (2.22), (2.25), (2.28), (2.30) and (2.35), we have that:*

$$\sup_{\theta\in\Theta_n} \sup_{a_n\le h\le b_n} \left|\widehat{F}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right| = O_{a.co.}\left(\sqrt{\frac{\log n}{n f(a_n)}}\right).$$

Corollary 2.16 *Under assumptions of Lemma 2.15 together with Assumption (2.29), there exists $C > 0$ such that*

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\inf_{\theta\in\Theta_n} \inf_{a_n\le h\le b_n} \widehat{F}_{h,\theta}(\chi) < C\right) < \infty.$$

Lemma 2.17 *Under assumptions and (2.20), (2.22) and (2.26)–(2.29), we have that:*

$$\sup_{\theta\in\Theta_n} \sup_{a_n\le h\le b_n} \left|\widehat{B}_{h,\theta}(\chi)\right| = O\left(b_n^{\beta_0}\right).$$

Lemma 2.18 *Under assumptions (2.21), (2.22), (2.25), (2.28), (2.30) and (2.35), we have that:*

$$\sup_{\theta\in\Theta_n}\sup_{a_n\leq h\leq b_n}|\widehat{g}_{h,\theta}(\chi)-\mathbb{E}\left(\widehat{g}_{h,\theta}(\chi)\right)|=O_{a.co.}\left(\sqrt{\frac{\log n}{nf(a_n)}}\right).$$

2.7.3.1 Proof of Lemma 2.15

Following the definition of rate of almost-complete convergence (see 2.24), we need to prove that $\exists\eta_0>0$ and $b_0>0$ such that:

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{a_n\leq h\leq b_0}\sqrt{\frac{nf(a_n)}{\log n}}\left|\widehat{F}_{h,\theta}(\chi)-\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right|\geq\eta_0\right)<\infty.$$

Taking Assumption (2.28) into account, it suffices to prove that there exist $\eta_0>0$ and $b_0>0$ such that

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{a_n\leq h\leq b_0}\sqrt{\frac{n\phi_{\chi,\theta}(a_n)}{\log n}}\left|\widehat{F}_{h,\theta}(\chi)-\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right|\geq\eta_0\right)<\infty. \tag{2.43}$$

For bounding each addend in expression (2.43), Bernstein type inequality formulated in Lemma 2.13 will be used. For that, we will make some previous calculations.

First of all, if we define

$$h_j=2^j a_n,\quad L(n)=\max\{j:h_j\leq 2b_0\},$$

it allows us to write

$$\sup_{\theta\in\Theta_n}\sup_{a_n\leq h\leq b_0}\sqrt{\frac{n\phi_{\chi,\theta}(a_n)}{\log n}}\left|\widehat{F}_{h,\theta}(\chi)-\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right|\leq$$
$$\sup_{\theta\in\Theta_n}\max_{j=1,\ldots,L(n)}\sup_{h_{j-1}\leq h\leq h_j}\sqrt{\frac{n\phi_{\chi,\theta}(h)}{\log n}}\left|\widehat{F}_{h,\theta}(\chi)-\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right|. \tag{2.44}$$

In addition, since Assumption (2.25) is verified, $\Theta_n=\{\theta_1,\ldots,\theta_{n^\alpha}\}$. Furthermore,

if we denote

$$\alpha_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(g(\mathcal{X}_i) - \mathbb{E}\left(g(\mathcal{X}_i) \right) \right), \quad K_{h,\theta}(\mathcal{X}_i) = K(h^{-1} d_\theta(\chi, \mathcal{X}_i),$$

we can write

$$\widehat{F}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{F}_{h,\theta}(\chi) \right) = \frac{1}{\sqrt{n}\phi_{\chi,\theta}(h)} \alpha_n(K_{h,\theta}), \tag{2.45}$$

and, for $1 \leq j \leq L(n)$ and $1 \leq m \leq n^\alpha$,

$$\mathcal{G}_{j,m} = \left\{ \cdot \longrightarrow K\left(h^{-1} d_{\theta_m}(\chi, \cdot) \right) \text{ where } h_{j-1} \leq h \leq h_j \right\}. \tag{2.46}$$

Then, using (2.44), (2.45) and (2.46), we can establish the following chain of inequalities:

$$\begin{aligned}
& \mathbb{P}\left(\sup_{\theta\in\Theta_n} \sup_{a_n \leq h \leq b_0} \sqrt{\frac{n\phi_{\chi,\theta}(a_n)}{\log n}} \left| \widehat{F}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{F}_{h,\theta}(\chi) \right) \right| \geq \eta_0 \right) \\
\leq\ & \mathbb{P}\left(\sup_{\theta\in\Theta_n} \max_{j=1,\ldots,L(n)} \sup_{h_j \leq h \leq h_{j-1}} \sqrt{\frac{n\phi_{\chi,\theta}(h)}{\log n}} \left| \widehat{F}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{F}_{h,\theta}(\chi) \right) \right| \geq \eta_0 \right) \\
\leq\ & \mathbb{P}\left(\sup_{\theta\in\Theta_n} \max_{j=1,\ldots,L(n)} \sup_{h_{j-1} \leq h \leq h_j} \frac{1}{\sqrt{n \log n \phi_{\chi,\theta}(h)}} \left| \sqrt{n}\alpha_n(K_{h,\theta}) \right| \geq \eta_0 \right) \\
\leq\ & \mathbb{P}\left(\max_{m=1,\ldots,n^\alpha} \max_{j=1,\ldots,L(n)} \sup_{h_{j-1} \leq h \leq h_j} \frac{1}{\sqrt{n \log n \phi_{\chi,\theta_m}(h)}} \left| \left| \sqrt{n}\alpha_n(g) \right| \right|_{\mathcal{G}_{j,m}} \geq \eta_0 \right) \\
\leq\ & \sum_{m=1}^{n^\alpha} \sum_{j=1}^{L(n)} \mathbb{P}\left(\frac{1}{\sqrt{n \log n \phi_{\chi,\theta_m}(h_j/2)}} \left\| \sqrt{n}\alpha_n(g) \right\|_{\mathcal{G}_{j,m}} \geq \eta_0 \right) \\
\leq\ & n^\alpha L(n) \max_{m=1,\ldots,n^\alpha} \max_{j=1,\ldots,L(n)} \mathbb{P}\left(\max_{1\leq k\leq n} \left\| \sqrt{k}\alpha_k(g) \right\|_{\mathcal{G}_{j,m}} \geq \eta_0 \sqrt{n \log n \phi_{\chi,\theta_m}(h_j/2)} \right).
\end{aligned} \tag{2.47}$$

In order to bound the probability that appears in (2.47) by means of the the Bernstein's inequality formulated in Lemma 2.13, we first study the asymptotic behaviour of

$$\sigma^2 = \mathbb{E}\left(G_{j,m}^2(\mathcal{X}) \right) \quad \text{and} \quad \beta_n = \mathbb{E}\left(\left\| \sqrt{n}\alpha_n(g) \right\|_{\mathcal{G}_{j,m}} \right),$$

where $G_{j,m}$ denotes the minimal envelope function of the class $\mathcal{G}_{j,m}$. It follows from Assumption (2.22) that

$$G_{j,m}(z) \leq C_6 1_{(0,1/2)}\left(\frac{d_{\theta_m}(\chi, z)}{h_j}\right) = C_6 1_{B_{\theta_m}(\chi, h_j/2)}(z).$$

Hence, for all $p \geq 1$, we have that

$$\mathbb{E}\left(G_{j,m}(\mathcal{X})^p\right) \leq C_6^p \mathbb{E}\left(1_{B_{\theta_m}(\chi,h_j/2)}(\mathcal{X})\right) = C_6^p \mathbb{P}\left(d_{\theta_m}(\mathcal{X}, \chi) < h_j/2\right) = C_6^p \phi_{\chi,\theta_m}\left(h_j/2\right).$$

In particular,

$$\sigma^2 = O\left(\phi_{\chi,\theta_m}\left(h_j/2\right)\right)$$

holds. Focusing now on β_n, we obtain, by combining Assumption (2.35) together with Lemma 2.11, that

$$\mathbb{E}\left(\|\alpha_n(g)\|_{\mathcal{G}_{j,m}}\right) \leq \mathbb{E}\left(\|\alpha_n(g)\|_{\mathcal{K}_{\Theta_n}}\right) \leq C J(1, \mathcal{K}_{\Theta_n}) \|F_{\Theta_n}\|_2 \leq C \|F_{\Theta_n}\|_2.$$

Thus, the conditions of Lemma 2.12 are verified for the class $\mathcal{G}_{j,m}$ and the envelope function F_{Θ_n} (note that, in particular, F_{Θ_n} is an envelope function of the class $\mathcal{G}_{j,m}$). So, from such lemma it follows that

$$\mathbb{E}\left(\left\|\alpha_n\left(g1_{B_{\theta_m}(\chi,h_j/2)}\right)\right\|_{\mathcal{G}_{j,m}}\right) \leq C \left\|F1_{B_{\theta_m}(\chi,h_j/2)}\right\|_2.$$

Finally, taking into account (2.22), we obtain that:

$$\beta_n = \mathbb{E}\left(\left\|\sqrt{n}\alpha_n(g)\right\|_{\mathcal{G}_{j,m}}\right) = \mathbb{E}\left(\left\|\sqrt{n}\alpha_n\left(g1_{B_{\theta_m}(\chi,h_j/2)}\right)\right\|_{\mathcal{G}_{j,m}}\right) \leq C\sqrt{n\phi_{\chi,\theta_m}\left(h_j/2\right)}.$$

Now, we can apply the Bernstein's inequality (see Lemma 2.13) with

$$\beta_n = O\left(\sqrt{n\phi_{\chi,\theta_m}\left(h_j/2\right)}\right),\ \sigma^2 = O\left(\phi_{\chi,\theta_m}\left(h_j/2\right)\right) \text{ and } t = \frac{\eta_0}{2}\sqrt{n\log n\phi_{\chi,\theta_m}\left(h_j/2\right)}.$$

In first place,

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq k\leq n}\left\|\sqrt{k}\alpha_k(g)\right\|_{\mathcal{G}_{j,m}}\geq \eta_0\sqrt{n\log n\phi_{\chi,\theta_m}(h_j/2)}\right)\\
&\leq \mathbb{P}\left(\max_{1\leq k\leq n}\left\|\sqrt{k}\alpha_k(g)\right\|_{\mathcal{G}_{j,m}}\geq \beta_n+t\right)
\end{aligned}\tag{2.48}$$

(since, as $n\to\infty$, it is verified that

$$\begin{aligned}
\eta_0\sqrt{n\log n\phi_{\chi,\theta_m}(h_j/2)} &\geq \frac{\eta_0}{2}\sqrt{n\log n\phi_{\chi,\theta_m}(h_j/2)}+O\left(\sqrt{n\phi_{\chi,\theta_m}(h_j/2)}\right),\\
\frac{\eta_0}{2}\sqrt{n\log n\phi_{\chi,\theta_m}(h_j/2)} &\geq C\sqrt{n\phi_{\chi,\theta_m}(h_j/2)},\\
\frac{\eta_0}{2}\sqrt{\log n} &\geq C).
\end{aligned}$$

Then, using the Bernstein's inequality in Lemma 2.13,

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq k\leq n}\left\|\sqrt{k}\alpha_k(g)\right\|_{\mathcal{G}_{j,m}}\geq \beta_n+t\right)\\
&\leq \exp\left\{\frac{-\eta_0^2 n\log n\phi_{\chi,\theta_m}\left(\frac{h_j}{2}\right)}{8nC\phi_{\chi,\theta_m}\left(\frac{h_j}{2}\right)+4\eta_0 H\sqrt{n\log n\phi_{\chi,\theta_m}\left(\frac{h_j}{2}\right)}}\right\}\\
&\leq \exp\left\{-\eta_0^2\frac{\log n}{8C+C'\sqrt{\frac{\log n}{n\phi_{\chi,\theta_m}(h_j/2)}}}\right\}\\
&\leq \exp\left\{-\eta_0^2\frac{\log n}{8C+C'\sqrt{\frac{\log n}{nf(h_j/2)}}}\right\}\\
&\leq n^{-C''\eta_0^2}
\end{aligned}\tag{2.49}$$

(note that the penultimate inequality is consequence of Assumption (2.28) and the last one of Assumption (2.30)). Moreover, from (2.47) and (2.49) together with the fact that $L(n)\leq 2\log n$, we get that

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{a_n\leq h\leq b_0}\sqrt{\frac{n\phi_{\chi,\theta}(a_n)}{\log n}}\left|\widehat{F}_{h,\theta}(\chi)-\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right|\geq\eta_0\right)\leq Cn^{-C''\eta_0^2+\alpha}\log n. \tag{2.50}$$

Finally, by choosing now η_0 such that $C''\eta_0^2-\alpha>1$, (2.43) follows from (2.50). ■

2.7.3.2 Proof of Corollary 2.16

On the one hand, from Assumption (2.22), we obtain, $\forall h\in[a_n,b_n]$ and $\forall\theta\in\Theta_n$

$$\begin{aligned}\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right) &= \frac{1}{\phi_{\chi,\theta}(h)}\mathbb{E}\left(K\left(\frac{d_\theta(\chi,\mathcal{X})}{h}\right)\right)\\ &\geq \frac{C_5}{\phi_{\chi,\theta}(h)}\mathbb{E}\left(1_{(0,1/2)}\left(\frac{d_\theta(\chi,\mathcal{X})}{h}\right)\right)\\ &= C_5\frac{\mathbb{P}\left(d_\theta\left(\chi,\mathcal{X}\right)\leq h/2\right)}{\phi_{\chi,\theta}(h)}\\ &= C_5\frac{\phi_{\chi,\theta}(h/2)}{\phi_{\chi,\theta}(h)}.\end{aligned} \tag{2.51}$$

Using (2.51), and applying assumptions (2.28) and (2.29), we obtain that, for n large enough,

$$\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\geq C_5\frac{\phi_{\chi,\theta}(h/2)}{\phi_{\chi,\theta}(h)}\geq\frac{C_5C_9}{C_{10}}\frac{f(h/2)}{f(h)}\geq\frac{C_5C_9C_{11}}{C_{10}}=C'>0, \tag{2.52}$$

$\forall h\in[a_n,b_n]$ and $\forall\theta\in\Theta_n$.

Thus, denoting $C=C'/2$, it is verified that

$$\mathbb{P}\left(\inf_{\theta\in\Theta_n}\inf_{h\in[a_n,b_n]}\widehat{F}_{h,\theta}(\chi)\leq C\right)\leq\mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{h\in[a_n,b_n]}\left|\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)-\widehat{F}_{h,\theta}(\chi)\right|\geq C\right),$$

and Lemma 2.15 leads to the desired result. ■

2.7.3.3 Proof of Lemma 2.17

We have that

$$\begin{aligned}
\left|\widehat{B}_{h,\theta}(\chi)\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right| &= \left|\mathbb{E}\left(\widehat{g}_{h,\theta}(\chi)\right) - r_{\theta_0}(\chi)\mathbb{E}\left(\widehat{F}_{h,\theta}(\chi)\right)\right| \\
&= \left|\frac{1}{\phi_{\chi,\theta}(h)}\mathbb{E}\left(K\left(\frac{d_\theta(\chi,\mathcal{X})}{h}\right)Y\right)\right. \\
&\quad \left. - \frac{r_{\theta_0}(\chi)}{\phi_{\chi,\theta}(h)}\mathbb{E}\left(K\left(\frac{d_\theta(\chi,\mathcal{X})}{h}\right)\right)\right| \\
&= \left|\frac{1}{\phi_{\chi,\theta}(h)}\mathbb{E}\left[K\left(\frac{d_\theta(\chi,\mathcal{X})}{h}\right)\mathbb{E}\left(Y|\left\langle\mathcal{X},\theta_0\right\rangle\right)\right.\right. \\
&\quad \left.\left. - r_{\theta_0}(\chi)K\left(\frac{d_\theta(\chi,\mathcal{X})}{h}\right)\right]\right| \\
&= \left|\frac{1}{\phi_{\chi,\theta}(h)}\mathbb{E}\left[K\left(\frac{d_\theta(\chi,\mathcal{X})}{h}\right)\left(r_{\theta_0}(\mathcal{X}) - r_{\theta_0}(\chi)\right)\right]\right| \\
&\leq C_6\frac{1}{\phi_{\chi,\theta}(h)}\mathbb{E}\left[1_{B_\theta(\chi,h/2)}(\mathcal{X})d_{\theta_0}(\mathcal{X},\chi)^{\beta_0}\right]. \qquad (2.53)
\end{aligned}$$

(Note that the inequality in (2.53) is a consequence of assumptions (2.20) and (2.22)) In addition, assumptions (2.26) and (2.27), together with Cauchy-Schwarz inequality, allow us to write that, if $d_\theta(\mathcal{X},\chi) < h/2$ holds, then, for all $h \in [a_n, b_n]$,

$$\begin{aligned}
d_{\theta_0}(\mathcal{X},\chi) &= d_{\theta_0}(\mathcal{X},\chi) - d_\theta(\mathcal{X},\chi) + d_\theta(\mathcal{X},\chi) \leq |\langle\mathcal{X}-\chi,\theta_0-\theta\rangle| + d_\theta(\mathcal{X},\chi) \\
&\leq \langle\mathcal{X}-\chi,\mathcal{X}-\chi\rangle^{1/2}\langle\theta_0-\theta,\theta_0-\theta\rangle^{1/2} + d_\theta(\mathcal{X},\chi) \leq 2C_7C_8b_n + h/2 \leq Cb_n. \qquad (2.54)
\end{aligned}$$

Now, from (2.53) and (2.54) together with assumptions (2.28) and (2.29), we obtain that, for all $h \in [a_n, b_n]$ and for all $\theta \in \Theta_n$,

$$|\widehat{B}_{h,\theta}(\chi)\mathbb{E}(\widehat{F}_{h,\theta}(\chi))| \leq C\frac{\phi_{\chi,\theta}(h/2)}{\phi_{\chi,\theta}(h)}b_n^{\beta_0} \leq C\frac{f(h/2)}{f(h)}b_n^{\beta_0} \leq Cb_n^{\beta_0}. \qquad (2.55)$$

Finally, (2.52) and (2.55) complete the proof. ■

2.7.3.4 Proof of Lemma 2.18

This proof follows the same scheme as proof of Lemma 2.15. Taking Assumption (2.28) into account, it suffices to prove that there exist $\eta_0' > 0$ and $b_0 > 0$ such that

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{ \sup_{\theta\in\Theta_n} \sup_{a_n\leq h\leq b_0} \sqrt{\frac{n\phi_{\chi,\theta}(a_n)}{\log n}} \left|\widehat{g}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{g}_{h,\theta}(\chi)\right)\right| \geq \eta_0' \right\} < \infty. \tag{2.56}$$

For carrying out the proof, as in the proof of Lemma 2.15, we will make some calculations in order to apply Bernstein's inequality formulated in Lemma 2.13.

Firstly, let us define $h_j = 2^j a_n$ and $L(n) = \max\{j : h_j \leq 2b_0\}$. In addition, let us denote

$$\alpha_n'(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Y_i g(\mathcal{X}_i) - \mathbb{E}\left(Y_i g(\mathcal{X}_i)\right)\right),$$

$$\mathcal{G}_{j,m}' = \left\{(z,y) \longrightarrow yK\left(h^{-1} d_{\theta_m}(\chi,z)\right) \text{ where } h_{j-1} \leq h \leq h_j\right\}$$

and $G_{j,m}'$ denotes the minimal envelope function of the class $\mathcal{G}_{j,m}'$ ($1 \leq j \leq L(n)$ and $1 \leq m \leq n^\alpha$).

Similarly to the proof of Lemma 2.15, we obtain that

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{\theta\in\Theta_n} \sup_{a_n\leq h\leq b_0} \sqrt{\frac{n\phi_{\chi,\theta}(a_n)}{\log n}} \left|\widehat{g}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{g}_{h,\theta}(\chi)\right)\right| \geq \eta_0' \right\} \\ \leq\ & n^\alpha L(n) \max_{m=1,\ldots,n^\alpha} \max_{j=1,\ldots,L(n)} \mathbb{P}\left(\max_{1\leq k\leq n} \left\| \sqrt{k}\alpha_k'(g) \right\|_{\mathcal{G}_{j,m}} \geq \eta_0' \sqrt{n \log n \phi_{\chi,\theta_m}\left(h_j/2\right)} \right). \end{aligned} \tag{2.57}$$

Taking into account Assumption (2.21) and Assumption (2.22), we get:

$$\mathbb{E}\left(G_{j,m}'(\mathcal{X},Y)^p\right) \leq C^p \phi_{\chi,\theta_m}\left(h_j/2\right). \tag{2.58}$$

(Note that, we can write

$$\begin{aligned} \mathbb{E}\left(Y^p K\left(h_j^{-1} d_{\theta_m}(\chi,\mathcal{X})\right)^p\right) &\leq \mathbb{E}\left(\mathbb{E}\left(Y^p \middle| \langle\theta_0,\mathcal{X}\rangle\right) K\left(h_j^{-1} d_{\theta_m}(\chi,\mathcal{X})\right)^p\right) \\ &\leq \mathbb{E}\left(\mathbb{E}\left(|Y|^p \middle| \langle\theta_0,\mathcal{X}\rangle\right) K\left(h_j^{-1} d_{\theta_m}(\chi,\mathcal{X})\right)^p\right) \\ &\leq C_4 C_6^p \phi_{\chi,\theta_m}\left(h_j/2\right).) \end{aligned}$$

Let us denote

$$\sigma'^2 = \mathbb{E}\left(G'_{j,m}(\mathcal{X}, Y)^2\right), \quad \beta'_n = \mathbb{E}\left(\left\|\sqrt{n}\alpha'_n(g)\right\|_{\mathcal{G}'_{j,m}}\right).$$

Using (2.58), we get that

$$\sigma'^2 = O\left(\phi_{\chi,\theta_m}\left(h_j/2\right)\right),$$

and utilizing the same ideas as for the proof of Lemma 2.15, we obtain

$$\beta'_n = O\left(\sqrt{n\phi_{\chi,\theta_m}\left(h_j/2\right)}\right).$$

Now, from the Bernstein's inequality (see Lemma 2.13), it is obtained that

$$\begin{aligned}
&\mathbb{P}\left\{\max_{1\leq k\leq n}\left\|\sqrt{k}\alpha'_k(g)\right\|_{\mathcal{G}'_{j,m}} \geq \eta'_0\sqrt{n\log n\phi_{\chi,\theta_m}\left(h_j/2\right)}\right\} \\
&\leq \mathbb{P}\left\{\max_{1\leq k\leq n}\left\|\sqrt{k}\alpha'_k(g)\right\|_{\mathcal{G}'_{j,m}} \geq \beta'_n + t\right\} \\
&\leq n^{-C'\eta'^2_0}, \qquad (2.59)
\end{aligned}$$

while from (2.57) and (2.59) we have that

$$\mathbb{P}\left\{\sup_{\theta\in\Theta_n}\sup_{a_n\leq h\leq b_0}\sqrt{\frac{n\phi_{\chi,\theta}(a_n)}{\log n}}\left|\widehat{g}_{h,\theta}(\chi) - \mathbb{E}\left(\widehat{g}_{h,\theta}(\chi)\right)\right| \geq \eta'_0\right\} \leq n^{\alpha}n^{-C'\eta'^2_0}\log n. \quad (2.60)$$

Finally, by choosing η'_0 such that $C'\eta'^2_0 - \alpha > 1$, (2.56) follows from (2.60). ■

2.7.4 Proof of Theorem 2.5 (b)

We will follow the scheme of Kara-Zaitri et al. [66], but taking into account that in our setting, for a fixed k, the random bandwidth also depends on θ.

We have that

$$\begin{aligned}
&\sup_{\theta\in\Theta_n}\sup_{k_{1,n}\leq k\leq k_{2,n}}\left|\widehat{r}^{*}_{k,\theta}(\chi)-r_{\theta_0}(\chi)\right| \\
=\;&\sup_{\theta\in\Theta_n}\sup_{k_{1,n}\leq k\leq k_{2,n}}\left|\widehat{r}^{*}_{k,\theta}(\chi)-r_{\theta_0}(\chi)\right| 1_{\left\{\phi^{-1}_{\chi,\theta}\left(\frac{\rho_n k_{1,n}}{n}\right)\leq H_{k,\chi,\theta}\leq\phi^{-1}_{\chi,\theta}\left(\frac{k_{2,n}}{\rho_n n}\right)\right\}} \\
&+\sup_{\theta\in\Theta_n}\sup_{k_{1,n}\leq k\leq k_{2,n}}\left|\widehat{r}^{*}_{k,\theta}(\chi)-r_{\theta_0}(\chi)\right| 1_{\left\{H_{k,\chi,\theta}\notin\left(\phi^{-1}_{\chi,\theta}\left(\frac{\rho_n k_{1,n}}{n}\right),\phi^{-1}_{\chi,\theta}\left(\frac{k_{2,n}}{\rho_n n}\right)\right)\right\}},
\end{aligned}$$

where $\rho_n\in(0,1)$ was defined in Assumption (2.31). Thus, taking Assumption (2.25) into account, the proof of our theorem is completed once we prove the three following results:

$$\sup_{\theta\in\Theta_n}\sup_{\phi^{-1}_{\chi,\theta}\left(\frac{\rho_n k_{1,n}}{n}\right)\leq h\leq\phi^{-1}_{\chi,\theta}\left(\frac{k_{2,n}}{\rho_n n}\right)}|\widehat{r}_{h,\theta}(\chi)-r_{\theta_0}(\chi)|=O\left(f^{-1}\left(\frac{k_{2,n}}{\rho_n n}\right)^{\beta_0}\right)+O_{a.co.}\left(\sqrt{c_n}\right), \tag{2.61}$$

$$\sum_{n=1}^{\infty}\sum_{m=1}^{n^{\alpha}}\sum_{k=k_{1,n}}^{k_{2,n}}\mathbb{P}\left(H_{k,\chi,\theta_m}\leq\phi^{-1}_{\chi,\theta_m}\left(\frac{\rho_n k_{1,n}}{n}\right)\right)<\infty, \tag{2.62}$$

and

$$\sum_{n=1}^{\infty}\sum_{m=1}^{n^{\alpha}}\sum_{k=k_{1,n}}^{k_{2,n}}\mathbb{P}\left(H_{k,\chi,\theta_m}\geq\phi^{-1}_{\chi,\theta_m}\left(\frac{k_{2,n}}{n\rho_n}\right)\right)<\infty, \tag{2.63}$$

where we have denoted

$$c_n=\frac{\log n}{nf\left(\lambda f^{-1}(\rho_n k_{1,n}/n)\right)}.$$

On the one hand, the proof of (2.61) is a direct consequence of Theorem 2.5(a). Specifically, taking assumptions (2.31), (2.32) and (2.34) into account, it suffices to consider $a_n=\lambda f^{-1}\left(\frac{\rho_n k_{1,n}}{n}\right)$ and $b_n=\delta f^{-1}\left(\frac{k_{2,n}}{\rho_n n}\right)$ in Theorem 2.5(a). On the other hand, to prove (2.62) and (2.63) we will use Lemma 2.14 and we will proceed in a similar way as in Kara-Zaitri et al. [66]. Firstly, note that

$$\begin{aligned}\mathbb{P}\left(H_{k,\chi,\theta} \leq \phi^{-1}_{\chi,\theta_m}\left(\frac{\rho_n k_{1,n}}{n}\right)\right) &\leq \mathbb{P}\left(\sum_{i=1}^{n} 1_{B_{\theta_m}\left(\chi,\phi^{-1}_{\theta_m}\left(\frac{\rho_n k_{1,n}}{n}\right)\right)}(\mathcal{X}_i) \geq k\right) \\ &= \mathbb{P}\left(\sum_{i=1}^{n} 1_{B_{\theta_m}\left(\chi,\phi^{-1}_{\theta_m}\left(\frac{\rho_n k_{1,n}}{n}\right)\right)}(\mathcal{X}_i) \geq k\frac{k_{1,n}\rho_n}{k_{1,n}\rho_n}\right).\end{aligned} \tag{2.64}$$

In the same way,

$$\mathbb{P}\left(H_{k,\chi,\theta} \geq \phi^{-1}_{\chi,\theta_m}\left(\frac{k_{2,n}}{n\rho_n}\right)\right) \leq \mathbb{P}\left(\sum_{i=1}^{n} 1_{B_{\theta_m}\left(\chi,\phi^{-1}_{\theta_m}\left(\frac{k_{2,n}}{n\rho_n}\right)\right)}(\mathcal{X}_i) \leq k\frac{k_{2,n}\rho_n}{k_{2,n}\rho_n}\right). \tag{2.65}$$

In both expressions (2.64) and (2.65), we have a sum of independent variables following a Bernouilli distribution. In the first case, that sum of variables has mean

$$\begin{aligned}\mu = n\mathbb{E}\left(1_{B_{\theta_m}\left(\chi,\phi^{-1}_{\theta_m}\left(\frac{\rho_n k_{1,n}}{n}\right)\right)}(\mathcal{X})\right) &= n\mathbb{P}\left(d_{\theta_m}(\chi,\mathcal{X}) \leq \phi^{-1}_{\theta_m}\left(\frac{\rho_n k_{1,n}}{n}\right)\right) \\ &= n\phi_{\chi,\theta_m}\left(\phi^{-1}_{\chi,\theta_m}\left(\frac{\rho_n k_{1,n}}{n}\right)\right) \\ &= \rho_n k_{1,n},\end{aligned} \tag{2.66}$$

while in the second case, following the same procedure, the mean is

$$\mu = n\mathbb{E}\left(1_{B_{\theta_m}\left(\chi,\phi^{-1}_{\theta_m}\left(\frac{k_{2,n}}{\rho_n n}\right)\right)}(\mathcal{X})\right) = \frac{k_{2,n}}{\rho_n}. \tag{2.67}$$

Then, by means of Lemma 2.14, taking $\omega = k/(k_{1,n}\rho_n) - 1$ for expression (2.64), and $\omega = 1 - k\rho_n/k_{2,n}$ for expression (2.65), one can obtain:

$$\mathbb{P}\left(H_{k,\chi,\theta} \leq \phi^{-1}_{\chi,\theta_m}\left(\frac{\rho_n k_{1,n}}{n}\right)\right) \leq \exp\left\{-\frac{(1-\rho_n)k_{1,n}}{4}\right\} + \exp\left\{-\frac{(1-\rho_n)^2 k_{1,n}}{4\rho_n}\right\} \tag{2.68}$$

and

$$\mathbb{P}\left(H_{k,\chi,\theta} \geq \phi^{-1}_{\chi,\theta_m}\left(\frac{k_{2,n}}{n\rho_n}\right)\right) \leq \exp\left\{-\frac{(1-\rho_n)^2 k_{2,n}}{2\rho_n}\right\}. \tag{2.69}$$

Note that the first addend in the bound (2.68) is related to the case where $\min\{\omega^2, \omega\} = \omega$, while the second one, which was forgotten in Kara-Zaitri et al. [66], corresponds to the case where $\min\{\omega^2, \omega\} = \omega^2$; for details on the role of ω, see Lemma 2.14. Therefore, it is obtained that

$$\begin{aligned}\sum_{m=1}^{n^\alpha}\sum_{k=k_{1,n}}^{k_{2,n}} \mathbb{P}\left(H_{k,\chi,\theta_m} \leq \phi^{-1}_{\chi,\theta_m}\left(\frac{\rho_n k_{1,n}}{n}\right)\right) &\leq n^\alpha k_{2,n}\left(n^{-\frac{1-\rho_n}{4}\frac{k_{1,n}}{\ln n}} + n^{-\frac{(1-\rho_n)^2}{4\rho_n}\frac{k_{1,n}}{\ln n}}\right) \\ &\leq n^{\alpha+1-\frac{1-\rho_n}{4}\frac{k_{1,n}}{\ln n}} + n^{\alpha+1-\frac{(1-\rho_n)^2}{4\rho_n}\frac{k_{1,n}}{\ln n}},\end{aligned} \tag{2.70}$$

and

$$\sum_{m=1}^{n^\alpha}\sum_{k=k_{1,n}}^{k_{2,n}} \mathbb{P}\left(H_{k,\chi,\theta_m} \geq \phi^{-1}_{\chi,\theta_m}\left(\frac{k_{2,n}}{n\rho_n}\right)\right) \leq n^\alpha k_{2,n} n^{-\frac{(1-\rho_n)^2}{2\rho_n}\frac{k_{2,n}}{\ln n}} \leq n^{\alpha+1-\frac{(1-\rho_n)^2}{2\rho_n}\frac{k_{2,n}}{\ln n}}. \tag{2.71}$$

Finally, from Assumption (2.33) together with the bounds (2.70) and (2.71), we obtain (2.62) and (2.63). This completes the proof of the theorem. ■

2.7.5 Proof of Corollary 2.6

It is enough to check that the assumptions used in Theorem 2.5(b) hold and then, to write the corresponding rate of convergence for the particular case considered in Corollary 2.6. ■

2.7.6 Proof of Corollary 2.8

Trivial. ■

Chapter 3

Contributions on the Semi-Functional Partial Linear Single-Index Model

3.1 Introduction

In the previous chapter, uniform rates of consistency over all the parameters involved in kNN and kernel estimators of the FSIM were stated. Such results, in addition to the value on their own, have at least two important applications. On the one hand, as formulated in Section 2.3.4, they give theoretical validation to data-driven choices of the parameters of the model. On the other hand, these results are, somehow, a pillar to obtain analogous asymptotics for more complex models which are extension of the FSIM.

Precisely, in this chapter we are going to take advantage of this second fact, dealing with the SFPLSIM, briefly presented in Section 1.3.2. Specifically, the SFPLSIM is given by the expression

$$Y = X_1\beta_{01} + \cdots + X_p\beta_{0p} + r\left(\langle\theta_0, \mathcal{X}\rangle\right) + \varepsilon, \tag{3.1}$$

where X_j $(j = 1, \ldots, p)$ and Y are real random variables, while $\mathcal{X}$ is a functional random variable valued in a separable Hilbert space $\mathcal{H}$ with inner product denoted by

$\langle\cdot,\cdot\rangle$. In expression (3.1), ε denotes a random error verifying $\mathbb{E}\left(\varepsilon|X_1,\ldots,X_p,\mathcal{X}\right)=0$. The vector $\boldsymbol{\beta}_0=(\beta_{01},\ldots,\beta_{0p})^\top\in\mathbb{R}^p$, the functional direction $\theta_0\in\mathcal{H}$ and the link real-valued function $r(\cdot)$ are supposed unknown. Moreover, to ensure identifiability of model (3.1) we assume conditions (2.2) and (2.3) (see e.g. Ait-Saïdi et al. [3]; see also Wang et al. [110] for other ways to ensure identifiability).

One of the nice features of the model (3.1) is to allow sets of predictors to be a mixture of functional and multivariate ones. In addition, the model combines single-index ideas (for dealing with the functional predictor) together with partial linear ideas (for dealing with the multivariate one). Therefore, it is a semiparametric model, which are more interesting candidates for reducing dimensionality effects, but being able to capture, as wide as possible, information on the data (see the Introduction of this dissertation).

The aim of this chapter is to develop a kNN procedure for estimating the smooth component of the model (3.1) (see Section 3.2). In Section 3.3, rates of uniform consistency are obtained in a general way allowing for fully automatic estimates. As a by-product, we state similar results for usual Nadaraya-Watson functional kernel regression. A short simulation study is reported along Section 3.4 for highlighting the advantages of the kNN procedure. Finally, the Tecator's dataset is analysed in Section 3.5 and a comparative study will show the interest of semiparametrics. Technical proofs are gathered in the Section 3.6.

3.2 The statistics

First of all, assume that we have a statistical sample of n vectors $\{(X_{i1},\ldots,X_{ip},\mathcal{X}_i,Y_i)\}_{i=1}^n$ i.i.d. as $(X_1,\ldots,X_p,\mathcal{X},Y)$ verifying model (3.1). That is,

$$Y_i=X_{i1}\beta_{01}+\cdots+X_{ip}\beta_{0p}+r\left(\langle\theta_0,\mathcal{X}_i\rangle\right)+\varepsilon_i\quad(i=1,\ldots,n).$$

For each $\theta\in\mathcal{H}$, we consider the operator $r_\theta(\cdot):\mathcal{H}\longrightarrow\mathbb{R}$ defined in (2.4). Note that, in this case

$$r_{\theta_0}(\mathcal{X})=\mathbb{E}\left(Y-\boldsymbol{X}^\top\boldsymbol{\beta}_0|\mathcal{X}\right),\tag{3.2}$$

where $\boldsymbol{X}=(X_1,\ldots,X_p)^\top$ and $\boldsymbol{\beta}_0=(\beta_{01},\ldots,\beta_{0p})^\top$; kNN ideas are used for es-

timating $r_{\theta_0}(\cdot)$ from a smoothing factor $k = k_n \in \mathbb{Z}^+$ and a kernel function K as follows:

$$\widehat{r}^*_{k,\theta,\boldsymbol{\beta}}(\chi) = \sum_{i=1}^{n} w^*_{n,k,\theta}(\chi, \mathcal{X}_i) \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\right), \tag{3.3}$$

where $w^*_{n,k,\theta}(\cdot, \mathcal{X}_i)$ was defined in (2.7). It is worth being noted that this kNN statistic is an extension of the usual Nadaraya-Watson one,

$$\widehat{r}_{h,\theta,\boldsymbol{\beta}}(\chi) = \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\right), \tag{3.4}$$

where $w_{n,h,\theta}(\cdot, \mathcal{X}_i)$ was defined in (2.11) with $h \in \mathbb{R}^+$ being the bandwidth ($h = h_n$ depends on n).

We should emphasize that, to estimate $r(\cdot)$ in (3.1) by means of (3.3) and (3.4), it is needed to introduce in (3.3) and (3.4) estimates not only for θ_0 (as in the case of the FSIM; see Chapter 2) but also for $\boldsymbol{\beta}_0$. This fact is the major difficulty for the theoretical study of the estimator of $r(\cdot)$ presented in this chapter compared to that of the FSIM (2.1).

3.3 Asymptotic theory

3.3.1 Additional assumptions

In order to state results of uniform (over k, θ and $\boldsymbol{\beta}$) almost-complete consistency for $\widehat{r}^*_{k,\theta,\boldsymbol{\beta}}(\chi)$ and $\widehat{r}_{h,\theta,\boldsymbol{\beta}}(\chi)$, in addition, to assumptions presented in Section 2.3, we need to formulate the following technical assumptions:

About the model. We assume that:

- The conditional moments of the errors of the linear regression verify

$$\exists m \geq 2,\ \exists C > 0 \text{ such that } \mathbb{E}\left(|Y - \boldsymbol{X}^\top \boldsymbol{\beta}_0|^m | \mathcal{X}\right) < C < \infty,\ a.s. \tag{3.5}$$

- Furthermore, let us denote by N_{χ,θ_0} a fixed neighbourhood of $\chi \in \mathcal{H}$ in

the topological space induced by the semi-metric $d_{\theta_0}(\cdot,\cdot)$ (2.5), and denote

$$g_{j,\theta_0}(\chi) = \mathbb{E}\left(X_{ij} | \langle \theta_0, \mathcal{X}_i \rangle = \langle \theta_0, \chi \rangle\right) (j = 1, \ldots, p),$$

that is, the functional single-index regression operators of each X_j over $\mathcal{X}$. Hölder type conditions are verified for regression operators, in the sense that exist constants $0 \leq C < \infty$ and $\alpha_0 > 0$ such that, $\forall \chi_1, \chi_2 \in N_{\chi,\theta_0}$, $\forall z \in \{r_{\theta_0}, g_{1,\theta_0}, \ldots, g_{p,\theta_0}\}$,

$$|z(\chi_1) - z(\chi_2)| \leq C d_{\theta_0}(\chi_1, \chi_2)^{\alpha_0}. \tag{3.6}$$

Furthermore, for fixed $\chi \in \mathcal{H}$ it is verified that

$$\max_{j=1,\ldots,p} |g_{j,\theta_0}(\chi)| = O(1). \tag{3.7}$$

About the space of linear parameters. It is assumed that vectors $\boldsymbol{\beta}$ are not far from the target vector $\boldsymbol{\beta}_0$, in the sense that there exists a sequence $\{c_n\}$, with $c_n \to 0$ as $n \to \infty$, such that

$$\Psi_n = \{\boldsymbol{\beta} \in \mathbb{R}^p; ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| = O(c_n)\}. \tag{3.8}$$

Remark 3.1 *These hypotheses allow to deal with the complexity of the model and to obtain general results, but they are actually not very restrictive. On one hand, (3.5), (3.6), (3.7) are standard assumptions in regression models mixing linear and nonparametric structures (see e.g. Aneiros-Pérez and Vieu [11]). On the other hand, Assumption (3.8) is added for controlling the bias in the estimation of the linear coefficients in model (3.1).*

3.3.2 Main results

The next Theorem 3.2 is the main part of this chapter.

Theorem 3.2 *Assume that conditions (2.22), (2.25)-(2.29), (2.35), (3.1) and (3.5)-(3.8) hold*

(a) If in addition Assumption (2.30) holds, then we have that

$$\sup_{\boldsymbol{\beta}\in\Psi_n}\sup_{\theta\in\Theta_n}\sup_{a_n\leq h\leq b_n}|\widehat{r}_{h,\theta,\boldsymbol{\beta}}(\chi)-r_{\theta_0}(\chi)|=O\left(b_n^{\alpha_0}\right)+O_{a.co.}\left(\sqrt{\frac{\log n}{nf(a_n)}}\right)+O\left(c_n\right).$$

(b) If in addition assumptions (2.31)-(2.34) hold, then we have that

$$\sup_{\boldsymbol{\beta}\in\Psi_n}\sup_{\theta\in\Theta_n}\sup_{k_{1,n}\leq k\leq k_{2,n}}|\widehat{r}^*_{k,\theta,\boldsymbol{\beta}}(\chi)-r_{\theta_0}(\chi)|=O\left(f^{-1}\left(\frac{k_{2,n}}{\rho_n n}\right)^{\alpha_0}\right)$$
$$+O_{a.co.}\left(\sqrt{\frac{\log n}{nf\left(\lambda f^{-1}(\rho_n k_{1,n}/n)\right)}}\right)+O\left(c_n\right).$$

Remark 3.3 *Note that the first two terms in both rates of convergence are the same as those gotten in Theorem 2.5 for the FSIM (2.1). The third term in the rates corresponds to the bias of estimating the linear coefficients of the model. For small enough values of c_n, this third term could be much smaller than both previous ones. This fact highlights that the presence of linear component in the SFPLSIM does not deteriorate the asymptotics. Note also that, under standard additional conditions on $f(\cdot)$, ρ_n, $k_{1,n}$ and $k_{2,n}$ (or a_n and b_n) (see Remark 2.7), the rates in Theorem 3.2 are the same as if $\mathcal{X}$ were one-dimensional. In other words, the semiparametric model has achieved its goal of being insensitive to dimensionality effects.*

Theorem 3.2 confirms the well-known fact that practical using of both, kernel- and kNN-based methods, is linked with the choice of a smoothing parameter (k or h, respectively) balancing bias and variance effects. One of the most important features of our result for kNN (respectively, for kernel) is to be uniform over $k\in[k_{1,n}k_{2,n}]$ (respectively, $h\in[a_n,b_n]$), $\boldsymbol{\beta}\in\Psi_n$ and $\theta\in\Theta_n$. That feature allows to say that the same asymptotics are available when k (h), $\boldsymbol{\beta}$ and θ are random variables valued in $[k_{1,n}k_{2,n}]$ ($[a_n,b_n]$), Ψ_n and Θ_n, respectively. In particular, when k (h), $\boldsymbol{\beta}$ and θ are data-driven selected. This property is formulated in the next corollary, whose proof is obvious (because of the uniform feature of previous theorem). That corollary makes the proposed methodology completely automatic, in the sense that the main parameter (i.e. k or h), as well as the other two (i.e. $\boldsymbol{\beta}$ and θ) can be selected from the sample without deteriorating its asymptotic behaviour.

3.3.3 Data-driven parameters selection

Corollary 3.4 • *Assume that* $\widehat{h}$, $\widehat{\boldsymbol{\beta}}_{\widehat{h}}$ *and* $\widehat{\theta}_{\widehat{h}}$ *are random variables taking values in* $[a_n, b_n]$, Ψ_n *and* Θ_n, *respectively, being data-driven in the sense that they depend on the statistical sample* $\mathcal{D}_n = \{(X_{i1}, \ldots, X_{ip}, \mathcal{X}_i, Y_i),\ i = 1, \ldots, n\}$. *Under assumptions of Theorem 3.2(a), we have that*

$$\left|\widehat{r}_{\widehat{h},\widehat{\theta}_{\widehat{h}},\widehat{\boldsymbol{\beta}}_{\widehat{h}}}(\chi) - r_{\theta_0}(\chi)\right| = O\left(b_n^{\alpha_0}\right) + O_{a.co.}\left(\sqrt{\frac{\log n}{nf(a_n)}}\right) + O\left(c_n\right).$$

• *Assume that* $\widehat{k}$, $\widehat{\boldsymbol{\beta}}^*_{\widehat{k}}$ *and* $\widehat{\theta}^*_{\widehat{k}}$ *are random variables taking values in* $[k_{1,n}k_{2,n}]$, Ψ_n *and* Θ_n, *respectively, being data-driven in the sense that they depend on the statistical sample* $\mathcal{D}_n$. *Under assumptions of Theorem 3.2(b), we have that*

$$\begin{aligned}|\widehat{r}^*_{\widehat{k},\widehat{\theta}^*_{\widehat{k}},\widehat{\boldsymbol{\beta}}^*_{\widehat{k}}}(\chi) - r_{\theta_0}(\chi)| &= O\left(f^{-1}\left(\frac{k_{2,n}}{\rho_n n}\right)^{\alpha_0}\right) + O_{a.co.}\left(\sqrt{\frac{\log n}{nf\left(\lambda f^{-1}(\rho_n k_{1,n}/n)\right)}}\right) \\ &+ O\left(c_n\right).\end{aligned}$$

This corollary allows to have asymptotics for any automatic data-driven parameters. To fix the ideas let us just mention one example. Kernel-based estimators $\widehat{\theta}_{\widehat{h}}$ and $\widehat{\boldsymbol{\beta}}_{\widehat{h}}$, and kNN-based estimators $\widehat{\theta}^*_{\widehat{k}}$ and $\widehat{\boldsymbol{\beta}}^*_{\widehat{k}}$ could be constructed from the ordinary least squares (OLS) procedure applied to a linear model in which the effects of the functional covariate have been extracted. That is, estimators in the kernel case were constructed by minimizing the score function

$$\mathcal{Q}_h\left(\boldsymbol{\beta}, \theta\right) = \frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_{h,\theta} - \widetilde{\boldsymbol{X}}_{h,\theta}\boldsymbol{\beta}\right)^\top \left(\widetilde{\boldsymbol{Y}}_{h,\theta} - \widetilde{\boldsymbol{X}}_{h,\theta}\boldsymbol{\beta}\right), \tag{3.9}$$

while in the kNN case we minimize the score function

$$\mathcal{Q}^*_k\left(\boldsymbol{\beta}, \theta\right) = \frac{1}{2}\left(\widetilde{\boldsymbol{Y}}^*_{k,\theta} - \widetilde{\boldsymbol{X}}^*_{k,\theta}\boldsymbol{\beta}\right)^\top \left(\widetilde{\boldsymbol{Y}}^*_{k,\theta} - \widetilde{\boldsymbol{X}}^*_{k,\theta}\boldsymbol{\beta}\right). \tag{3.10}$$

In expressions (3.9) and (3.10), $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$, with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, while for any $(n \times q)$-matrix $\boldsymbol{A}$ $(q \geq 1)$, any $\theta \in \Theta_n$ and

bandwidth h or number of neighbours k, respectively, we denote

$$\widetilde{\boldsymbol{A}}_{h,\theta} = (\boldsymbol{I} - \boldsymbol{W}_{h,\theta})\,\boldsymbol{A} \quad \text{with} \quad \boldsymbol{W}_{h,\theta} = (w_{n,h,\theta}(\mathcal{X}_i, \mathcal{X}_j))_{i,j}\,, \tag{3.11}$$

and

$$\widetilde{\boldsymbol{A}}^*_{k,\theta} = \left(\boldsymbol{I} - \boldsymbol{W}^*_{k,\theta}\right)\boldsymbol{A} \quad \text{with} \quad \boldsymbol{W}^*_{k,\theta} = \left(w^*_{n,k,\theta}(\mathcal{X}_i, \mathcal{X}_j)\right)_{i,j}\,.$$

Then cross-validation ideas (either leave-one-out or k-fold cross-validation) could be used to obtain an estimate $\widehat{k}$ and $\widehat{h}$ (see Section 2.3.4).

3.4 Simulation study

3.4.1 The design

Samples of i.i.d. data $\mathcal{D}_n = \{(X_{i1}, X_{i2}, X_{i3}, \mathcal{X}_i, Y_i)\}_{i=1}^{n+25}$ were generated from the model

$$Y_i = X_{i1}\beta_{01} + X_{i2}\beta_{02} + X_{i3}\beta_{03} + \alpha r\left(\langle\theta_0, \mathcal{X}_i\rangle\right) + (1-\alpha)m(\mathcal{X}_i) + \varepsilon_i \quad (i = 1, \ldots, n+25). \tag{3.12}$$

Note that the case $\alpha = 1$ gives the SFPLSIM studied in this chapter, while values $\alpha \in [0, 1)$ allow to show a sensitivity analysis of the proposed method; in particular, case $\alpha = 0$ provides the SFPLM (1.7). In model (3.12):

- The functional covariate, $\mathcal{X}_i$ $(i = 1, \ldots, n+25)$, was generated following expression (2.38). As in that expression, to build a dataset of heterogeneous curves, the random variables a_i, b_i and c_i were independent variables, uniformly distributed either on $[5, 10]$ with probability 0.5 or on $[20, 20.5]$ with probability 0.5 (note that independence means both between and within vectors $(a_i, b_i, c_i)^\top$). These curves were discretized on the same grid of 100 equispaced points in $[0, 1]$.

- The vector of real covariates, $(X_{i1}, X_{i2}, X_{i3})^\top$ $(i = 1, \ldots, n+25)$, was generated from a multivariate normal distribution with zero mean and covariance matrix given by $(\rho^{|j-k|})_{jk}$ $(j, k = 1, 2, 3)$. Two values for ρ were considered: $\rho = 0$ (independence between linear covariates) and $\rho = 0.5$.

- The i.i.d. random errors, ε_i $(i = 1, \ldots, n + 25)$, were simulated from a $N(0, \sigma_\varepsilon^2 = c\sigma_r^2)$ where σ_r^2 is the empirical variance of $X_{i1}\beta_{01} + X_{i2}\beta_{02} + X_{i3}\beta_{03} + \alpha r\left(\langle\theta_0, \mathcal{X}_i\rangle\right) + (1-\alpha)m(\mathcal{X}_i)$. The signal-to-noise ratio c has been taken equal to $c = 0.025$.
- The true vector of linear coefficients was $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \beta_{03})^\top = (-1, 0.5, 1.5)^\top$.
- The true direction of projection, θ_0, was constructed as described in Section 2.4. Values $l = 3$ and $m_n = 3$ were considered and the vector of coefficients of θ_0 in expression (2.36) was that given in (2.39).
- The inner product and the link function in the semiparametric component were $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$ and $r(\langle\theta_0, \chi\rangle) = \langle\theta_0, \chi\rangle^3$, respectively.
- As link function of the nonparametric component we have considered $m(\chi_i) = 2\sqrt{c_i}$ (note that $\mathcal{X}_i = \mathcal{X}_{a_i,b_i,c_i}$).

For each simulation case $(n, \rho, \alpha) \in \{50, 100, 200\} \times \{0, 0.5\} \times \{0.8, 0.9, 1\}$, $M = 100$ independent samples were generated from (3.12). Each sample $\mathcal{D}_n$ was split into two subsamples: a training sample, $\mathcal{D}_{n,train} = \{(X_{i1}, X_{i2}, X_{i3}, \mathcal{X}_i, Y_i)\}_{i=1}^{n}$, and a testing sample, $\mathcal{D}_{n,test} = \{(X_{i1}, X_{i2}, X_{i3}, \mathcal{X}_i, Y_i)\}_{i=n+1}^{n+25}$. The tuning parameters ($\widehat{h}$ and $\widehat{k}$) were constructed from the training sample by means of the 10-fold cross-validation procedure. In addition, we only use the training sample for getting estimates of θ_0 ($\widehat{\theta}_0$ with the kernel-based method and $\widehat{\theta}_0^*$ with the kNN-based one) and of $\boldsymbol{\beta}_0$ ($\widehat{\boldsymbol{\beta}}_0$ with the kernel-based procedure and $\widehat{\boldsymbol{\beta}}_0^*$ with the kNN-based one). These estimates were obtained by minimizing the score functions (3.9) and (3.10), respectively, as suggested at the end of Section 3.3.2. The set of eligible directions Θ_n was constructed as recommended in Section 2.4.

For measuring the performance of the proposed estimators we computed

$$||\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0||^2 = \sum_{j=1}^{3}\left(\widehat{\beta}_{0j} - \beta_{0j}\right)^2, \quad ||\widehat{\boldsymbol{\beta}}_0^* - \boldsymbol{\beta}_0||^2 = \sum_{j=1}^{3}\left(\widehat{\beta}_{0j}^* - \beta_{0j}\right)^2,$$

$$||\widehat{\theta}_0 - \theta_0||^2 = \int_0^1\left(\widehat{\theta}_0(t) - \theta_0(t)\right)^2 dt, \quad ||\widehat{\theta}_0^* - \theta_0||^2 = \int_0^1\left(\widehat{\theta}_0^*(t) - \theta_0(t)\right)^2 dt, \quad (3.13)$$

and the MSEP (2.40), with $n_{test} = 25$.

3.4.2 Results

The results are summarized in Tables 3.1, 3.2 and 3.3 below. It appears that both methods are benefited by the increase of the sample size. Furthermore, the dependence between the covariates ($\rho = 0.5$) with linear effect is detrimental in the estimation of the linear coefficients, but benefits the estimation of θ_0 and it has a positive effect on the MSEP, as it decreases with respect to the independent case ($\rho = 0$) (this behaviour has been observed in other contexts like variable selection; see Huang et al. [65] or Bühlmann and Meier [21] for the SLM, Xie and Huang [116] for the sparse nonfunctional partial linear model or Aneiros et al. [7] for the SSFPLM).

More importantly, it seems that for both independent covariates and correlated ones, the kNN-based procedure clearly outperforms results obtained with the kernel-based procedure by capturing heterogeneous structure of the data. Finally, the proposed procedure is not very sensitive, at least in this example, to slight modifications (high values of α) in the effect of the functional covariate.

Table 3.1: Averaged MSEPs with 10-fold cross-validation selectors for h and k.

		$n = 50$		$n = 100$		$n = 200$	
α	ρ	kernel	kNN	kernel	kNN	kernel	kNN
1	0	0.1959	0.1626	0.1619	0.1297	0.1239	0.1024
	0.5	0.1791	0.1393	0.1458	0.1154	0.1068	0.0893
0.9	0	0.2088	0.1785	0.1674	0.1431	0.1350	0.1121
	0.5	0.1838	0.1583	0.1500	0.1278	0.1187	0.0992
0.8	0	0.2193	0.1976	0.1858	0.1591	0.1473	0.1200
	0.5	0.2016	0.1767	0.1654	0.1426	0.1307	0.1067

Table 3.2: Averaged squared errors for $\boldsymbol{\beta}_0$ with 10-fold cross-validation selectors for h and k.

		$n = 50$		$n = 100$		$n = 200$	
α	ρ	kernel	kNN	kernel	kNN	kernel	kNN
1	0	0.0133	0.0097	0.0043	0.0041	0.0021	0.0018
	0.5	0.0181	0.0120	0.0059	0.0058	0.0025	0.0021
0.9	0	0.0140	0.0105	0.0047	0.0044	0.0022	0.0020
	0.5	0.0183	0.0138	0.0063	0.0064	0.0026	0.0024
0.8	0	0.0141	0.0117	0.0049	0.0047	0.0025	0.0022
	0.5	0.0187	0.0154	0.0067	0.0069	0.0029	0.0028

Table 3.3: Averaged squared errors for θ_0 with 10-fold cross-validation selectors for h and k.

		$n = 50$		$n = 100$		$n = 200$	
α	ρ	kernel	kNN	kernel	kNN	kernel	kNN
1	0	0.0950	0.0507	0.0715	0.0413	0.0603	0.0070
	0.5	0.0933	0.0463	0.0659	0.0389	0.0618	0.0061
0.9	0	0.0958	0.0656	0.0713	0.0595	0.0679	0.0330
	0.5	0.0931	0.0622	0.0697	0.0586	0.0643	0.0302
0.8	0	0.0921	0.0781	0.0871	0.0759	0.0732	0.0757
	0.5	0.0895	0.0758	0.0851	0.0746	0.0756	0.0751

3.5 Application to real data

This section is devoted to illustrate the usefulness of the SFPLSIM (3.1), as well as to compare the performance of kernel and kNN procedures.

3.5.1 The data

As in Chapter 2, in this real data application we will analyse Tecator's data (see Sections 1.1 and 2.6.1), but in this case we will consider two new scalar covariates, and we will only study the benchmark partition (see Section 2.6.2.2). As remember, "Tecator's data" contains measurements of contents of fatness (Y_i) for 215 pieces of meat, as well as the near-infrared absorbance spectras ($\mathcal{X}_i$) observed on 100 equally wavelengths in the range $850-1050$ nm. The new scalar covariates considered in this section are the contents of protein (X_{1i}) and the contents of moisture (X_{2i}) of each piece. Remember that left panel in Figure 1.2 showed a sample of 100 absorbance curves.

Our purpose in this real data application is to model the link between fat content and the other variables, with aim to predict the fat content. We will split the original sample into two subsamples: a training sample, $\mathcal{D}_{train} = \{(X_{i1}, X_{i2}, \mathcal{X}_i, Y_i)\}_{i=1}^{160}$, and a testing one, $\mathcal{D}_{test} = \{(X_{i1}, X_{i2}, \mathcal{X}_i, Y_i)\}_{i=161}^{215}$. The estimation task is made only by means of the training sample, while the testing sample is used to measure the quality of the predictions. Then, to quantify the prediction error we use the MSEP (see (2.40)) with $n = 160$ and $n_{test} = 55$.

3.5.2 Results

Firstly, we predict the fat content of meat using two simple models involving only the two scalar covariates: a bivariate linear model (LM) and an additive spline model (ASM). Both models give similar results, which are reported in Table 3.4.

Table 3.4: MSEP for models with two scalar covariates.

	Model	MSEP
LM	$Y = \beta_{01}X_1 + \beta_{02}X_2 + \varepsilon$	1.95
ASM	$Y = r(X_1) + r(X_2) + \varepsilon$	1.93

In addition, we report in Table 3.5 the results obtained in Section 2.6.2.2 with simple models involving only the functional covariate, such as the FLM (1.1), the

FNM (1.2), the FSIM (2.1), and the FSIM combined with the application of a full nonparametric boosting step to its residuals (FSIM & FNM). It can be observed that kNN-based estimation outperforms kernel-based one in each case. However, even using the kNN procedure, each model gives results more or less similar to those obtained with models in Table 3.4.

Table 3.5: Values of the MSEPs for some functional models.

	Model	Method	MSEP
FLM	$Y = \gamma_0 + \int_{850}^{1050} \mathcal{X}^{(2)}(t)\gamma(t)dt + \varepsilon$	FPC	7.17
FNM	$Y = m(\mathcal{X}^{(2)}) + \varepsilon$	kernel	4.06
		kNN	1.79
FSIM	$Y = r\left(\langle \theta_0, \mathcal{X}^{(2)} \rangle\right) + \varepsilon$	kernel	3.49
		kNN	2.69
FSIM & FNM	$Y = r\left(\langle \theta_0, \mathcal{X}^{(2)} \rangle\right) + m(\mathcal{X}^{(1)}) + \varepsilon$	kernel	1.74
		kNN	1.53

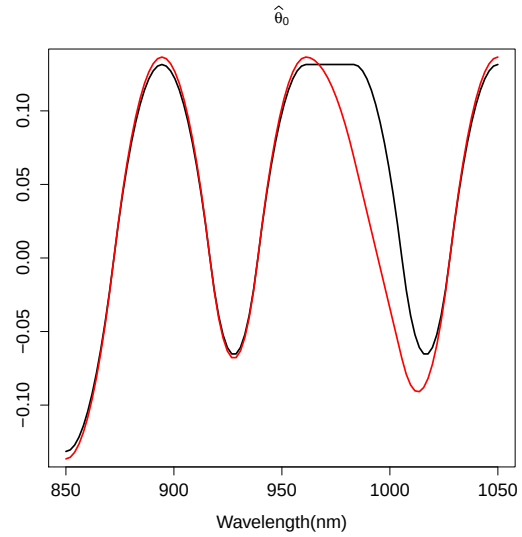
Finally, we used models incorporating both scalar and functional covariates, namely the SFPLM (1.7) and the SFPLSIM (3.1) proposed in this chapter. For both models, we use OLS-based estimators for estimating $\boldsymbol{\beta}_0$ (and also θ_0 in the SFPLSIM case) and 10-fold cross-validation for selecting k, h, the order q of the derivatives of the absorbance curves ($\mathcal{X}_i^{(q)}$) and the number m_n of regularly interior knots of the B-spline basis functions considered to construct the set of eligible directions Θ_n (for details, see Section 3.4.1). Table 3.6 summarizes the results. In both cases, the kNN-based estimation procedures outperform the kernel-based ones and the SFPLSIM offers lower MSEP than the SFPLM. More importantly, these models involving both kinds of covariate give a much smaller prediction error than models using only one kind of variables (as those in Tables 3.4 and 3.5). All in all, the SFPLSIM model with kNN estimates leads to the lowest MSEP among all models/estimates studied.

Table 3.6: Values of the MSEPs for some functional partial linear models.

	Model	Method	MSEP
SFPLM	$Y = \beta_{01}X_1 + \beta_{02}X_2 + m(\mathcal{X}^{(1)}) + \varepsilon$	kernel	0.87
		kNN	0.69
SFPLSIM	$Y = \beta_{01}X_1 + \beta_{02}X_2 + r\left(\langle\theta_0, \mathcal{X}^{(1)}\rangle\right) + \varepsilon$	kernel	0.77
		kNN	0.60

To conclude, it should be noted that, in addition to this good predictive behaviour, another great advantage of the SFPLSIM is that the functional variable enters in the model through an interpretable parameter: θ_0. The obtained estimations of this functional direction in the SFPLSIM, using both kNN and kernel-based estimation procedures, can be seen in Figure 3.1. The estimated directions show two peaks and two bumps that could give information on which wavelength ranges have the highest influence on the fat content.

Figure 3.1: Estimates of the functional direction θ_0 using kNN-based (red line) and kernel-based (black line) estimators.



3.5.3 Conclusions

The SFPLSIM studied in this chapter together with the kNN-based estimation procedure offered the best results in terms of predictive power in this real data application. In addition, the semiparametric feature of the model allows the interpretability of the derived estimations. We also would like to remember that, to obtain our estimate of θ_0, our method proposes to minimize on a predefined index set Θ_n. Therefore, its computational cost is higher than that required by efficient proposals based on functional dimension reduction techniques, as that in Wang et al. [110]. The advantage of our method against such proposals is (at least in this example) its great predictive power: considering the same Tecator subsamples and measure of the predictive performance as in Wang et al. [110], our procedure improves in a 35% the predictive power of the method in Wang et al. [110].

3.6 Appendix Chapter 3: Proofs

Before starting the proofs, we are going to introduce some additional notation. The kernel statistic and the kNN statistic associated with the estimation of $g_{j,\theta_0}(\cdot)$ $(j = 1, \ldots, p)$, will be defined for each $\theta \in \Theta_n$, respectively, as:

$$\widehat{g}_{j,h,\theta}(\chi) = \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) X_{ij} \quad \text{and} \quad \widehat{g}^*_{j,k,\theta}(\chi) = \sum_{i=1}^{n} w^*_{n,k,\theta}(\chi, \mathcal{X}_i) X_{ij} \quad \forall \chi \in \mathcal{H}.$$

3.6.1 Proof of Theorem 3.2 (a)

The main idea of the proof consists in applying existing results for kernel estimates in the FSIM (see section 2.3.3) without additional multivariate predictors. Then, we will have to deal with the estimation of the linear coefficients $\boldsymbol{\beta}_0$.

For fixed $\chi \in \mathcal{H}$, the following decomposition can be made:

$$
\begin{aligned}
|\widehat{r}_{h,\theta,\boldsymbol{\beta}}(\chi) - r_{\theta_0}(\chi)| &= \left| \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) \left(Y_i - \boldsymbol{X}_i^\top \left(\boldsymbol{\beta} - \boldsymbol{\beta}_0 + \boldsymbol{\beta}_0 \right) \right) - r_{\theta_0}(\chi) \right| \\
&= \left| \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 \right) - r_{\theta_0}(\chi) \right. \\
&\quad \left. + \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) \boldsymbol{X}_i^\top \left(\boldsymbol{\beta}_0 - \boldsymbol{\beta} \right) \right| \\
&\leq \left| \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 \right) - r_{\theta_0}(\chi) \right| \\
&\quad + \left| \sum_{j=1}^{p} \left(\widehat{g}_{j,h,\theta}(\chi) + g_{j,\theta_0}(\chi) - g_{j,\theta_0}(\chi) \right) \left(\beta_{0j} - \beta_j \right) \right| \\
&\leq \left| \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) \left(r_{\theta_0}(\mathcal{X}_i) + \varepsilon_i \right) - r_{\theta_0}(\chi) \right| \\
&\quad + \left| \sum_{j=1}^{p} \left(\widehat{g}_{j,h,\theta}(\chi) - g_{j,\theta_0}(\chi) \right) \left(\beta_{0j} - \beta_j \right) \right| \\
&\quad + \left| \sum_{j=1}^{p} g_{j,\theta_0}(\chi) \left(\beta_{0j} - \beta_j \right) \right| . \qquad (3.14)
\end{aligned}
$$

Now, using Theorem 2.5 (a) it is obtained that

$$
\sup_{\theta \in \Theta_n} \sup_{a_n \leq h \leq b_n} \left| \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) \left(r_{\theta_0}(\mathcal{X}_i) + \varepsilon_i \right) - r_{\theta_0}(\chi) \right| = O\left(b_n^{\alpha_0} \right) + O_{\text{a.co.}} \left(\sqrt{\frac{\log n}{n f(a_n)}} \right) . \qquad (3.15)
$$

Using again Theorem 2.5 together with Condition (3.8),

$$
\begin{aligned}
&\left| \sum_{j=1}^{p} \left(\widehat{g}_{j,h,\theta}(\chi) - g_{j,\theta_0}(\chi) \right) \left(\beta_{0j} - \beta_j \right) \right| \\
&\leq p^{1/2} \max_{j=1,\ldots,p} \sup_{\theta \in \Theta_n} \sup_{a_n \leq h \leq b_n} \left| \widehat{g}_{j,h,\theta}(\chi) - g_{j,\theta_0}(\chi) \right| \left| \left| \boldsymbol{\beta} - \boldsymbol{\beta}_0 \right| \right| \\
&= O\left(b_n^{\alpha_0} c_n \right) + O_{\text{a.co.}} \left(c_n \sqrt{\frac{\log n}{n f(a_n)}} \right) . \qquad (3.16)
\end{aligned}
$$

In addition, from conditions (3.7) and (3.8) we get:

$$\max_{j=1,\ldots,p} |g_{j,\theta_0}(\chi)| \, ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| = O(c_n). \tag{3.17}$$

Finally, applying (3.15)–(3.17) in (3.14), and using that $c_n \to 0$ as $n \to \infty$ we obtain the claimed result:

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\Psi_n} \sup_{\theta\in\Theta_n} \sup_{a_n\leq h\leq b_n} |\widehat{r}_{h,\theta,\boldsymbol{\beta}}(\chi) - r_{\theta_0}(\chi)| &= O\left(b_n^{\alpha_0}\right) + O_{\text{a.co.}}\left(\sqrt{\frac{\log n}{nf(a_n)}}\right) \\
&\quad + O\left(b_n^{\alpha_0} c_n\right) + O_{\text{a.co.}}\left(c_n\sqrt{\frac{\log n}{nf(a_n)}}\right) + O(c_n) \\
&= O\left(b_n^{\alpha_0}\right) + O_{\text{a.co.}}\left(\sqrt{\frac{\log n}{nf(a_n)}}\right) + O(c_n). \ \blacksquare
\end{aligned}$$

3.6.2 Proof of Theorem 3.2 (b)

Following an analogous reasoning to (3.14), we can obtain

$$\begin{aligned}
\left|\widehat{r}^*_{k,\theta,\boldsymbol{\beta}}(\chi) - r_{\theta_0}(\chi)\right| &\leq \left|\sum_{i=1}^{n} w^*_{n,k,\theta}(\chi, \mathcal{X}_i)\left(r_{\theta_0}(\mathcal{X}_i) + \varepsilon_i\right) - r_{\theta_0}(\chi)\right| \\
&\quad + \left|\sum_{j=1}^{p} \left(\widehat{g}^*_{j,k,\theta}(\chi) - g_{j,\theta_0}(\chi)\right)(\beta_{0j} - \beta_j)\right| \\
&\quad + \left|\sum_{j=1}^{p} g_{j,\theta_0}(\chi)\,(\beta_{0j} - \beta_j)\right|.
\end{aligned} \tag{3.18}$$

Now, using Theorem 2.5, it is obtained that

$$\begin{aligned}
&\sup_{\theta\in\Theta_n} \sup_{k_{1,n}\leq h\leq k_{2,n}} \left|\sum_{i=1}^{n} w^*_{n,k,\theta}(\chi, \mathcal{X}_i)\left(r_{\theta_0}(\mathcal{X}_i) + \varepsilon_i\right) - r_{\theta_0}(\chi)\right| \\
&= O\left(f^{-1}\left(\frac{k_{2,n}}{\rho_n n}\right)^{\alpha_0}\right) + O_{a.co.}\left(\sqrt{\frac{\log n}{nf\left(\lambda f^{-1}(\rho_n k_{1,n}/n)\right)}}\right).
\end{aligned} \tag{3.19}$$

Using again Theorem 2.5 and Condition (3.8),

$$\left|\sum_{j=1}^{p}\left(\widehat{g}^{*}_{k,j,\theta}(\chi)-g_{j,\theta_0}(\chi)\right)(\beta_{0j}-\beta_j)\right| = O\left(c_n f^{-1}\left(\frac{k_{2,n}}{\rho_n n}\right)^{\alpha_0}\right) + O_{a.co.}\left(c_n\sqrt{\frac{\log n}{nf\left(\lambda f^{-1}(\rho_n k_{1,n}/n)\right)}}\right). \tag{3.20}$$

The desired result is obtained by combining (3.19), (3.20) and (3.17) with (3.18), and using that $c_n \to 0$ as $n \to \infty$. ■

Chapter 4

Contributions on the sparse semi-functional partial linear single-index model

4.1 Introduction

Today, most of multivariate data analysis methodologies have been adapted to functional data, as it has been evidenced by several recent surveys on FDA (see e.g. Cuevas [29], Goia and Vieu [55] and Aneiros et al. [8]). Due to the infinite-dimensionality of random variables in FDA, one of the main issues to ensure the good performance of any statistical procedure is to control, in one way or another, the dimensionality of the model (see the Introduction of this dissertation). In fact, this dimensionality challenge is not so far from what exists in the related field of Big Data analysis, in which traditionally the statistical variable is a high-dimensional vector. In the recent past, the necessary links between the two fields have been highlighted by both the Big Data community (see e.g. Scott [100]) and the FDA one (see e.g. Goia and Vieu [55] and Aneiros et al. [8]). Moreover, in many fields of applications, one could find data consisting of mixtures of functional and high-dimensional variables. Then, the statistical methodologies to be built have to cross both fields of FDA and Big Data.

This chapter is part of this category, since our purpose is to develop a new model

for regression problems involving some scalar response, Y, and predictors composed of some functional variable, $\mathcal{X}$, and some high-dimensional vector, $(X_1, \ldots, X_{p_n})$. This new model must take into account three important features of our problem: i) firstly, additive ideas are needed to separate the effects of the functional predictor, $\mathcal{X}$, from those of the multivariate one, $(X_1, \ldots, X_{p_n})$; ii) secondly, sparse ideas are needed to control the high number of variables, p_n (which is allowed to go to infinity as n does), involved in the multivariate predictor; iii) finally, functional semiparametric ideas are required to model the effect of the infinite-dimensional predictor, $\mathcal{X}$. These features lead us to the SSFPLSIM, which has been briefly presented in Section 1.4.3. Specifically, the SSFPLSIM is given by the expression

$$Y = X_1\beta_{01} + \cdots + X_{p_n}\beta_{0p_n} + r\left(\langle\theta_0, \mathcal{X}\rangle\right) + \varepsilon, \tag{4.1}$$

where Y denotes a scalar response, $X_1, \ldots, X_{p_n}$ are random covariates taking values in $\mathbb{R}$ and $\mathcal{X}$ is a functional random covariate valued in a separable Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle$. In this equation, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^\top \in \mathbb{R}^{p_n}$, $\theta_0 \in \mathcal{H}$ and $r(\cdot)$ are a vector of unknown real parameters, an unknown functional direction and an unknown smooth real-valued function, respectively. In addition, ε is the random error, which verifies

$$\mathbb{E}\left(\varepsilon | X_1, \ldots, X_{p_n}, \mathcal{X}\right) = 0. \tag{4.2}$$

Finally, we will consider the same conditions (2.2) and (2.3) presented in the FSIM (2.1) to ensure identifiability.

Model (4.1) is a generalization of the SFPLSIM (3.1) studied in the previous chapter focusing on the estimation of the semiparametric component. The difference between SFPLSIM and model (4.1) is that the latter incorporates the possibility of having a divergent number of linear parameters and sparseness in the linear component. Accordingly, in this chapter we put special attention on the linear component. A variable selection method and estimators of the components of the model will be constructed along Section 4.2, while a wide set of asymptotics will be provided in Section 4.3. Finite sample behaviour of the method will be assessed through Monte Carlo experiments in Section 4.4. In addition, Section 4.5 provides an application to Tecator's data. Technical proofs and lemmas are gathered in the Section 4.6.

4.2 The penalized least-squares estimators

For simultaneously estimating the linear β-parameters and selecting the relevant X-covariates in the SSFPLSIM (4.1) we will use a *penalized least-squares* (PLS) approach. That is, assume that we have a statistical sample of n vectors

$$\{(X_{i1}, \ldots, X_{ip_n}, \mathcal{X}_i, Y_i)\}_{i=1}^n,$$

i.i.d. as $(X_1, \ldots, X_{p_n}, \mathcal{X}, Y)$ verifying the SSFPLSIM (4.1):

$$Y_i = X_{i1}\beta_{01} + \cdots + X_{ip_n}\beta_{0p_n} + r(\langle \theta_0, \mathcal{X}_i \rangle) + \varepsilon_i \quad (i = 1, \ldots, n).$$

Step 1 The first idea is to transform the SSFPLSIM (4.1) into a linear model by extracting from Y_i and X_{ij} $(j = 1, \ldots, p_n)$ the effect of the functional covariate $\mathcal{X}_i$ when is projected on the direction θ_0. Specifically, if we denote $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip_n})^\top$ $(i = 1, \ldots, n)$, the fact that

$$Y_i - \mathbb{E}\left(Y_i | \langle \theta_0, \mathcal{X}_i \rangle\right) = \left(\boldsymbol{X}_i - \mathbb{E}\left(\boldsymbol{X}_i | \langle \theta_0, \mathcal{X}_i \rangle\right)\right)^\top \boldsymbol{\beta}_0 + \varepsilon_i \ (i = 1, \ldots, n) \tag{4.3}$$

allows to consider the following approximate linear model:

$$\widetilde{\boldsymbol{Y}}_{\theta_0} \approx \widetilde{\boldsymbol{X}}_{\theta_0}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \tag{4.4}$$

where, as in Chapters 2 and/or 3, the following notations were used:

- $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$.
- For any $(n \times q)$-matrix $\boldsymbol{A}$ $(q \geq 1)$ and $\theta \in \mathcal{H}$, we denote

  $$\widetilde{\boldsymbol{A}}_\theta = (\boldsymbol{I} - \boldsymbol{W}_{h,\theta})\,\boldsymbol{A}, \text{ where } \boldsymbol{W}_{h,\theta} = \left(w_{n,h,\theta}(\mathcal{X}_i, \mathcal{X}_j)\right)_{i,j}$$

  with $w_{n,h,\theta}(\cdot, \cdot)$ being the weight function defined in (2.11).
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$.

Note that, to obtain (4.4), the conditional expectations in (4.3) were estimated by means of functional nonparametric techniques (kernel-based procedures).

Step 2 The penalized least-squares approach is applied to model (4.4). In this way, the considered penalized profile least-squares function is defined as

$$\mathcal{Q}(\boldsymbol{\beta},\theta)=\frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_{\theta}-\widetilde{\boldsymbol{X}}_{\theta}\boldsymbol{\beta}\right)^{\top}\left(\widetilde{\boldsymbol{Y}}_{\theta}-\widetilde{\boldsymbol{X}}_{\theta}\boldsymbol{\beta}\right)+n\sum_{j=1}^{p_n}\mathcal{P}_{\lambda_{j_n}}\left(|\beta_j|\right), \tag{4.5}$$

where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_{p_n})^{\top}$, $\mathcal{P}_{\lambda_{j_n}}(\cdot)$ is a penalty function and $\lambda_{j_n}>0$ is a tuning parameter (as commented in Section 1.4.1, the role of the sum in (4.5) is to penalize the presence of non zero β-parameters; in fact, under suitable conditions on $\mathcal{P}_{\lambda}$ (see e.g. Fan and Li [38]), the penalized least-squares estimators produce sparse solutions (many estimated coefficients are zero)). At this moment it is noteworthy, on the one hand, the fact that the objective function $\mathcal{Q}$ in (4.5) is not necessarily convex. This is the reason why, as usual in the related literature (see e.g. Fan and Li [38], Fan and Peng [41] and Wang and Zhu [109]), our asymptotic results in next section are focused on a local minimizer, $(\widehat{\boldsymbol{\beta}}_0,\widehat{\theta}_0)$, of $\mathcal{Q}$ (in particular, existence of such local minimizer will be established in Theorem 4.2). On the other hand, this parameter estimation procedure can be used also as a variable selection method in a simple way: if $\widehat{\beta}_{0j}$ is a non-null component of $\widehat{\boldsymbol{\beta}}_0$, then X_j is selected as an influential variable.

Step 3 Finally, after estimating $\boldsymbol{\beta}_0$ and θ_0, we can deal with the estimation of the nonlinear function $r_{\theta_0}(\cdot)\equiv r\left(\langle\theta_0,\cdot\rangle\right)$ in (4.1). A natural way is employing again nonparametric procedures and smoothing (using kernel-based estimators) the partial residuals $Y_i-\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}_0$; that is,

$$\widehat{r}_{\theta}(\chi)=\sum_{i=1}^{n}w_{n,h,\theta}(\chi,\mathcal{X}_i)\left(Y_i-\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}_0\right); \tag{4.6}$$

then, the estimator of $r_{\theta_0}(\chi)$ will be $\widehat{r}_{\widehat{\theta}_0}(\chi)$. Note that the same bandwidth, h, is used to estimate both the functional index θ_0 and the parameter $\boldsymbol{\beta}_0$ from (4.5), as well as to estimate the smooth real-valued function $r(\cdot)$ from (4.6). Although the partial residuals in (4.6) could be smoothed by considering a different bandwidth, we have adopted the more usual procedure of using the same bandwidth twice (see e.g. Liang et al. [75] for the case of a non-functional par-

tial linear single-index model). Extension to the case with different bandwidths does not involve any additional difficulties.

4.3 Asymptotic theory

4.3.1 Some initial notation

Let us first introduce some notation to be used in the results of this chapter, as well as in their proofs:

- The set of indices of the covariates with linear effect will be denoted as J_n, that is $J_n = \{1, \ldots, p_n\}$, and the set of indices corresponding to the influential variables will be referred as S_n, that is $S_n = \{j \in J_n;\ \beta_{0j} \neq 0\}$ (the complement of S_n will be denoted as $\overline{S}_n$). In addition, s_n will mean $\text{card}(S_n)$.

- Given any vector $\boldsymbol{v} \in \mathbb{R}^{p_n}$ and any $p_n \times p_n$ matrix $\boldsymbol{M}$, $\boldsymbol{v}_{S_n}$ and $\boldsymbol{M}_{S_n \times S_n}$ denote the vector and the matrix obtained from $\boldsymbol{v}$ and $\boldsymbol{M}$ retaining only the components corresponding to the index sets S_n and $S_n \times S_n$, respectively.

- For any $\theta \in \mathcal{H}$, and for $1 \leq i \leq n$, $1 \leq j \leq p_n$, we denote

$$g_{0,\theta}(\mathcal{X}_i) = \mathbb{E}\left(Y_i | \langle \theta, \mathcal{X}_i \rangle\right),$$

and

$$g_{j,\theta}(\mathcal{X}_i) = \mathbb{E}\left(X_{ij} | \langle \theta, \mathcal{X}_i \rangle\right),$$

that is, $g_{0,\theta}(\cdot)$ and $g_{j,\theta}(\cdot)$ are the functional single-index regression operators of Y_i over $\mathcal{X}_i$, and of X_{ij} $(j = 1, \ldots, p_n)$ over $\mathcal{X}_i$, respectively. Finally, the errors of these regressions are denoted by η_{ij,θ_0} with $1 \leq i \leq n$, $1 \leq j \leq p_n$:

$$\eta_{ij,\theta_0} = X_{ij} - g_{j,\theta_0}(\mathcal{X}_i),$$

and

$$\boldsymbol{\eta}_{i,\theta_0} = \left(\eta_{i1,\theta_0}, \ldots, \eta_{ip_n,\theta_0}\right)^\top.$$

- $\Delta_{min}(\boldsymbol{M})$ and $\Delta_{max}(\boldsymbol{M})$ denote the smallest and the largest eigenvalues of the matrix $\boldsymbol{M}$, respectively.

- The symbol $||\cdot||$ is used for denoting the L_2 norm of vectors and matrices. The same symbol is also employed for denoting the norm induced by the inner product $\langle\cdot,\cdot\rangle$. Specifically:

$$||\boldsymbol{a}|| = \left(a_1^2 + \cdots + a_q^2\right)^{1/2} \text{ for } \boldsymbol{a} = (a_1, \ldots, a_q)^\top \in \mathbb{R}^q,$$
$$||\boldsymbol{A}|| = \max_{0 \neq \boldsymbol{x} \in \mathbb{R}^q} \frac{||\boldsymbol{A}\boldsymbol{x}||}{||\boldsymbol{x}||} \text{ for any } r \times q \text{ matrix } \boldsymbol{A}$$
and
$$||\chi|| = \langle \chi, \chi \rangle^{1/2} \text{ for any } \chi \in \mathcal{H}.$$

- $\forall\ \chi, \theta \in \mathcal{H}$ and $\forall \epsilon > 0$, we will use the notation:

$$B(\theta, \epsilon) = \{\theta' \in \mathcal{H};\ d(\theta, \theta') < \epsilon\},$$

where, $\forall\ \chi, \chi' \in \mathcal{H}$, $d(\chi, \chi') = ||\chi - \chi'||$.

4.3.2 Assumptions

In order to state the rates of convergence of the proposed estimators and the model selection consistency, we will use a large number of assumptions (some of them very technical). Such number is directly related to the complexity of the model and the results to be obtained. These assumptions, which will be justified in next Remark 4.1, are the following:

Conditions on the set of values of $\mathcal{X}$ and the topologies induced by $d_\theta(\cdot,\cdot)$.
The functional variable $\mathcal{X}$ is valued in some subset $\mathcal{C}$ of $\mathcal{H}$ such that can be covered in the following way

$$\mathcal{C} \subset \bigcup_{k=1}^{N_{\mathcal{C},\epsilon}^{\theta}} B_\theta(\chi_{\epsilon,k}^{\theta}, \epsilon),\ \forall \theta \in \Theta_n. \tag{4.7}$$

In (4.7), $N^{\theta}_{\mathcal{C},\epsilon}$ is the minimal number of open balls in $(\mathcal{H}, d_\theta(\cdot,\cdot))$ of radius ϵ which are necessary to cover $\mathcal{C}$ (note that the number $N^{\theta}_{\mathcal{C},\epsilon}$ and the centres of the balls, $\chi^{\theta}_{\epsilon,k}$, depend on θ and ϵ). In addition,

$$\Theta_n = \{\theta \in \mathcal{H};\ d(\theta,\theta_0) \leq v_n\} \text{ with } v_n \to 0 \text{ as } n \to \infty. \tag{4.8}$$

That is, directions belonging to Θ_n are relatively close to the target direction (and much closer as $n \to \infty$).

Conditions on the entropies and the balls in (4.7). Let us denote

$$N_{\mathcal{C},\epsilon} = \sup_{\theta\in\Theta_n} N^{\theta}_{\mathcal{C},\epsilon}, \quad \psi_{\mathcal{C}}(\epsilon) = \log(N_{\mathcal{C},\epsilon}), \quad k^0_{(\theta,k,\epsilon)} = \arg \min_{k' \in \left\{1,\ldots,N^{\theta_0}_{\mathcal{C},\epsilon}\right\}} d(\chi^{\theta}_{\epsilon,k}, \chi^{\theta_0}_{\epsilon,k'}). \tag{4.9}$$

That is, $\psi_{\mathcal{C}}(\epsilon)$ denotes the Kolmogorov ϵ-entropy of the subset $\mathcal{C}$ in the topology induced by $d_{\theta^*}(\cdot,\cdot)$, where $\theta^* = \arg\sup_{\theta\in\Theta_n} \mathcal{N}^{\theta}_{\mathcal{C},\epsilon}$ (for details on the Kolmogorov ϵ-entropy, see last item in Section 2.3.1). In addition, $k^0_{(\theta,k,\epsilon)}$ is the subscript associated to the centre of the θ_0-ball with radius ϵ ($\chi^{\theta_0}_{\epsilon,k}$) which minimizes the "distance"(measured with the semi-metric induced by the inner product) to the centre of the k^{th} θ-ball with the same radius ($\chi^{\theta}_{\epsilon,k}$). In the sake of brevity, for the particular case $\epsilon = 1/n$, we will use the notation

$$\chi^{\theta}_k = \chi^{\theta}_{1/n,k} \text{ and } k^0 = k^0_{(\theta,k,1/n)}.$$

On the one hand, it is assumed that

$$\exists \beta > 1 \text{ such that } p_n \exp\left\{(1-\beta\log p_n)\,\psi_{\mathcal{C}}\left(\frac{1}{n}\right)\right\} \to 0 \text{ as } n \to \infty. \tag{4.10}$$

On the other hand, we assume that the maximum "distance" (measured with the semi-metric associated to the inner product) between the centre of the k^{th} θ-ball with radius $1/n$ and the nearest θ_0-ball of the same radius verifies

$$\sup_{\theta\in\Theta_n} \max_{k\in\{1,\ldots,N^{\theta}_{\mathcal{C},1/n}\}} d(\chi^{\theta}_k, \chi^{\theta_0}_{k^0}) = O(1/n). \tag{4.11}$$

Conditions on the small-ball probabilities. There exist constants $C_1 > 0$, $0 < C_2 \leq C_3 < \infty$ and a function $f : \mathbb{R} \longrightarrow (0, \infty)$ such that

$$\int_0^1 f\left(hs\right) ds > C_1 f\left(h\right), \tag{4.12}$$

and

$$\forall \chi \in \mathcal{C} \text{ and } \forall \theta \in \Theta_n, \quad C_2 f(h) \leq \phi_{\chi,\theta}(h) \leq C_3 f(h) \tag{4.13}$$

(that is, the function $f(\cdot)$ bounds small-ball probability functions, avoiding dependence on χ and θ).

Conditions linking the entropies and the small-ball probabilities. In order to control the entropy of the set $\mathcal{C}$, it is assumed that there exists a constant $C_4 > 0$ such that, for n large enough,

$$\psi_{\mathcal{C}}\left(\frac{1}{n}\right) \leq \frac{C_4 n f(h)}{\alpha_n \log p_n}, \text{ where } \alpha_n \to \infty \text{ as } n \to \infty. \tag{4.14}$$

Conditions on the kernel K.

For all $u \in [0,1)$, $K(u) \neq 0$ and for all $u \in (-\infty, 0) \cup (1, +\infty)$, $K(u) = 0$.
In addition, K is Lipschitz continuous on $[0,1)$,
and, if $K(1) = 0$, K also satisfies for all $u \in [0,1)$ that
$-\infty < C_5 < K'(u) < C_6 < 0$, where C_5 and C_6 denote constants. (4.15)

Note that from a theoretical point of view, we have to differentiate the case where K is a continuous kernel ($K(1) = 0$) from the case where $K(\cdot)$ is not continuous. The case $K(1) = 0$ is more delicate and additional assumptions are needed.

Conditions on the smoothness. Hölder type conditions are assumed for involved functional single-index regressions; that is, for some constants $0 \leq C_7 < \infty$ and $\alpha > 0$, $\forall (\chi, \chi') \in \mathcal{C} \times \mathcal{C}$, and $\forall z \in \{g_{0,\theta_0}, g_{1,\theta_0}, \ldots, g_{p_n,\theta_0}\}$, it is verified that

$$|z(\chi) - z(\chi')| \leq C_7 d_{\theta_0}(\chi, \chi')^{\alpha}. \tag{4.16}$$

Conditions on the random variables.

$$\{(Y_i, X_{i1}, \ldots, X_{ip_n}, \mathcal{X}_i)\} \text{ are random vectors i.i.d. verifying model (4.1).} \tag{4.17}$$

$$\{\boldsymbol{\eta}_{i,\theta_0}\} \text{ and } \{\varepsilon_i\} \text{ are independents.} \tag{4.18}$$

$$\langle \mathcal{X}, \mathcal{X} \rangle^{1/2} < C_8, \text{where } C_8 \text{ denotes a positive constant.} \tag{4.19}$$

Conditions on the moments. Let C_9, $C_{\eta_{\theta_0}}$ and C_{m_ε} be positive constants. On the one hand, conditional moments of the involved regressions are bounded, in the sense that, $\forall m \geq 2$, there exists a continuous operator in $\mathcal{C}$, $\sigma_m(\cdot)$, such that $\forall \chi \in \mathcal{C}$,

$$\begin{aligned} \max_{j \in \{1,\ldots,p_n\}} \left\{ \mathbb{E}\left(|Y_1|^m \left| \langle \theta_0, \mathcal{X}_1 \rangle = \langle \theta_0, \chi \rangle \right. \right), \mathbb{E}\left(|X_{1j}|^m \left| \langle \theta_0, \mathcal{X}_1 \rangle = \langle \theta_0, \chi \rangle \right. \right) \right\} &< \sigma_m(\chi) \\ &< C_9. \end{aligned} \tag{4.20}$$

On the other hand, the errors of the regressions also verify some moment conditions:

$$\forall\ m \geq 2 \text{ and } \forall 1 \leq j \leq p_n,\ \mathbb{E}|\eta_{1j,\theta_0}|^m \leq C_{\eta_{\theta_0}} \left(\frac{m!}{2} \right). \tag{4.21}$$

$$\exists m_\varepsilon > 4 \text{ such that } \mathbb{E}\left|\varepsilon_1\right|^{m_\varepsilon} \leq C_{m_\varepsilon}. \tag{4.22}$$

In addition, there exists a constant C_{10} such that

$$0 < C_{10} < \Delta_{min}\left(\boldsymbol{B}_{\theta_0 S_n \times S_n} \right), \tag{4.23}$$

where $\boldsymbol{B}_{\theta_0} = \mathbb{E}\left(\boldsymbol{\eta}_{1\theta_0} \boldsymbol{\eta}_{1\theta_0}^\top \right)$. In particular, $\boldsymbol{B}_{\theta_0 S_n \times S_n}$ is a definite positive matrix.

Conditions on the non null parameters and the penalty functions. Let C_{11} and C_{12} be positive constants. The penalty function verifies the following con-

ditions:

$$\mathcal{P}_{\lambda_{jn}}(\cdot) \text{ is a continuous and nonnegative function verifying } \mathcal{P}_{\lambda_{jn}}(0) = 0, \tag{4.24}$$

$$\mathcal{P}_{\lambda_{jn}}(\cdot) \text{ is differentiable excepted perhaps at 0,} \tag{4.25}$$

with second derivative verifying Lipschitz continuity

$$\left|\mathcal{P}''_{\lambda_{jn}}(a) - \mathcal{P}''_{\lambda_{jn}}(b)\right| \leq C_{11}|a-b|, \ \forall a, b > C_{12}\lambda_{jn}, \tag{4.26}$$

and the first derivative verifying

$$\lim \inf_{n\to\infty} \min_{j\in\overline{S}_n} \left\{ \lim \inf_{d\to 0^+} \frac{\mathcal{P}'_{\lambda_{jn}}(d)}{\lambda_{jn}} \right\} > 0. \tag{4.27}$$

Finally, the non null parameters verify

$$\min_{j\in S_n} \left\{ \frac{|\beta_{0j}|}{\lambda_{jn}} \right\} \to \infty \text{ as } n \to \infty, \tag{4.28}$$

which explicitly shows the rate at which the PLS approach can distinguish nonvanishing parameters from 0. In addition, it is also assumed that linear coefficients are bounded

$$\max_{j\in S_n} \{|\beta_{0j}|\} = O(1). \tag{4.29}$$

Remark 4.1 *The hypotheses listed above are, in general, usual (or natural extensions of those) in the related literature. For instance, conditions (4.7), (4.10) and (4.14) are linked with the topology of* $(\mathcal{C}, d_\theta)$ *and, in the particular case of known* θ_0*, they are common when it is needed to obtain uniform orders over* $\mathcal{C}$ *(see e.g. Ferraty et al. [49] or Aneiros et al. [7]). In the general case dealt here, where* θ_0 *is unknown and one needs to control the behaviour of the profile function* $\mathcal{Q}(\cdot,\cdot)$ *(4.5) around* θ_0*, conditions (4.7), (4.10) and (4.14) are the natural extension of the corresponding to such particular case. In the same way, conditions (4.8), (4.12) and (4.13) also allow to control the effect of* θ*. Specifically, Condition (4.8) establishes the set of values of* θ *where the profile function* $\mathcal{Q}(\cdot,\cdot)$ *achieves a local minimum (see Ma [81]), while*

conditions (4.12) and (4.13) are natural extensions of usual assumptions (related to the concentration properties of the probability measure of the functional variable $\mathcal{X}$) from the case of known θ_0 (see e.g. Ferraty et al. [49]) to that of unknown θ_0 (see e.g. Chapter 2 or Ait-Saïdi et al. [3]). In addition, conditions (4.15)-(4.23) are standard ones in nonparametric and semiparametric regression estimation when functional covariates are present (see e.g. Ferraty et al. [49], Aneiros et al. [7], Wang et al. [110]). Basically, they are mild conditions on the kernel, on the smoothness of the nonparametric components involved (related to both the response variable and the scalar covariates), on the dependence within the model and on the moments of the variables. Focusing now on the conditions directly linked to the penalty procedure (conditions (4.24)-(4.29)), they are usual assumptions in the topic of variable selection using nonconcave penalized functions (see e.g. Fan and Li [38], Fan and Peng [41], Aneiros et al. [7]). Note that a main role of these conditions is to produce sparse solutions, i.e., automatically to set to zero small estimated coefficients to reduce model complexity. In addition, under some specific condition (see, for instance, Aneiros et al. [7]), the SCAD penalty function (1.10) verifies our assumptions. Finally, Condition (4.11) is really specific to the functional framework addressed here and, therefore, requires a deeper reasoning. It will be discussed in a more general setting in Section 4.6.7 (see Remark 4.11).

4.3.3 Results

Our first result focuses on both the existence and rate of convergence of a local minimizer of the penalized least-squares objective function $\mathcal{Q}(\boldsymbol{\beta}, \theta)$ (see (4.5)). Let us denote

$$\delta_n = \max_{j \in S_n} \left\{ \left| \mathcal{P}'_{\lambda_{j_n}} (|\beta_{0j}|) \right| \right\}, \quad \rho_n = \max_{j \in S_n} \left\{ \left| \mathcal{P}''_{\lambda_{j_n}} (|\beta_{0j}|) \right| \right\} \quad \text{and} \quad u_n = \sqrt{s_n} \left(n^{-1/2} + \delta_n \right). \tag{4.30}$$

Theorem 4.2 *Assume that conditions (4.2), (4.7), (4.8) and (4.10)-(4.29) hold. Assume, in addition, that* $p_n \to \infty$ *as* $n \to \infty$, $p_n = o\left(n^{1/2}\right)$ *and*

$$\max\left\{ns_n^2 h^{4\alpha}, s_n h^{\alpha}\log n\right\} = O(1),$$
$$s_n^2 \log p_n \log^2 n = O\left(\frac{nf(h)}{\psi_{\mathcal{C}}(1/n)}\right),$$
$$s_n^2 \log^2 n = O\left(n\left(\frac{f(h)}{\psi_{\mathcal{C}}(1/n)}\right)^2\right),$$
$$ns_n v_n = O(hf(h))$$
and
$$\max\left\{\rho_n, \frac{u_n}{\min_{j\in S_n}\{\lambda_{jn}\}}, \frac{u_n \Delta_{max}^{1/2}(\boldsymbol{B}_{\theta_0})}{\min_{j\in \overline{S}_n}\{\lambda_{jn}\}}, \frac{n^{-1/2+1/m_\epsilon}\log n}{\min_{j\in \overline{S}_n}\{\lambda_{jn}\}}\right\} = o(1).$$

Then, there exists a local minimizer $\left(\widehat{\boldsymbol{\beta}}_0, \widehat{\theta}_0\right)$ *of* $\mathcal{Q}(\boldsymbol{\beta}, \theta)$ *such that*

$$\left\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right\| = O_p(u_n) \text{ and } \left\|\widehat{\theta}_0 - \theta_0\right\| = O_p(v_n)$$

(Note that v_n *was defined in (4.8)).*

Remark 4.3 *Theorem 4.2 can be seen, in a certain sense, as an extension of Theorem 3.1 in Aneiros et al. [7] from the case* $\Theta_n = \{\theta_0\}$ *(i.e.,* $v_n = 0$ *in (4.8); equivalently,* θ_0 *known) to the case where* $\{\theta_0\} \subset \Theta_n$ *(i.e.,* θ_0 *unknown). For verifying that, we only need to consider the results in Aneiros et al. [7] when the semi-metric* $d_{\theta_0}(\cdot,\cdot)$ *is used. From Theorem 4.2 we obtain that the rate of convergence* (u_n) *achieved by the local minimizer* $\widehat{\boldsymbol{\beta}}_0$ *is the same as that reached in the least complex scenario studied in Aneiros et al. [7] (as well as in the linear model considered in Fan and Lv [40], where* $\delta_n = 0$*), and this was one of our main aims. Naturally, for this to be possible, it is necessary to have a very good estimator of the parameter* θ_0*. Such estimator is obtained by means of the local minimizer* $\widehat{\theta}_0$ *(note that the local feature of the minimizers plays a main role to obtain fast rates of convergence).*

Our second result states the model selection consistency. Let us denote

$$\widehat{S}_n = \left\{j \in J_n; \widehat{\beta}_{0j} \neq 0\right\},$$

where $\widehat{\boldsymbol{\beta}}_0 = (\widehat{\beta}_{01}, \ldots, \widehat{\beta}_{0p_n})^\top$ is the estimator in Theorem 4.2.

Theorem 4.4 (Model selection consistency) *Under assumptions in Theorem 4.2, we have that*

$$\mathbb{P}\left(\widehat{S}_n = S_n\right) \to 1 \text{ as } n \to \infty.$$

The following result focuses on the asymptotic distribution of certain projections of $\widehat{\boldsymbol{\beta}}_0$. First, let us introduce some additional notation. We will denote

$$\boldsymbol{c} = (c_1, \ldots, c_{p_n})^\top, \text{ being } c_j = \mathcal{P}'_{\lambda_{jn}}(|\beta_{0j}|)\operatorname{sgn}(\beta_{0j}) 1_{\{j \in S_n\}},$$

and

$$\boldsymbol{V} = \operatorname{diag}\{V_1, \ldots, V_{p_n}\}, \text{ where } V_j = \mathcal{P}''_{\lambda_{jn}}(|\beta_{0j}|)\, 1_{\{j \in S_n\}}.$$

In addition, we will denote $\sigma_\varepsilon^2 = \mathbb{E}(\varepsilon_i^2)$, while $\boldsymbol{A}_n$ will be any $q \times s_n$ matrix such that $\boldsymbol{A}_n \boldsymbol{A}_n^\top \to \boldsymbol{A}$ as $n \to \infty$, where $\boldsymbol{A}$ is a $q \times q$ definite positive matrix.

Theorem 4.5 (Asymptotic normality) *Adding the following conditions to assumptions in Theorem 4.2 (where, if $s_n = 1$, $\log s_n$ must be interpreted as 1):*

$$\exists \beta' > 1 \text{ such that } s_n \exp\left\{(1 - \beta' \log s_n)\, \psi_{\mathcal{C}}\left(\frac{1}{n}\right)\right\} \to 0 \text{ as } n \to \infty,$$

$$\max\left\{n s_n^3 h^{4\alpha}, n^{2/m_\varepsilon} s_n h^{2\alpha} \log^2 n, s_n^3 h^{2\alpha} \log^2 n, n^{-1} s_n^3, n s_n^3 \delta_n^4\right\} = o(1)$$

and

$$\max\left\{n^{2/m_\varepsilon} s_n \log s_n \log^2 n, s_n^3 \log s_n \log^2 n\right\} = o\left(\frac{n f(h)}{\psi_{\mathcal{C}}(1/n)}\right),$$

the following result can be established:

$$n^{1/2} \boldsymbol{A}_n \boldsymbol{C}_{\theta_0, S_n}\left(\widehat{\boldsymbol{\beta}}_{0S_n} - \boldsymbol{\beta}_{0S_n} + (\boldsymbol{B}_{\theta_0 S_n \times S_n} + \boldsymbol{V}_{S_n \times S_n})^{-1} \boldsymbol{c}_{S_n}\right) \xrightarrow{\mathrm{d}} N(\boldsymbol{0}, \boldsymbol{A}),$$

where we have denoted $\boldsymbol{C}_{\theta_0, S_n} = \sigma_\varepsilon^{-1} \boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2} (\boldsymbol{B}_{\theta_0 S_n \times S_n} + \boldsymbol{V}_{S_n \times S_n})$.

Remark 4.6 *Theorems 4.4 and 4.5 show that $\widehat{\boldsymbol{\beta}}_0$ enjoys the oracle property with the meaning given, for instance, in Xie and Huang [116]: "the estimator can correctly*

select the nonzero coefficients with probability converging to one, and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariances that they would have if the zero coefficients were known in advance". In the setting of multivariate regression (not functional), the interested reader can find estimators verifying such property in Fan and Peng [41] (linear regression), Xie and Huang [116] (partially linear regression) or Wang and Zhu [109] (partial linear single-index regression), among others. See also Aneiros et al. [7] for the case of the semi-functional partial linear regression.

Finally, the next theorem states the uniform rate of convergence of the statistic $\widehat{r}_\theta(\chi)$ in (4.6).

Theorem 4.7 *Under assumptions of Theorem 4.2, if in addition the following conditions are verified:*

A) $\forall(\chi,\chi') \in \mathcal{C}\times\mathcal{C}$, $|r_{\theta_0}(\chi) - r_{\theta_0}(\chi')| \leq C_{13} d_{\theta_0}(\chi,\chi')^\alpha$, *where* α *was defined in (4.16),*

B) $\sup_{\chi\in\mathcal{C}, j\in S_n} |g_{j,\theta_0}(\chi)| = O(1)$

and

C) $\psi_{\mathcal{C}}(1/n) \to \infty$ *as* $n\to\infty$,

then, we have that

$$\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}} |\widehat{r}_\theta(\chi) - r_{\theta_0}(\chi)| = O_p\left(h^\alpha + \sqrt{\frac{\psi_{\mathcal{C}}(1/n)}{nf(h)}}\right) + O_p\left(\sqrt{s_n}u_n\right).$$

Corollary 4.8 *Under assumptions of Theorem 4.7, it is verified that*

$$\sup_{\chi\in\mathcal{C}} \left|\widehat{r}_{\widehat{\theta}_0}(\chi) - r_{\theta_0}(\chi)\right| = O_p\left(h^\alpha + \sqrt{\frac{\psi_{\mathcal{C}}(1/n)}{nf(h)}}\right) + O_p\left(\sqrt{s_n}u_n\right).$$

Corollary 4.9 *Under assumptions of Theorem 4.7, if in addition the following conditions hold:*

A) $\forall \theta \in \Theta_n$, *the random variables* $\langle \theta, \mathcal{X} \rangle$ *are valued in the same compact subset,* $\mathcal{R}$, *of* $\mathbb{R}$, *and are absolutely continuous with respect to the Lebesgue measure, with density* f_θ *satisfying*

$$0 < \inf_{\theta \in \Theta_n,\ u \in \mathcal{R}} f_\theta(u) \leq \sup_{\theta \in \Theta_n,\ u \in \mathcal{R}} f_\theta(u) < \infty,$$

B) $h \approx C(\log n/n)^{1/(2\alpha+1)}$,

C) $s_n \approx cn^\gamma$ *with* $0 < 2\gamma < 1 - 2\alpha/(2\alpha+1)$

and

D) $\delta_n = O(n^{-1/2})$ *(*δ_n *was defined in (4.30)),*

then we have that

$$\sup_{\theta \in \Theta_n} \sup_{\chi \in \mathcal{C}} |\widehat{r}_\theta(\chi) - r_{\theta_0}(\chi)| = O_p\left(\left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)}\right)$$

and

$$\sup_{\chi \in \mathcal{C}} \left|\widehat{r}_{\widehat{\theta}_0}(\chi) - r_{\theta_0}(\chi)\right| = O_p\left(\left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)}\right).$$

Remark 4.10 *Theorem 4.7 extends, in the sense commented in Remark 4.3, Theorem 3.3 in Aneiros et al. [7] from the case* $\Theta_n = \{\theta_0\}$ *(i.e.,* $v_n = 0$ *in (4.8); equivalently,* θ_0 *known) to the case where* $\{\theta_0\} \subset \Theta_n$ *(i.e.,* θ_0 *unknown). Corollary 4.9 shows a nice property of dimensionality reduction: the semi-functional nonparametric component* $r_{\theta_0}(\cdot) \equiv r(\langle \theta_0, \cdot \rangle)$ *is estimated with univariate nonparametric rate (note that Condition A imposed in Corollary 4.9 was used, e.g., in Ferraty et al. [50], while Condition D is satisfied, for instance, for the SCAD penalty function; finally, Condition B considers a bandwidth with optimal rate for univariate nonparametric regression while Condition C is a non-restrictive technical assumption).*

4.4 Simulation study

The aim of this section is to show the finite sample behaviour of the two statistical procedures presented before for the SSFPLSIM (4.1); that is, firstly, the penal-

ized least-squares procedure (for both variable selection and estimation of the linear parameters $\boldsymbol{\beta}_0$) and secondly, the single-index approach for estimating the functional semiparametric component of the model.

4.4.1 The design

For $(n, p_n) \in \{(100, 50), (200, 100)\}$, samples of i.i.d. data $\{(X_{i1}, \ldots, X_{ip_n}, \mathcal{X}_i, Y_i)\}_{i=1}^n$ were constructed according to the following model:

$$Y_i = X_{i1}\beta_{01} + \cdots + X_{ip_n}\beta_{0p_n} + r\left(\langle \theta_0, \mathcal{X}_i \rangle\right) + \varepsilon_i, \quad (i = 1, \ldots, n), \tag{4.31}$$

where:

- The vectors of real covariates, $(X_{i1}, \ldots, X_{ip_n})^\top$ $(i = 1, \ldots, n)$, were generated from a multivariate normal distribution with zero mean and covariance matrix given by $(\rho^{|j-k|})_{jk}$ $(j, k = 1, \ldots, p_n)$. Two values for ρ (namely $\rho = 0$ and $\rho = 0.5$) were considered.

- The functional covariate, $\mathcal{X}_i$ $(i = 1, \ldots, n)$, was generated from expression

$$\mathcal{X}_i(t) = a_i \cos(2\pi t) + b_i \sin(4\pi t) + 2c_i(t - 0.25)(t - 0.5) \quad \forall t \in [0, 1], \tag{4.32}$$

  where now the random variables a_i, b_i and c_i $(i = 1, \ldots, n)$ were independent and uniformly distributed on the interval $[0, 10]$ (note that we refer to independence both between and within vectors $(a_i, b_i, c_i)^\top$). These curves were discretized on the same grid of 100 equispaced points in $[0, 1]$.

- The i.i.d. random errors, ε_i $(i = 1, \ldots, n)$, were simulated from a $N(0, \sigma_\varepsilon)$ distribution, where $\sigma_\varepsilon^2 = c\sigma_r^2$ with σ_r^2 denoting the empirical variance of the regression $X_{i1}\beta_{01} + \cdots + X_{ip_n}\beta_{0p_n} + r\left(\langle \theta_0, \mathcal{X}_i \rangle\right)$. Note that c is the signal-to-noise ratio, and two values (namely $c = 0.01$ and $c = 0.05$) were considered.

- The true vector of linear coefficients was

$$\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^\top = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^\top.$$

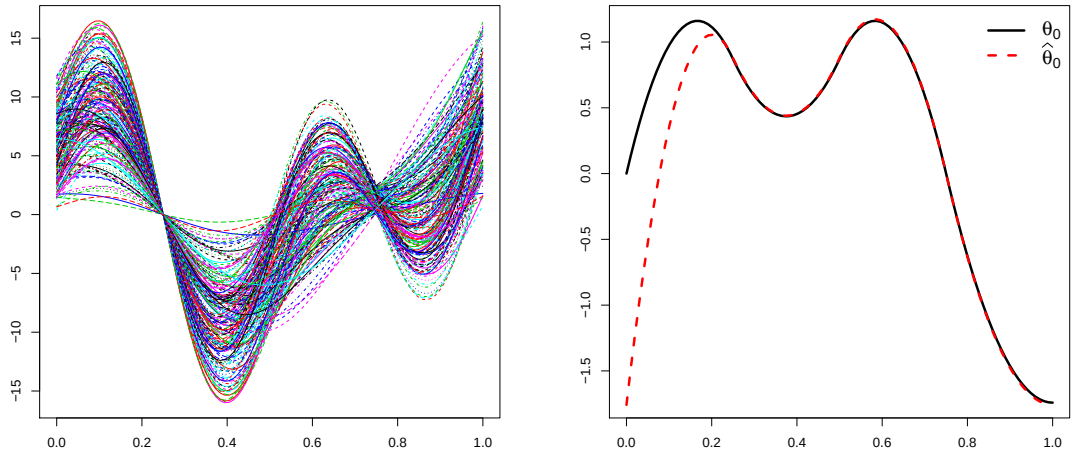
- The true direction of projection, θ_0, was constructed as described in Section 2.4. Values $l = 3$ and $m_n = 3$ were considered and the vector of coefficients of θ_0 in expression (2.36) was:

$$(\alpha_1, \ldots, \alpha_{d_n})^\top = (0, 1.741539, 0, 1.741539, -1.741539, -1.741539)^\top. \quad (4.33)$$

- The inner product and the link function considered were $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$ and $r(\langle\theta_0, \chi\rangle) = \langle\theta_0, \chi\rangle^3$, respectively.

Figure 4.1 shows a sample of 200 curves generated from (4.32) in its left panel, while in the right panel, in black colour and solid line, the functional direction θ_0 is displayed.

Figure 4.1: Sample of 200 curves generated from (4.32) (left panel) and functional direction θ_0 (right panel). In addition, in right panel, it is displayed the estimation, $\widehat{\theta}_0$, of θ_0 obtained from a particular sample in the scenario $(n, p_n, \rho, c) = (100, 50, 0, 0.05)$.



For each simulation case $((n, p_n), \rho, c) \in \{(100, 50), (200, 100)\} \times \{0, 0.5\} \times \{0.01, 0.05\}$, $M = 100$ independent samples were generated from model (4.31). For each sample, we obtained an estimator of the pair $(\boldsymbol{\beta}_0, \theta_0)$ by minimizing the penalized profile least-squares function $\mathcal{Q}(\boldsymbol{\beta}, \theta)$ (see (4.5)). For that, we considered

eligible functional directions as explained in Section 2.4. Epanechnikov kernel was used while the penalty function considered was the SCAD (1.10). The value $a = 3.7$ is usually considered in literature (see Fan [37], Fan and Li [38] or Aneiros et al. [7]). To reduce the quantity of tuning parameters, λ_j, to be selected for each sample, we consider $\lambda_j = \lambda\widehat{\sigma}_{\beta_{0,j,OLS}}$, where $\beta_{0,j,OLS}$ denotes the OLS estimate of $\beta_{0,j}$ in (4.31) and $\widehat{\sigma}_{\beta_{0,j,OLS}}$ is the estimated standard deviation. This tuning parameter, λ, as well as the bandwidth, h, were selected by means of the BIC procedure. More specifically, the BIC value corresponding to $(\widehat{\boldsymbol{\beta}}_{0,h,\lambda}, \widehat{\theta}_{0,h,\lambda})$ (the estimate of the parameter $(\boldsymbol{\beta}_0, \theta_0)$ in the linear model (4.4) obtained by minimizing the profile least-squares function (4.5)) was computed from the routine `select` of the R package *grpreg* (see Breheny and Huang [19]). The main reason why we have used this selector is its low computational cost compared to cross-validation-based selectors (which are time consuming procedures). Moreover, this BIC selector shows good behaviour both in this simulation study and in the real data application reported in Section 4.5. Then, it is noteworthy that its low computational cost takes a main relevance in the estimation of such a complex model as the SSFPLSIM.

4.4.2 Results

First results of the simulation study are presented in Table 4.1 and Figure 4.2 (variable selection) and Table 4.2 ($\boldsymbol{\beta}_0$ estimation).

Table 4.1 shows both the average percentage (restricted only to the true zero coefficients) of coefficients correctly set to zero and the average percentage (restricted only to the true non-zero coefficients) of coefficients erroneously set to zero.

From Table 4.1 we can observe that, as sample size increases, our procedure can detect a greater percentage of non-significant variables. In addition, the percentage of significant variables erroneously set to zero decreases. It is noteworthy that positive dependence between variables gives some advantage in detecting non-significant variables, but it is detrimental to the detection of the significant ones (similar conclusions were obtained in nonfunctional both linear (Huang et al. [65]) and partial linear (Xie and Huang [116]) models, as well as in the semi-functional partial linear model (Aneiros et al. [7])). We can also observe that results are better for $c = 0.01$ than $c = 0.05$, especially for finding the true relevant variables.

Table 4.1: Column "Correct": Average percentage (restricted only to the true zero coefficients) of coefficients correctly set to zero. Column "Incorrect": Average percentage (restricted only to the true non-zero coefficients) of coefficients erroneously set to zero.

			$\rho = 0$		$\rho = 0.5$	
n	p_n	c	Correct	Incorrect	Correct	Incorrect
100	50	0.05	77.404	16.667	84.447	24.333
		0.01	92.830	1.000	96.319	7.667
200	100	0.05	85.052	2.667	91.072	11.333
		0.01	98.619	0.000	99.732	2.667

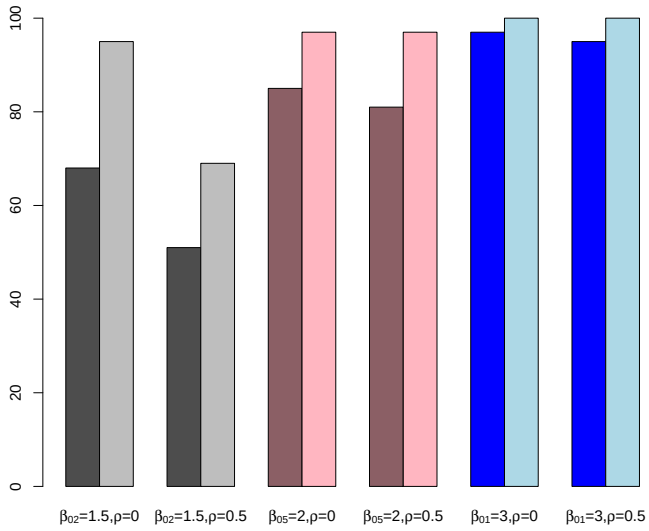
Figure 4.2 shows barplots with the percentage of times that each non-zero coefficient ($\beta_{01} = 3$, $\beta_{02} = 1.5$ and $\beta_{05} = 2$) is not set to zero. Therefore, since the linear covariates are identically distributed, Figure 4.2 allows to analyse the influence of the size of each β_{0j} non-zero coefficient in the detection of the j^{th} variable as influential one. A first conclusion is that, as intuition says, as bigger is the value of β_{0j}, greater is the percentage of success. In general, results also improve if we increase the sample size or if we reduce c (and then, σ_ε^2). In addition, positive dependence between variables makes more difficult the detection of the significant variables, especially for smaller values of β_{0j}.

Table 4.2 reports information about the performance of the penalized least-squares (PLS) estimator of $\boldsymbol{\beta}_0$ in the SSFPLSIM (4.31). Specifically, on the one hand it shows both the mean and standard deviation of the squared errors,

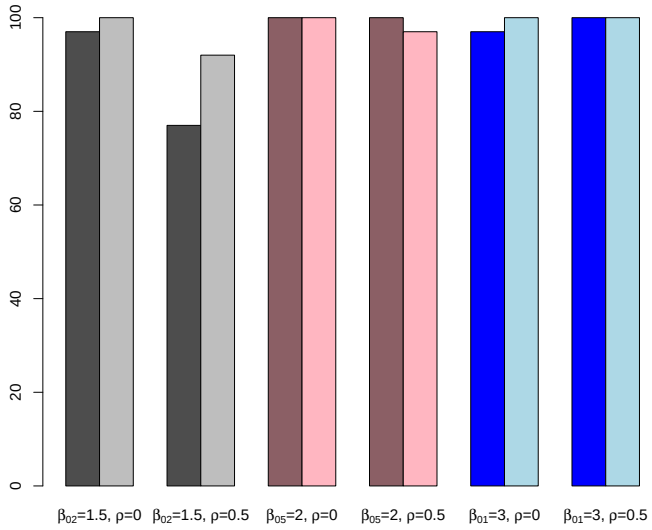
$$\left|\left|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right|\right|^2 = \sum_{j=1}^{p_n} \left(\widehat{\beta}_{0j} - \beta_{0j}\right)^2, \tag{4.34}$$

obtained from the M replicates when both the proposed PLS approach and the ordinary least-squares (OLS) estimator are applied to the SSFPLSIM.

Figure 4.2: Percentage of times that each non-zero coefficient of $\boldsymbol{\beta}_0$ is not set to zero. We use grey for $\beta_{02} = 1.5$, pink for $\beta_{05} = 2$ and blue for $\beta_{01} = 3$. Dark colours correspond to $n = 100$, while light colours match $n = 200$. Values $\rho = 0$ and $\rho = 0.5$ are considered.



(a) $c = 0.05$



(b) $c = 0.01$

On the other hand, Table 4.2 reports the corresponding mean and standard deviation of the squared errors obtained from the OLS approach assuming that one knows in advance what are the non-null coefficients (oracle estimator). Note that the oracle estimator can't be used in practice, but can be seen as a benchmark method in simulation. As expected, the oracle estimator performs the best and the OLS one performs the worst. The proposed PLS estimator shows a good behaviour, since its performance is much closer to the oracle estimator than to the OLS one. In addition, its behaviour improves significantly when the sample size increases (note that the estimation of θ_0 is needed, which, given the complexity of the SSFPLSIM, requires a sufficient sample size) or the signal-to-noise ratio (c) decreases. Note also that the dependence in the linear covariates has effects on both estimators (oracle and OLS) and on the variable selection procedure (PLS): in general, small values of ρ provide better results.

Table 4.2: Mean and standard deviation (SD) of the squared errors (4.34) obtained from ORACLE, (the proposed) PLS and OLS procedures.

			ORACLE		PLS		OLS	
c	ρ	n	Mean	SD	Mean	SD	Mean	SD
0.05	0	100	1.879	1.855	12.742	13.241	89.209	51.619
		200	0.796	0.628	4.055	2.807	61.023	26.082
	0.5	100	2.347	2.882	11.014	9.610	148.431	91.162
		200	1.051	0.956	4.804	3.821	99.872	41.521
0.01	0	100	0.440	0.502	1.359	1.063	22.425	17.874
		200	0.175	0.140	0.656	0.519	12.847	4.738
	0.5	100	0.540	0.733	2.645	1.720	37.210	30.573
		200	0.231	0.207	1.825	1.080	21.067	7.868

In addition, in this simulation study we are going to show the practical behaviour of the estimators related to the semiparametric component. For that, Table 4.3 reports mean and standard deviation of the squared errors (3.13) obtained in

the estimation of θ_0 (by means of kernel-based procedures). From Table 4.3 we can conclude that the performance of the proposed estimator for the single-index (infinite-dimensional) parameter, θ_0, clearly improves when the sample size increases or the signal-to-noise ratio (c) decreases (in a similar way as happened for estimation of the linear (finite-dimensional) parameter $\boldsymbol{\beta}_0$). In addition, it appears that, once the variable selection is performed, the estimation errors are not really affected by higher values of ρ.

Table 4.3: Mean and standard deviation (SD) of the squared errors (3.13) obtained from the proposed procedure (by means of kernel-based estimation).

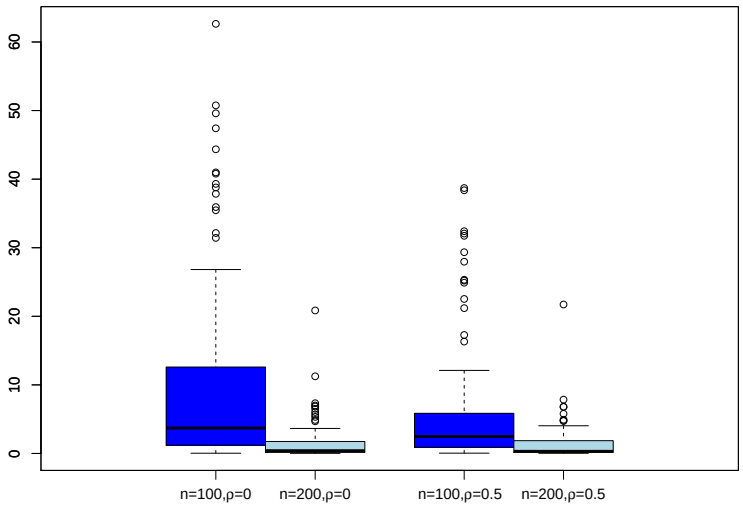
c	ρ	n	Mean	SD
0.05	0	100	0.010	0.011
		200	0.003	0.007
	0.5	100	0.009	0.011
		200	0.004	0.004
0.01	0	100	0.004	0.008
		200	0.001	0.004
	0.5	100	0.004	0.007
		200	0.001	0.004

In order to measure the performance of the proposed estimator ($\widehat{r}(\cdot)$) for the nonparametric component ($r(\cdot)$), M independent test samples with sample size $n_{test} = 100$,

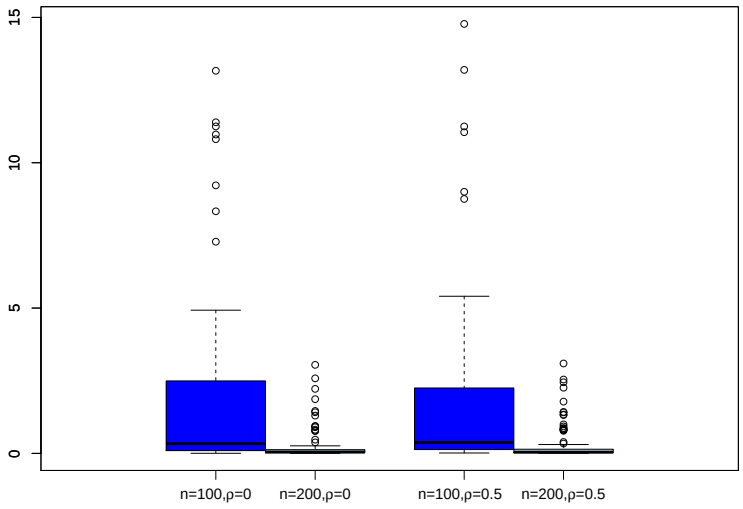
$$\left\{\mathcal{X}_j^{(k)}\right\}_{j=1}^{n_{test}},\ k = 1, \ldots, M,$$

were constructed in a similar way as in Section 4.4.1 (note that these M test samples were also independent of the M (training) samples considered until now).

Figure 4.3: Boxplots of the squared errors (4.35) obtained from the proposed procedure for the several considered scenarios.



(a) $c = 0.05$



(b) $c = 0.01$

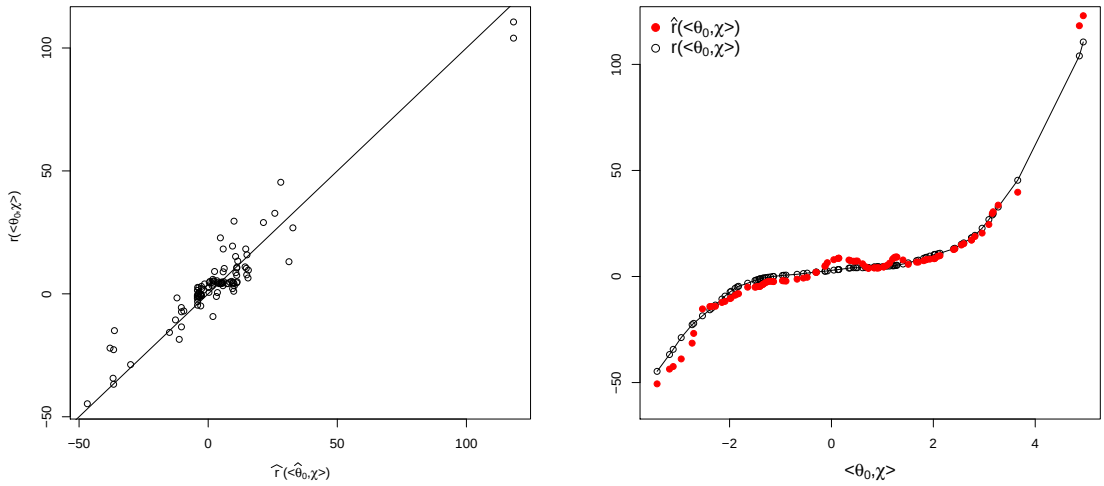
Then, the performance of the estimate for $r(\cdot)$ constructed from the k-th training sample was measured by means of the MSEP

$$MSEP_k = \frac{1}{n_{test}} \sum_{j=1}^{n_{test}} \left(\widehat{r}_k \left(\left\langle \widehat{\theta}_{0k}, \mathcal{X}_j^{(k)} \right\rangle \right) - r \left(\left\langle \theta_0, \mathcal{X}_j^{(k)} \right\rangle \right) \right)^2, \tag{4.35}$$

where $\widehat{\theta}_{0k}$ and $\widehat{r}_k(\cdot)$ denote estimators for θ_0 and $r(\cdot)$, respectively, constructed from information in the k-th training sample. Figure 4.3 displays, for each considered scenario, boxplots with the corresponding $MSEP_k$ values. In a similar way as for estimation of both the linear (finite-dimensional) parameter $\boldsymbol{\beta}_0$ (see Table 4.2) and the single-index (infinite-dimensional) parameter θ_0 (see Table 4.3), Figure 4.3 shows that the performance of the estimate of the nonparametric component, $\widehat{r}(\cdot)$, clearly improves when the sample size increases or the signal-to-noise ratio (c) decreases.

Finally, Figure 4.4 displays, for a particular replicate, values of $r\left(\langle \theta_0, \cdot \rangle\right)$ versus $\widehat{r}\left(\left\langle \widehat{\theta}_0, \cdot \right\rangle\right)$, as well as both $r\left(\langle \theta_0, \cdot \rangle\right)$ and $\widehat{r}\left(\langle \theta_0, \cdot \rangle\right)$ versus $\langle \theta_0, \cdot \rangle$. For a graphic representation of the estimate of θ_0 obtained from such particular replicate, see right panel in Figure 4.1 (red color and dashed line).

Figure 4.4: Real and estimated values, from a particular sample in the scenario $(n, p_n, \rho, c) = (100, 50, 0, 0.05)$, related to the semiparametric component, $r\left(\langle \theta_0, \cdot \rangle\right)$, of the SSFPLSIM (4.31). The curve in the right panel is the true $r(\cdot)$.



4.5 Application to real data

In this section, Tecator's dataset is modelled through different functional regression models, including the SSFPLSIM (4.1) proposed in this chapter. The results obtained show the usefulness of both the SSFPLSIM and the proposed PLS estimation procedure.

Before beginning the next sections dedicated to present variables, modelling, variable selection and prediction, we indicate that, in the estimation of the three models that require variable selection (that is, SLM (1.8), SSFPLM (1.11) and SSFPLSIM (4.1) in Table 4.5), both the tuning parameter, λ, and the bandwidth, h, were selected by means of the BIC procedure. Moreover, the Epanechnikov kernel and the penalty function SCAD (with parameter $a = 3.7$) were used to estimate them (in a similar way as in the simulation study in Section 4.4). In addition, in the SSFPLSIM the order of the splines was $l = 3$, while the number of regularly interior knots, m_n, was selected by means of the BIC procedure (resulting $\widehat{m}_n = 4$); for details on the role of the splines, see (2.36) in Section 2.4.

4.5.1 The data

In this real data application, as happened in Section 3.5, we will consider as variables the percentages of fat, protein and moisture contents (Y_i, X_{1i} and X_{2i}, respectively) and the near-infrared absorbance spectra, $\mathcal{X}_i$, of 215 finely chopped pieces of meat. The variables Y_i, X_{1i} and X_{2i} are scalar, while the corresponding near-infrared absorbance spectra (observed on 100 equally spaced wavelengths (t_j, $j = 1, \ldots, 100$) in the range 850–1050 nm) can be considered as a continuous curve. As usual when one deals with Tecator's dataset, we will use the second derivatives of the absorbance curves, $\mathcal{X}_i^{(2)}$, as functional covariate instead of the original curve (see e.g. Ferraty and Vieu [47] for details). Figure 2.4 displays samples of both the absorbance curves and their second derivatives. For making comparisons with results obtained in previous chapters for the Tecator's dataset, note that in Chapter 2 we also have used second derivatives of the absorbance curves; however, in Chapter 3 the first derivatives were employed (the order of the derivative was selected by means of the 10-folds-CV procedure).

Our purpose in this section is modelling the relationship between the fat content (response), the protein and moisture contents (scalar covariates), and the absorbance spectra (functional covariate) and then, use the model to predict the fat content. In addition, we are interested in whether there are any interaction effects, quadratic effects and/or cubic effects between these scalar covariates.

In order to compare the behaviour of each considered model and estimation procedure, we will split the original sample into two subsamples: a training sample,

$$\mathcal{D}_{train} = \{(X_{i1}, X_{i2}, \mathcal{X}_i^{(2)}, Y_i)\}_{i=1}^{160},$$

and a testing one,

$$\mathcal{D}_{test} = \{(X_{i1}, X_{i2}, \mathcal{X}_i^{(2)}, Y_i)\}_{i=161}^{215}$$

(that is, benchmark partition will be used). In this way, all the estimation task is made only by means of the training sample, while the testing sample is used to measure the quality of the predictions. To quantify the error in the prediction task, the MSEP (2.40) will be used with $n = 160$ and $n_{test} = 55$.

4.5.2 Results

In literature, several models have been used to describe the relation between the fat content and the absorbance spectra (see e.g. Ferraty and Vieu [47] for a functional nonparametric model, and Chen et al. [27] for a multiple index functional model). In Chapter 2, we modelled this dataset using the FSIM (2.1) and compared the performance of the obtained predictions to that provided by the FLM (1.1) and the FNM (1.2). Such three models as well as the corresponding MSEPs obtained from kernel-based estimation procedures (in the case of FNM and FSIM) and functional principal components regression (in the case of FLM) are summarized in Table 4.4.

To improve the performance of the FNM and FSIM, as we have done in Section 3.5, Aneiros-Pérez and Vieu [11] and Wang et al. [110] included in such models, respectively, information from the scalar covariates X_1 and X_2. Nevertheless, in Chapter 3, as well as in those two papers, only linear effects of X_1 and X_2 were considered: no interaction effects, and neither quadratic nor cubic effects. In order to take into account such potential effects, one can extend the case studies of Section 3.5,

Aneiros-Pérez and Vieu [11] and Wang et al. [110] by considering as linear covariates $X_{2j-1} = X_1^j$ and $X_{2j} = X_2^j$ $(j = 1, \cdots, q_n)$, and $X_{p_n} = X_1 X_2$ (we have denoted $p_n = 2q_n + 1$). The corresponding semi-functional partial linear model (SFPLM) and SSFPLSIM for the particular case of $p_n = 7$ (equivalently, for models allowing linear, quadratic and cubic effects, as well as interaction between the covariates X_1 and X_2) are shown in Table 4.5. The SLM (1.8) is also included in such table. Table 4.5 also reports the selected variables when the PLS procedures in Fan and Peng [41] (SLM), Aneiros et al. [7] (SSFPLM) and our proposal (SSPLSIM) are applied, as well as the corresponding MSEPs.

Table 4.4: Values of the MSEPs from some functional models.

	Model	MSEP
FLM	$Y = \gamma_0 + \int_{850}^{1050} \mathcal{X}^{(2)}(t)\gamma(t)dt + \varepsilon$	7.17
FNM	$Y = m(\mathcal{X}^{(2)}) + \varepsilon$	4.06
FSIM	$Y = r\left(\left\langle \theta_0, \mathcal{X}^{(2)} \right\rangle\right) + \varepsilon$	3.49

Table 4.5: Values of the MSEPs from some scalar parametric and functional semi-parametric models when PLS variable selection methods are used. The selected variables are also shown.

	Model	Selected variables	MSEP
SLM	$Y = \sum_{j=1}^{7} X_j \beta_{0j} + \varepsilon$	X_1, X_2, X_7	1.95
SFPLM	$Y = \sum_{j=1}^{7} X_j \beta_{0j} + m(\mathcal{X}^{(2)}) + \varepsilon$	X_1, X_2	1.48
SSFPLSIM	$Y = \sum_{j=1}^{7} X_j \beta_{0j} + r\left(\left\langle \theta_0, \mathcal{X}^{(2)} \right\rangle\right) + \varepsilon$	X_1, X_2, X_4, X_5	1.29

Several conclusions can be drawn from Tables 4.4 and 4.5. First, Table 4.4 shows that the FSIM improves results of both the FLM and FNM. Second, Table 4.5 indicates that to add scalar linear effects in the FNM and FSIM (or, equivalently, to add functional nonparametric or semiparametric effects in the SLM) improves the predictive power of these simpler models. In addition, the percentages of protein (X_1)

and moisture (X_2) contents have a linear influence on the percentage of fat content (Y). X_1 and X_2 also present cubic and quadratic influence on Y, respectively, when the SSFPLSIM is considered, while interaction effects (the covariate X_7 is selected) are only detected in the SLM. Finally, Table 4.5 also shows that our proposed model (SSFPLSIM), which is a mix of all these ideas (semiparametric and partial linear ideas), presents the better performance.

Figure 4.5: Predicted values from the SSFPLSIM vs Observed values.

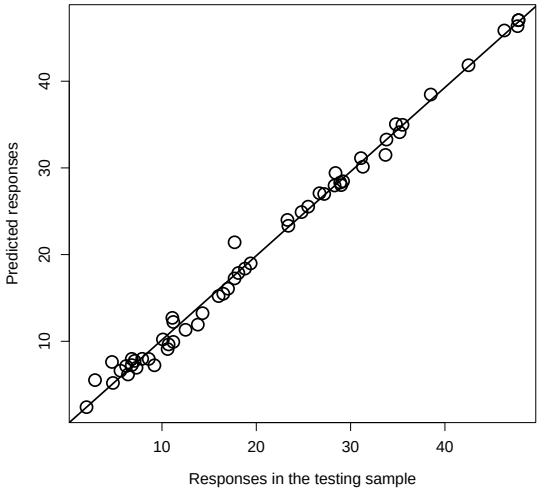
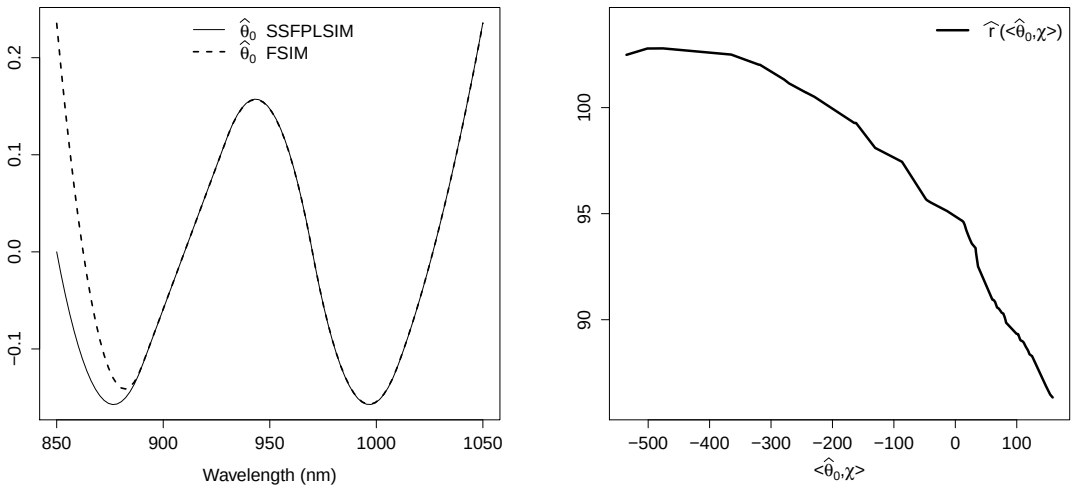


Figure 4.5 displays the predicted values ($\widehat{Y}_i,\ i = 161, \ldots, 215$) from the SSFPLSIM versus the observed ones ($Y_i,\ i = 161, \ldots, 215$). The high predictive power of the SSFPLSIM is evident. The estimates of the functional directions, θ_0, in the FSIM and SSFPLSIM are displayed in Figure 4.6 (left panel). It is worth being noted that both graphics of $\widehat{\theta}_0$ suggests that the two bumps around wavelengths 880 and 1000, as well as the peak around wavelength 940, could be important indicators of the fat content (note that this suggestion is compatible with the findings in Section 2.6). Finally, Figure 4.6 (right panel) shows the estimate of the smooth real-valued function, $r(\cdot)$, in the SSFPLSIM.

Figure 4.6: Left panel: Estimates of the functional directions (θ_0) in the FSIM and SSFPLSIM. Right panel: Estimate of the function $r(\cdot)$ in the SSFPLSIM.



4.5.3 Conclusions

Our real data application evidences the advantages of using the SSFPLSIM (4.1) together with the proposed PLS procedure in terms of accuracy of predictions. In addition, as in the case of FSIM (2.1), the SSFPLSIM presents the advantage of the interpretation of the estimated direction of projection, $\widehat{\theta}_0$, which could also complement the information about how the (second derivative of the) spectrometric curves affect to the fat content.

4.6 Appendix Chapter 4: Proofs

This section presents the proofs of our main results. For that, a major role is played by the technical lemmas provided in Section 4.6.7.3. Note that the Remark 4.11 in Section 4.6.7.1 justifies that such lemmas can be applied under the conditions of our theorems.

Without loss of generality, we will assume that $S_n = \{1, \ldots, s_n\}$.

4.6.1 Proof of Theorem 4.2

Before starting the proof, let us complete in the following way the notations introduced in Section 4.3.1 of the chapter:

$$\boldsymbol{g}_{j,\theta_0} = (g_{j,\theta_0}(\mathcal{X}_1), \ldots, g_{j,\theta_0}(\mathcal{X}_n))^\top \ (0 \leq j \leq p_n),\ \boldsymbol{G}_{\theta_0} = (\boldsymbol{g}_{1,\theta_0}, \ldots, \boldsymbol{g}_{p_n,\theta_0}),$$

and for each $\theta \in \Theta_n$:

$$\widehat{g}_{j,\theta}(\chi) = \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) Z_{ij} \text{ with } Z_{i0} = Y_i \text{ and } Z_{ij} = X_{ij} \ (1 \leq j \leq p_n),$$
$$\widehat{\boldsymbol{g}}_{j,\theta} = (\widehat{g}_{j,\theta}(\mathcal{X}_1), \ldots, \widehat{g}_{j,\theta}(\mathcal{X}_n))^\top \ (0 \leq j \leq p_n) \text{ and } \widehat{\boldsymbol{G}}_\theta = \left(\widehat{\boldsymbol{g}}_{1,\theta}, \ldots, \widehat{\boldsymbol{g}}_{p_n,\theta}\right).$$

In addition, we denote

$$\boldsymbol{\eta}_{\theta_0} = (\boldsymbol{\eta}_{1,\theta_0}, \ldots, \boldsymbol{\eta}_{n,\theta_0})^\top,$$

and

$$\mathcal{Q}^*\left(\boldsymbol{\beta}_{S_n}, \theta\right) = \frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_\theta - \widetilde{\boldsymbol{X}}_{\theta S_n}\boldsymbol{\beta}_{S_n}\right)^\top \left(\widetilde{\boldsymbol{Y}}_\theta - \widetilde{\boldsymbol{X}}_{\theta S_n}\boldsymbol{\beta}_{S_n}\right) + n\sum_{j=1}^{s_n} \mathcal{P}_{\lambda_{j_n}}\left(|\beta_j|\right).$$

To obtain the desired result, it suffices to prove that there exists a local minimizer $\left(\widehat{\boldsymbol{\beta}}_{0S_n}, \widehat{\theta}_0\right)$ of $\mathcal{Q}^*(\boldsymbol{\beta}_{S_n}, \theta)$ such that

$$\left\|\widehat{\boldsymbol{\beta}}_{0S_n} - \boldsymbol{\beta}_{0S_n}\right\| = O_p(u_n),\ \left\|\widehat{\theta}_0 - \theta_0\right\| = O_p(v_n) \tag{4.36}$$

and

$$\left(\widehat{\boldsymbol{\beta}}_0, \widehat{\theta}_0\right) = \left(\left(\widehat{\boldsymbol{\beta}}_{0S_n}^\top, \mathbf{0}_{p_n - s_n}^\top\right)^\top, \widehat{\theta}_0\right) \text{ is a local minimizer of } \mathcal{Q}(\cdot,\cdot), \tag{4.37}$$

where $\mathbf{0}_{p_n - s_n}$ is a vector of zero components with dimension $p_n - s_n$.

First, we will obtain the results in (4.36). For that, it suffices to show that, for any given $\gamma > 0$, there exists a constant C such that, for n large enough,

$$\mathbb{P}\left(\inf_{||\boldsymbol{u}||=C, \theta \in \Theta_n^*} \mathcal{Q}^*\left(\boldsymbol{\beta}_{0S_n} + u_n\boldsymbol{u}, \theta\right) > \mathcal{Q}^*\left(\boldsymbol{\beta}_{0S_n}, \theta_0\right)\right) \geq 1 - \gamma,$$

where $\boldsymbol{u} = (u_1, \ldots, u_{s_n})^\top \in \mathbb{R}^{s_n}$ and $\Theta_n^* = \{\theta \in \Theta_n;\ d(\theta, \theta_0) = v_n\}$. Let us denote

$$\mathcal{Q}^*(\boldsymbol{\beta}_{S_n}, \theta) = \mathcal{L}^*(\boldsymbol{\beta}_{S_n}, \theta) + \mathcal{P}^*(\boldsymbol{\beta}_{S_n}), \tag{4.38}$$

where

$$\mathcal{L}^*(\boldsymbol{\beta}_{S_n}, \theta) = \frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_\theta - \widetilde{\boldsymbol{X}}_{\theta S_n}\boldsymbol{\beta}_{S_n}\right)^\top \left(\widetilde{\boldsymbol{Y}}_\theta - \widetilde{\boldsymbol{X}}_{\theta S_n}\boldsymbol{\beta}_{S_n}\right) \text{ and } \mathcal{P}^*(\boldsymbol{\beta}_{S_n}) = n\sum_{j\in S_n} \mathcal{P}_{\lambda_{j_n}}(|\beta_j|).$$

We have that

$$\mathcal{Q}^*(\boldsymbol{\beta}_{0S_n}, \theta_0) - \mathcal{Q}^*(\boldsymbol{\beta}_{0S_n} + u_n\boldsymbol{u}, \theta) = A_1 + A_2, \tag{4.39}$$

where

$$A_1 = \mathcal{L}^*(\boldsymbol{\beta}_{0S_n}, \theta_0) - \mathcal{L}^*(\boldsymbol{\beta}_{0S_n} + u_n\boldsymbol{u}, \theta) \text{ and } A_2 = \mathcal{P}^*(\boldsymbol{\beta}_{0S_n}) - \mathcal{P}^*(\boldsymbol{\beta}_{0S_n} + u_n\boldsymbol{u}). \tag{4.40}$$

Focusing on A_1, we can write

$$\begin{aligned} 2A_1 \;=\; & \left(\widetilde{\boldsymbol{Y}}_{\theta_0}^\top\widetilde{\boldsymbol{Y}}_{\theta_0} - 2\widetilde{\boldsymbol{Y}}_{\theta_0}^\top\widetilde{\boldsymbol{X}}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n}\right) + \left(\widetilde{\boldsymbol{X}}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n}\right)^\top\widetilde{\boldsymbol{X}}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n} \\ & - \left(\widetilde{\boldsymbol{Y}}_{\theta}^\top\widetilde{\boldsymbol{Y}}_{\theta} - 2\widetilde{\boldsymbol{Y}}_{\theta}^\top\widetilde{\boldsymbol{X}}_{\theta S_n}\boldsymbol{\beta}_{0S_n}\right) - \left(\widetilde{\boldsymbol{X}}_{\theta S_n}\boldsymbol{\beta}_{0S_n}\right)^\top\widetilde{\boldsymbol{X}}_{\theta S_n}\boldsymbol{\beta}_{0S_n} \\ & + 2u_n\left(\widetilde{\boldsymbol{Y}}_{\theta}^\top\widetilde{\boldsymbol{X}}_{\theta S_n} - \left(\widetilde{\boldsymbol{X}}_{\theta S_n}\boldsymbol{\beta}_{0S_n}\right)^\top\widetilde{\boldsymbol{X}}_{\theta S_n}\right)\boldsymbol{u} \\ & - u_n^2\boldsymbol{u}^\top\widetilde{\boldsymbol{X}}_{\theta S_n}^\top\widetilde{\boldsymbol{X}}_{\theta S_n}\boldsymbol{u} \equiv A_{11} + A_{12} - A_{13} - A_{14} + 2u_nA_{15} - A_{16}. \end{aligned} \tag{4.41}$$

Taking into account that

$$\widetilde{\boldsymbol{Y}}_{\theta_0} = \boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta_0} + \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n} + \boldsymbol{\varepsilon} \text{ and } \widetilde{\boldsymbol{Y}}_{\theta} = \boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta} + \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n} + \boldsymbol{\varepsilon},$$

as well as that

$$\widetilde{\boldsymbol{X}}_{\theta_0 S_n} = \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n} + \boldsymbol{\eta}_{\theta_0 S_n} \text{ and } \widetilde{\boldsymbol{X}}_{\theta S_n} = \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n} + \boldsymbol{\eta}_{\theta_0 S_n},$$

we obtain that

$$\begin{aligned} A_{11} &= \left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta_0}\right)^\top \left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta_0}\right) + 2\left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta_0}\right)^\top \boldsymbol{\varepsilon} \\ &\quad -2\left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta_0}\right)^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n} \boldsymbol{\beta}_{0S_n} - 2\boldsymbol{\varepsilon}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n} \boldsymbol{\beta}_{0S_n} \\ &\quad +\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - \left(\boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n}\right)^\top \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n} - 2\left(\boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n}\right)^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n} \boldsymbol{\beta}_{0S_n}, \end{aligned} \tag{4.42}$$

$$\begin{aligned} A_{12} &= \boldsymbol{\beta}_{0S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n} \boldsymbol{\beta}_{0S_n} \\ &\quad +2\boldsymbol{\beta}_{0S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n}^\top \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n} + \left(\boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n}\right)^\top \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n}, \end{aligned} \tag{4.43}$$

$$\begin{aligned} A_{13} &= \left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta}\right)^\top \left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta}\right) + 2\left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta}\right)^\top \boldsymbol{\varepsilon} \\ &\quad -2\left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta}\right)^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n} \boldsymbol{\beta}_{0S_n} - 2\boldsymbol{\varepsilon}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n} \boldsymbol{\beta}_{0S_n} \\ &\quad +\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - \left(\boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n}\right)^\top \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n} - 2\left(\boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n}\right)^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n} \boldsymbol{\beta}_{0S_n}, \end{aligned} \tag{4.44}$$

$$\begin{aligned} A_{14} &= \boldsymbol{\beta}_{0S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n} \boldsymbol{\beta}_{0S_n} \\ &\quad +2\boldsymbol{\beta}_{0S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}^\top \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n} + \left(\boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n}\right)^\top \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n} \end{aligned} \tag{4.45}$$

and

$$\begin{aligned} A_{15} &= \left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta}\right)^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n} \boldsymbol{u} + \left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta}\right)^\top \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{u} \\ &\quad -\boldsymbol{\beta}_{0S_n}^\top \left(\left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n} + \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}^\top \boldsymbol{\eta}_{\theta_0 S_n}\right)\boldsymbol{u} \\ &\quad +\boldsymbol{\varepsilon}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta}\right)_{S_n} \boldsymbol{u} + \boldsymbol{\varepsilon}^\top \boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{u}. \end{aligned} \tag{4.46}$$

Let us denote

$$B = A_{11} + A_{12} - A_{13} - A_{14}. \tag{4.47}$$

From decompositions (4.42)-(4.45), it is easy to obtain that

$$\begin{aligned} B &= \left(\widehat{\boldsymbol{g}}_{0,\theta_0}-\widehat{\boldsymbol{g}}_{0,\theta}\right)^{\top}\left(\widehat{\boldsymbol{g}}_{0,\theta_0}+\widehat{\boldsymbol{g}}_{0,\theta}\right)+2\boldsymbol{g}_{0,\theta_0}^{\top}\left(\widehat{\boldsymbol{g}}_{0,\theta}-\widehat{\boldsymbol{g}}_{0,\theta_0}\right)+2\boldsymbol{\varepsilon}^{\top}\left(\widehat{\boldsymbol{g}}_{0,\theta}-\widehat{\boldsymbol{g}}_{0,\theta_0}\right) \\ &\quad +2\boldsymbol{g}_{0,\theta_0}^{\top}\left(\widehat{\boldsymbol{G}}_{\theta_0}-\widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}\boldsymbol{\beta}_{0S_n}+2\left(\widehat{\boldsymbol{g}}_{0,\theta_0}-\widehat{\boldsymbol{g}}_{0,\theta}\right)^{\top}\boldsymbol{G}_{\theta_0 S_n}\boldsymbol{\beta}_{0S_n} \\ &\quad +2\left(\widehat{\boldsymbol{g}}_{0,\theta}^{\top}\widehat{\boldsymbol{G}}_{\theta S_n}-\widehat{\boldsymbol{g}}_{0,\theta_0}^{\top}\widehat{\boldsymbol{G}}_{\theta_0 S_n}\right)^{\top}\boldsymbol{\beta}_{0S_n}+2\boldsymbol{\varepsilon}^{\top}\left(\widehat{\boldsymbol{G}}_{\theta_0}-\widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}\boldsymbol{\beta}_{0S_n} \\ &\quad +2\boldsymbol{\beta}_{0S_n}^{\top}\widehat{\boldsymbol{G}}_{\theta_0 S_n}^{\top}\left(\widehat{\boldsymbol{G}}_{\theta}-\widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n}\boldsymbol{\beta}_{0S_n}+\boldsymbol{\beta}_{0S_n}^{\top}\left(\widehat{\boldsymbol{G}}_{\theta_0}-\widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}^{\top}\left(\widehat{\boldsymbol{G}}_{\theta_0}+\widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}\boldsymbol{\beta}_{0S_n} \\ &= B_1+B_2+B_3+B_4+B_5+B_6+B_7+B_8+B_9. \end{aligned} \tag{4.48}$$

Now, we are going to obtain bounds (in probability) for each term, B_k $(k=1,\ldots,9)$, in (4.48). Let us denote, for $0\leq j\leq p_n$,

$$\left(\widehat{\boldsymbol{g}}_{j,\theta_0}-\widehat{\boldsymbol{g}}_{j,\theta}\right)=\left(d'_{j1},\ldots,d'_{jn}\right)^{\top} \text{ and } \left(\widehat{\boldsymbol{g}}_{j,\theta_0}+\widehat{\boldsymbol{g}}_{j,\theta}\right)=\left(d''_{j1},\ldots,d''_{jn}\right)^{\top}.$$

On the one hand, from Lemma 4.18 we have that

$$\max_{0\leq j\leq p_n}\sup_{\theta\in\Theta_n^*}\max_{1\leq i\leq n}|d'_{ji}|=O_p\left(\frac{v_n}{hf(h)}\right). \tag{4.49}$$

On the other hand, from the uniform convergence of $\widehat{g}_{j,\theta}(\chi)$ to $g_{j,\theta_0}(\chi)$ (see Lemma 4.17) together with the fact that

$$\max_{0\leq j\leq p_n}\max_{1\leq i\leq n}|g_{j,\theta_0}(\mathcal{X}_i)|=O(1) \tag{4.50}$$

(see Assumption (4.20)), we obtain that

$$\max_{0\leq j\leq p_n}\sup_{\theta\in\Theta_n^*}\max_{1\leq i\leq n}|\widehat{g}_{j,\theta}(\mathcal{X}_i)|=O_p(1); \tag{4.51}$$

then,

$$\max_{0\leq j\leq p_n}\sup_{\theta\in\Theta_n^*}\max_{1\leq i\leq n}|d''_{ji}|=O_p\left(1\right). \tag{4.52}$$

Taking into account (4.49) and (4.52), we have that

$$|B_1| = \left| \sum_{i=1}^{n} d'_{0i} d''_{0i} \right| \leq n \max_{1\leq i\leq n} |d'_{0i}| \max_{1\leq i\leq n} |d''_{0i}| = O_p\left(n \frac{v_n}{hf(h)} \right) \text{ uniformly on } \theta \in \Theta_n^*. \tag{4.53}$$

From (4.49) and (4.50) we obtain that

$$\begin{aligned} |B_2| &= 2 \left| \sum_{i=1}^{n} g_{0,\theta_0}(\mathcal{X}_i)\, d'_{0i} \right| \leq 2n \max_{1\leq i\leq n} |g_{0,\theta_0}(\mathcal{X}_i)| \max_{1\leq i\leq n} |d'_{0i}| \\ &= O_p\left(n \frac{v_n}{hf(h)} \right) \text{ uniformly on } \theta \in \Theta_n^*. \end{aligned} \tag{4.54}$$

From Lemma 4.12 and expression (4.49) we obtain that

$$|B_3| = 2 \left| \sum_{i=1}^{n} d'_{0i} \varepsilon_i \right| = O_p\left(n^{1/2+1/m_\varepsilon} \frac{v_n}{hf(h)} \log n \right) \text{ uniformly on } \theta \in \Theta_n^*. \tag{4.55}$$

Let us denote $\left(\widehat{\boldsymbol{G}}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta} \right)_{S_n} \boldsymbol{\beta}_{0S_n} = (e'_1, \ldots, e'_n)^\top$. From Lemma 4.18 and Assumption (4.29) we obtain:

$$\sup_{\theta\in\Theta_n^*} \max_{1\leq i\leq n} |e'_i| = O_p\left(s_n \frac{v_n}{hf(h)} \right). \tag{4.56}$$

As a consequence, using (4.56) and expression (4.50):

$$\begin{aligned} |B_4| &= \left| \boldsymbol{g}_{0,\theta_0}^\top \left(\widehat{\boldsymbol{G}}_{\theta_0} - \widehat{\boldsymbol{G}}_{\theta} \right)_{S_n} \boldsymbol{\beta}_{0S_n} \right| = \left| \sum_{i=1}^{n} g_{0,\theta_0}(\mathcal{X}_i)\, e'_i \right| \leq n \max_{1\leq i\leq n} |g_{0,\theta_0}(\mathcal{X}_i)| \max_{1\leq i\leq n} |e'_i| \\ &= O_p\left(n s_n \frac{v_n}{hf(h)} \right) \text{ uniformly on } \theta \in \Theta_n^*. \end{aligned} \tag{4.57}$$

Let us denote $\boldsymbol{G}_{\theta_0 S_n} \boldsymbol{\beta}_{0S_n} = (g_1, \ldots, g_n)^\top$. If we take into account hypotheses

(4.29) and (4.50), we obtain

$$\max_{1\leq i\leq n}|g_i| = O_p(s_n). \tag{4.58}$$

Therefore, using (4.49) and (4.58) we have

$$\begin{aligned} |B_5| &= 2\left|\left(\widehat{\boldsymbol{g}}_{0,\theta_0}-\widehat{\boldsymbol{g}}_{0,\theta}\right)^\top \boldsymbol{G}_{\theta_0}\boldsymbol{\beta}_{0S_n}\right| = 2\left|\sum_{i=1}^{n} d'_{0i}g_i\right| \leq n\max_{1\leq i\leq n}|d'_{0i}|\max_{1\leq i\leq n}|g_i| \\ &= O_p\left(ns_n\frac{v_n}{hf(h)}\right) \text{ uniformly on } \theta\in\Theta_n^*. \end{aligned} \tag{4.59}$$

Let us use the notation $\left(\widehat{\boldsymbol{G}}_{\theta_0}-\widehat{\boldsymbol{G}}_{\theta}\right)_{S_n} = \left(b'_{ij}\right)_{\substack{1\leq i\leq n\\ 1\leq j\leq s_n}}$. The application of Lemma 4.18 gives us

$$\sup_{\theta\in\Theta_n^*}\max_{1\leq i\leq n}\max_{1\leq j\leq s_n}\left|b'_{ij}\right| = O_p\left(\frac{v_n}{hf(h)}\right). \tag{4.60}$$

Therefore, Assumption (4.29), Lemma 4.12 and expression (4.60) give us:

$$\begin{aligned} |B_7| &= \left|\boldsymbol{\varepsilon}^\top\left(\widehat{\boldsymbol{G}}_{\theta_0}-\widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}\boldsymbol{\beta}_{0S_n}\right| = \left|\sum_{j=1}^{s_n}\sum_{i=1}^{n}\varepsilon_i b'_{ij}\beta_{0j}\right| \leq s_n\max_{1\leq j\leq s_n}\left|\sum_{i=1}^{n}\varepsilon_i b'_{ij}\right|\max_{1\leq j\leq s_n}|\beta_{0j}| \\ &= O_p\left(n^{1/2+1/m_\varepsilon}s_n\frac{v_n}{hf(h)}\log n\right) \text{ uniformly on } \theta\in\Theta_n^*. \end{aligned} \tag{4.61}$$

Let us denote $\widehat{\boldsymbol{G}}_{\theta_0 S_n}^\top\boldsymbol{\beta}_{0S_n}^\top = (g'_1,\ldots,g'_n)^\top$. If we take into account assumptions (4.29) and (4.51), we obtain

$$\max_{1\leq i\leq n}|g'_i| = O_p(s_n). \tag{4.62}$$

Using expressions (4.62) and (4.56), we have:

$$\begin{aligned} |B_8| &= \left|\boldsymbol{\beta}_{0S_n}^\top\widehat{\boldsymbol{G}}_{\theta_0 S_n}^\top\left(\widehat{\boldsymbol{G}}_{\theta}-\widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n}\boldsymbol{\beta}_{0S_n}\right| = \left|\sum_{i=1}^{n}g'_i e'_i\right| = n\max_{1\leq i\leq n}|g'_i|\max_{1\leq i\leq n}|e'_i| \\ &= O_p\left(\frac{ns_n^2 v_n}{hf(h)}\right) \text{ uniformly on } \theta\in\Theta_n^*. \end{aligned} \tag{4.63}$$

Let us denote $\left(\widehat{\boldsymbol{G}}_{\theta_0}+\widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}\boldsymbol{\beta}_{0S_n}=(e_1'',\ldots,e_n'')^\top$. By virtue of expression (4.51) and Assumption (4.29), we can write:

$$\sup_{\theta\in\Theta_n^*}\max_{1\leq i\leq n}|e_i''|=O_p(s_n). \tag{4.64}$$

Now application of (4.56) and (4.64) gives us

$$\begin{aligned}|B_9| &= \left|\boldsymbol{\beta}_{0S_n}^\top\left(\widehat{\boldsymbol{G}}_{\theta_0}-\widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}^\top\left(\widehat{\boldsymbol{G}}_{\theta_0}+\widehat{\boldsymbol{G}}_{\theta}\right)_{S_n}\boldsymbol{\beta}_{0S_n}\right|=\left|\sum_{i=1}^n e_i'e_i''\right|\leq n\max_{1\leq i\leq n}|e_i''|\max_{1\leq i\leq n}|e_i'|\\ &= O_p\left(ns_n^2\frac{v_n}{hf(h)}\right)\text{ uniformly on }\theta\in\Theta_n^*. \end{aligned}\tag{4.65}$$

The term B_6 can be re-written in the following manner:

$$B_6=\left(\widehat{\boldsymbol{g}}_{0,\theta}-\widehat{\boldsymbol{g}}_{0,\theta_0}\right)^\top\widehat{\boldsymbol{G}}_{\theta S_n}\boldsymbol{\beta}_{0S_n}+\widehat{\boldsymbol{g}}_{0,\theta_0}^\top\left(\widehat{\boldsymbol{G}}_{\theta}-\widehat{\boldsymbol{G}}_{\theta_0}\right)_{S_n}\boldsymbol{\beta}_{0S_n}=B_{61}+B_{62}.$$

If one considers (4.51) instead of (4.50), then similar arguments as those used to obtain the orders of B_5 and B_4 (see (4.59) and (4.57), respectively) give

$$B_{61}=O_p\left(ns_n\frac{v_n}{hf(h)}\right)\text{ and }B_{62}=O_p\left(ns_n\frac{v_n}{hf(h)}\right)\text{ uniformly on }\theta\in\Theta_n^*,$$

respectively; so we have that

$$B_6=O_p\left(ns_n\frac{v_n}{hf(h)}\right)\text{ uniformly on }\theta\in\Theta_n^*. \tag{4.66}$$

It is noteworthy that, as consequence of our assumptions, all the $O_p(\cdot)$ in (4.53)-(4.55), (4.57), (4.59), (4.61), (4.63), (4.65) and (4.66) are $O_p(nu_n^2)$. Therefore, we have proved that

$$B=O_p(nu_n^2)\text{ uniformly on }\theta\in\Theta_n^*. \tag{4.67}$$

The term A_{15} (see (4.46)) can be studied in a similar way as (A6) in Aneiros et al. [7], but considering our Lemma 4.17 instead of Lemma A.3 of Aneiros et al. [7]. Specifically, let us denote

$$r_n=\frac{\log p_n\psi_{\mathcal{C}}\left(1/n\right)}{nf(h)}, \tag{4.68}$$

and $\left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta}\right)^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_\theta\right)_{S_n} = (a_1, \ldots, a_{s_n})$. From Lemma 4.17 we obtain that

$$\sup_{\theta \in \Theta_n^*} \max_{1 \leq j \leq s_n} |a_j| = O_p\left(n\left(h^{2\alpha} + r_n\right)\right) \tag{4.69}$$

(take into account that, because it is assumed that $ns_n v_n = O(hf(h))$, it is verified that $r_n^* = r_n$, where v_n was defined in (4.8) while r_n^* was defined in (4.114) and used in Lemma 4.17). Then, Cauchy-Schwarz inequality and (4.69) give:

$$\left|\left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta}\right)^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_\theta\right)_{S_n} \boldsymbol{u}\right| = \left|\sum_{j=1}^{s_n} a_j u_j\right| = O_p\left(n s_n^{1/2}\left(h^{2\alpha} + r_n\right)\right) \|\boldsymbol{u}\|$$
$$\text{uniformly on } \theta \in \Theta_n^*. \tag{4.70}$$

Furthermore, if in addition we use Assumption (4.29), we obtain

$$\left|\boldsymbol{\beta}_{0S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_\theta\right)_{S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_\theta\right)_{S_n} \boldsymbol{u}\right| = O_p\left(n s_n^{3/2}(h^{2\alpha} + r_n)\right) \|\boldsymbol{u}\|$$
$$\text{uniformly on } \theta \in \Theta_n^*. \tag{4.71}$$

Let us denote, $\left(\boldsymbol{g}_{j,\theta_0} - \widehat{\boldsymbol{g}}_{j,\theta}\right) = (d_{j1}, \ldots, d_{jn})^\top$ for $0 \leq j \leq p_n$. From Lemma 4.17 we have that

$$\max_{0 \leq j \leq p_n} \sup_{\theta \in \Theta_n^*} \max_{1 \leq i \leq n} |d_{ji}| = O_p\left(h^\alpha + \sqrt{r_n}\right). \tag{4.72}$$

In addition, as a consequence of Cauchy-Schwarz inequality, Lemma 4.13 and (4.72), we obtain that

$$\begin{aligned}\left|\left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\theta}\right)^\top \boldsymbol{\eta}_{\theta_0 S_n} \boldsymbol{u}\right| &= \left|\sum_{j=1}^{s_n}\sum_{i=1}^{n} d_{0i}\eta_{ij,\theta_0} u_j\right| \leq s_n^{1/2} \max_{1 \leq j \leq s_n}\left|\sum_{i=1}^{n} d_{0i}\eta_{ij,\theta_0}\right| \|\boldsymbol{u}\| \\ &= O_p\left(n^{1/2} s_n^{1/2}(h^\alpha + \sqrt{r_n}) \log n\right) \|\boldsymbol{u}\| \text{ uniformly on } \theta \in \Theta_n^*.\end{aligned} \tag{4.73}$$

Let us denote $\boldsymbol{\beta}_{0S_n}^\top \left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_\theta\right)_{S_n}^\top = (e_1, \ldots, e_n)^\top$. From Assumption (4.29) and

Lemma 4.17 it is obtained

$$\sup_{\theta\in\Theta_n^*}\max_{1\leq i\leq n}|e_i| = O_p\left(s_n\left(h^\alpha+\sqrt{r_n}\right)\right). \tag{4.74}$$

The use of Cauchy-Schwarz inequality, Lemma 4.13 and (4.74) gives us

$$\begin{aligned}\left|\boldsymbol{\beta}_{0S_n}^\top\left(\boldsymbol{G}_{\theta_0}-\widehat{\boldsymbol{G}}_\theta\right)_{S_n}^\top\boldsymbol{\eta}_{\theta_0 S_n}\boldsymbol{u}\right| &= \left|\sum_{j=1}^{s_n}\sum_{i=1}^{n}e_i\eta_{ij,\theta_0}u_j\right| \leq s_n^{1/2}\max_{1\leq j\leq s_n}\left|\sum_{i=1}^{n}e_i\eta_{ij,\theta_0}\right|\|\boldsymbol{u}\| \\ &= O_p\left(n^{1/2}s_n^{3/2}(h^\alpha+\sqrt{r_n})\log n\right)\|\boldsymbol{u}\| \\ &\quad \text{uniformly on } \theta\in\Theta_n^*. \end{aligned}\tag{4.75}$$

Let us use the notation $\left(\boldsymbol{G}_{\theta_0}-\widehat{\boldsymbol{G}}_\theta\right)_{S_n} = (b_{ij})_{\substack{1\leq i\leq n\\ 1\leq j\leq s_n}}$. Application of Lemma 4.17 gives us

$$\sup_{\theta\in\Theta_n^*}\max_{1\leq i\leq n}\max_{1\leq j\leq s_n}|b_{ij}| = O_p\left(h^\alpha+\sqrt{r_n}\right). \tag{4.76}$$

Finally, using Cauchy-Schwarz inequality, Lemma 4.12 and (4.76) we obtain

$$\begin{aligned}\left|\boldsymbol{\varepsilon}^\top\left(\boldsymbol{G}_{\theta_0}-\widehat{\boldsymbol{G}}_\theta\right)_{S_n}\boldsymbol{u}\right| &= \left|\sum_{j=1}^{s_n}\sum_{i=1}^{n}b_{ij}\varepsilon_i u_j\right| \leq s_n^{1/2}\max_{1\leq j\leq s_n}\left|\sum_{i=1}^{n}b_{ij}\varepsilon_i\right|\|\boldsymbol{u}\| \\ &= O_p\left(n^{1/2+1/m_\varepsilon}s_n^{1/2}(h^\alpha+r_n^{1/2})\log n\right)\|\boldsymbol{u}\| \\ &\quad \text{uniformly on } \theta\in\Theta_n^*. \end{aligned}\tag{4.77}$$

Then, from Lemma 4.23 together with the fact that all the orders $O_p(\cdot)$ involved in (4.70), (4.71), (4.73), (4.75) and (4.77) are $O_p(nu_n)$, we obtain that

$$A_{15} = O_p(nu_n)\|\boldsymbol{u}\| \text{ uniformly on } \theta\in\Theta_n^*. \tag{4.78}$$

Now we focus on the term A_{16} (see (4.41)). Using Lemma 4.20 (considering s_n, $\boldsymbol{X}_{\theta S_n}$ and $\boldsymbol{B}_{\theta_0 S_n\times S_n}$ instead of p_n, $\boldsymbol{X}_\theta$ and $\boldsymbol{B}_{\theta_0}$, respectively) we have that

$$A_{16} = nu_n^2\left(\boldsymbol{u}^\top\boldsymbol{B}_{\theta_0 S_n\times S_n}\boldsymbol{u}+o_p(1)\right) \tag{4.79}$$

uniformly over $\{\boldsymbol{u}\in\mathbb{R}^{p_n}, ||\boldsymbol{u}|| = C\}$ and over $\theta\in\Theta_n^*$. From (4.41), (4.47), (4.67),

(4.78) and (4.79), we obtain that

$$2A_1 = O_p(nu_n^2) + O_p\left(nu_n^2\right)||\boldsymbol{u}|| - nu_n^2\left(\boldsymbol{u}^\top \boldsymbol{B}_{\theta_0 S_n \times S_n}\boldsymbol{u} + o_p(1)\right), \tag{4.80}$$

where all the orders of convergence are uniform in $||\boldsymbol{u}|| = C$ and over $\theta \in \Theta_n^*$.

Focusing now on A_2 (see (4.40)), let us note that A_2 is only linked to the linear part of the model (4.1). Therefore, as in (A14) in Aneiros et al. [7], a Taylor expansion together with assumptions (4.27), (4.28) and $u_n/\min_{j\in S_n}\lambda_{j_n} = o(1)$ give

$$\begin{aligned} A_2 &= -\sum_{j=1}^{s_n}\left(nu_n\mathcal{P}'_{\lambda_{j_n}}(|\beta_{0j}|)sgn(\beta_{0j})u_j + nu_n^2\mathcal{P}''_{\lambda_{j_n}}(|\beta_{0j}|)u_j^2(1+o(1))\right) \\ &= O\left(nu_n s_n^{1/2}\delta_n\right)||\boldsymbol{u}|| + O\left(nu_n^2\rho_n\right)||\boldsymbol{u}||^2. \end{aligned} \tag{4.81}$$

Finally, from expressions (4.39), (4.80) and (4.81), we obtain that

$$\begin{aligned} \mathcal{Q}^*(\boldsymbol{\beta}_{0S_n},\theta_0) - \mathcal{Q}^*\left(\boldsymbol{\beta}_{0S_n} + u_n\boldsymbol{u}, \theta_0 + v_n v\right) &= O_p(nu_n^2) + O_p\left(nu_n^2\right)||\boldsymbol{u}|| \\ &\quad + O\left(nu_n s_n^{1/2}\delta_n\right)||\boldsymbol{u}|| + O\left(nu_n^2\rho_n\right)||\boldsymbol{u}||^2 \\ &\quad - nu_n^2\left(\boldsymbol{u}^\top \boldsymbol{B}_{\theta_0 S_n\times S_n}\boldsymbol{u} + o_p(1)\right). \end{aligned} \tag{4.82}$$

Therefore, taking into account that $s_n^{1/2}\delta_n = O(u_n)$ and $\rho_n \to 0$ as $n \to \infty$, together with Assumption (4.23), it is possible to choose a sufficiently large C in such a way that the last term in (4.82) dominates the other terms uniformly on $||\boldsymbol{u}|| = C$. This fact completes the proof of (4.36).

Now, we will obtain the result in (4.37). Because of $\left(\widehat{\boldsymbol{\beta}}_{0S_n}, \widehat{\theta}_0\right)$ is a local minimizer of $\mathcal{Q}^*(\boldsymbol{\beta}_{S_n},\theta)$ verifying (4.36), to prove (4.37) it suffices to obtain that:

$$\mathcal{Q}\left(\left(\widehat{\boldsymbol{\beta}}_{0S_n}^\top, \mathbf{0}_{p_n-s_n}^\top\right)^\top, \widehat{\theta}_0\right) = \min_{||\boldsymbol{\beta}_{\overline{S}_n}||\le Cu_n}\mathcal{Q}\left(\left(\widehat{\boldsymbol{\beta}}_{0S_n}^\top, \boldsymbol{\beta}_{\overline{S}_n}^\top\right)^\top, \widehat{\theta}_0\right), \tag{4.83}$$

where $\boldsymbol{\beta}_{\overline{S}_n} = (\beta_{s_n+1},\ldots,\beta_{p_n})^\top$. For that, we will show that both

$$\left.\frac{\partial\mathcal{Q}\left(\boldsymbol{\beta},\widehat{\theta}_0\right)}{\partial\beta_j}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{j\vartheta}} > 0 \text{ for } 0<\vartheta<Cu_n \tag{4.84}$$

and

$$\left.\frac{\partial \mathcal{Q}\left(\boldsymbol{\beta}, \widehat{\theta}_0\right)}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{j\vartheta}} < 0 \text{ for } -Cu_n < \vartheta < 0 \tag{4.85}$$

hold, where $j \in \{s_n+1, \ldots, p_n\}$ and $\boldsymbol{\beta}^{j\vartheta}$ denotes a vector with dimension p_n obtained from $\left(\widehat{\boldsymbol{\beta}}_{0S_n}^{\top}, \boldsymbol{\beta}_{\overline{S}_n}^{\top}\right)^{\top}$ by changing their jth component by ϑ. Simple calculations give, for $s_n + 1 \leq j \leq p_n$,

$$\begin{aligned}\left.\frac{\partial \mathcal{Q}\left(\boldsymbol{\beta}, \widehat{\theta}_0\right)}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{j\vartheta}} &= -\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j^{\top}\left(\widetilde{\boldsymbol{Y}}_{\widehat{\theta}_0} - \widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\boldsymbol{\beta}_0\right) + \left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j^{\top}\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\left(\boldsymbol{\beta}^{j\vartheta} - \boldsymbol{\beta}_0\right) \\ &\quad + n\mathcal{P}'_{\lambda_{jn}}\left(|\vartheta|\right) sgn(\vartheta),\end{aligned} \tag{4.86}$$

where $\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j$ denotes de jth column of $\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}$. Therefore, to prove (4.84) and (4.85) it suffices to show that the sign of (4.86) is determined by $sgn(\vartheta)$. On the one hand, we have that

$$\begin{aligned}\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j^{\top}\left(\widetilde{\boldsymbol{Y}}_{\widehat{\theta}_0} - \widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\boldsymbol{\beta}_0\right) &= \left(\boldsymbol{g}_{j,\theta_0} - \widehat{\boldsymbol{g}}_{j,\widehat{\theta}_0}\right)^{\top}\left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\widehat{\theta}_0}\right) + \left(\boldsymbol{g}_{j,\theta_0} - \widehat{\boldsymbol{g}}_{j,\widehat{\theta}_0}\right)^{\top}\boldsymbol{\varepsilon} \\ &\quad - \left(\boldsymbol{g}_{j,\theta_0} - \widehat{\boldsymbol{g}}_{j,\widehat{\theta}_0}\right)^{\top}\left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\widehat{\theta}_0}\right)\boldsymbol{\beta}_0 + \boldsymbol{\eta}_{j,\theta_0}^{\top}\left(\boldsymbol{g}_{0,\theta_0} - \widehat{\boldsymbol{g}}_{0,\widehat{\theta}_0}\right) \\ &\quad + \boldsymbol{\eta}_{j,\theta_0}^{\top}\boldsymbol{\varepsilon} - \boldsymbol{\eta}_{j,\theta_0}^{\top}\left(\boldsymbol{G}_{\theta_0} - \widehat{\boldsymbol{G}}_{\widehat{\theta}_0}\right)\boldsymbol{\beta}_0.\end{aligned} \tag{4.87}$$

The quantities in expression (4.87) can be bounded in the same way as we have done to bound expression A_{15} obtaining (4.78) but now using Lemma 4.12 instead of Lemma 4.23 (note that $\widehat{\theta}_0 \subset \Theta_n$). Therefore, it can be obtained that

$$\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j^{\top}\left(\widetilde{\boldsymbol{Y}}_{\widehat{\theta}_0} - \widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\boldsymbol{\beta}_0\right) = O_p\left(n^{1/2+1/m_\varepsilon}\log n\right). \tag{4.88}$$

On the other hand, we have that

$$\begin{aligned}\left|\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j^{\top}\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\left(\boldsymbol{\beta}^{j\vartheta}-\boldsymbol{\beta}_0\right)\right| &\leq \left|\left|\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j\right|\right|\left|\left|\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right|\right|\left|\left|\boldsymbol{\beta}^{\vartheta j}-\boldsymbol{\beta}_0\right|\right| \\ &= \left|\left|\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j\right|\right|\left|\left|\Delta_{\max}\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}^{\top}\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)\right|\right|\left|\left|\boldsymbol{\beta}^{j\vartheta}-\boldsymbol{\beta}_0\right|\right|. \end{aligned} \tag{4.89}$$

From Lemma 4.20 together with Assumption (4.21) we obtain that

$$\left|\left|\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j\right|\right| = O_p\left(n^{1/2}\right) \tag{4.90}$$

uniformly over $s_n+1\leq j\leq p_n$, while using a similar reasoning of that employed in proof of Lemma 4.20 we have that

$$\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}^{\top}\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0} = n\boldsymbol{B}_{\theta_0}+o_p(n). \tag{4.91}$$

Therefore, from (4.89), (4.90) and (4.91) together with the fact that $\left|\left|\boldsymbol{\beta}^{j\vartheta}-\boldsymbol{\beta}_0\right|\right| = O_p(u_n)$, we obtain that

$$\left|\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j^{\top}\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\left(\boldsymbol{\beta}^{j\vartheta}-\boldsymbol{\beta}_0\right)\right| = O_p\left(nu_n\Delta_{\max}^{1/2}(\boldsymbol{B}_{\theta_0})\right). \tag{4.92}$$

Finally, using (4.86), (4.88) and (4.92) we obtain that

$$\begin{aligned}\left.\frac{\partial\mathcal{Q}\left(\boldsymbol{\beta},\widehat{\theta}_0\right)}{\partial\beta_j}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{j\vartheta}} &= n\lambda_{jn}\left(O_p\left(n^{-1/2+1/m_\varepsilon}\lambda_{jn}^{-1}\log n\right)+O_p\left(\lambda_{jn}^{-1}u_n\Delta_{\max}^{1/2}(\boldsymbol{B}_{\theta_0})\right)\right. \\ &\quad \left.+\lambda_{jn}^{-1}\mathcal{P}'_{\lambda_{jn}}\left(|\vartheta|\right)sgn(\vartheta)\right).\end{aligned}$$

Thus, taking into account our assumptions, we have proved that the sign of $\partial\mathcal{Q}\left(\boldsymbol{\beta},\widehat{\theta}_0\right)/\partial\beta_j|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{j\vartheta}}$ is completely determined by that of ϑ. Therefore equations (4.84) and (4.85) are checked and, as a consequence, the proof of (4.37) is completed.

Because we have proven both (4.36) and (4.37), the proof of our Theorem 4.2 concludes. ■

4.6.2 Proof of Theorem 4.4

Taking our Theorem 4.2 into account, similar steps as those used to prove Theorem 3.2(a) in Aneiros et al. [7] can be followed to obtain the proof of our Theorem 4.4.

Specifically, by construction $\widehat{\boldsymbol{\beta}}_0^\top = (\widehat{\boldsymbol{\beta}}_{0S_n}^\top, \mathbf{0}_{p_n-s_n}^\top)$ and then,

$$\begin{aligned}
\mathbb{P}\left(S_n \neq \widehat{S}_n\right) &= \mathbb{P}\left(\exists j \in S_n \text{ such that } j \notin \widehat{S}_n\right) \\
&\leq \mathbb{P}\left(\exists j \in S_n \text{ such that } \left|\widehat{\beta}_{0j} - \beta_{0j}\right| = |\beta_{0j}|\right) \\
&\leq \mathbb{P}\left(\exists j \in S_n \text{ such that } \left|\widehat{\beta}_{0j} - \beta_{0j}\right| \geq \min_{\ell \in S_n} |\beta_{0\ell}|\right) \\
&\leq \mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right\| \geq \min_{\ell \in S_n} |\beta_{0\ell}|\right) \\
&\leq \mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right\| \geq \min_{\ell \in S_n} |\lambda_{\ell_n}|\right),
\end{aligned}$$

where the last inequality is derived from Assumption (4.28). Now using that $u_n / \min_{j \in S_n} \lambda_{j_n} = o(1)$ and Theorem 4.2 we obtain that

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right\| \geq \min_{\ell \in S_n} |\lambda_{\ell_n}|\right) \longrightarrow 0 \quad \text{as } n \to \infty,$$

concluding the proof.

■

4.6.3 Proof of Theorem 4.5

Because $\left(\widehat{\boldsymbol{\beta}}_{0S_n}, \widehat{\theta}_0\right)$ is a local minimum of $Q^*(\boldsymbol{\beta}_{S_n}, \theta)$, for each $j \in S_n$ it is verified that:

$$\left.\frac{\partial \mathcal{Q}^*\left(\boldsymbol{\beta}_{S_n}, \theta\right)}{\partial \beta_j}\right|_{(\boldsymbol{\beta}_{S_n}, \theta) = \left(\widehat{\boldsymbol{\beta}}_{0S_n}, \widehat{\theta}_0\right)} = 0. \tag{4.93}$$

After some Taylor expansion and using assumptions (4.26), (4.28), the fact that

$u_n/\min_{j\in S_n}\{\lambda_{jn}\} = o(1)$ and Theorem 4.2, we obtain that

$$\begin{aligned}\left.\frac{\partial \mathcal{Q}^*\left(\boldsymbol{\beta}_{S_n},\theta\right)}{\partial \beta_j}\right|_{(\boldsymbol{\beta}_{S_n},\theta)=\left(\widehat{\boldsymbol{\beta}}_{0S_n},\widehat{\theta}_0\right)} &= -\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j^{\top}\left(\widetilde{\boldsymbol{Y}}_{\widehat{\theta}_0}-\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0 S_n}\boldsymbol{\beta}_{0S_n}\right)\\ &\quad+\left(\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0}\right)_j^{\top}\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0 S_n}\left(\widehat{\boldsymbol{\beta}}_{0S_n}-\boldsymbol{\beta}_{0S_n}\right)\\ &\quad+n\mathcal{P}'_{\lambda_{jn}}\left(|\beta_{0j}|\right)\operatorname{sgn}(\beta_{0j})+n\mathcal{P}''_{\lambda_{jn}}\left(|\beta_{0j}|\right)\left(\widehat{\beta}_{0j}-\beta_{0j}\right)\\ &\quad+O_p(nu_n^2).\end{aligned}$$

Then, by virtue of (4.93), it can be written

$$\begin{aligned}\mathbf{0} &= -\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0 S_n}^{\top}\left(\widetilde{\boldsymbol{Y}}_{\widehat{\theta}_0}-\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0 S_n}\boldsymbol{\beta}_{0S_n}\right)+\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0 S_n}^{\top}\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0 S_n}\left(\widehat{\boldsymbol{\beta}}_{0S_n}-\boldsymbol{\beta}_{0S_n}\right)\\ &\quad+n\boldsymbol{c}_{S_n}+n\boldsymbol{V}_{S_n\times S_n}\left(\widehat{\boldsymbol{\beta}}_{0S_n}-\boldsymbol{\beta}_{0S_n}\right)+O_p\left(s_n^{1/2}nu_n^2\right).\end{aligned}\tag{4.94}$$

Now, from (4.94) and Lemma 4.22 we have that

$$\begin{aligned}&n^{1/2}\left(\boldsymbol{B}_{\theta_0 S_n\times S_n}+\boldsymbol{V}_{S_n\times S_n}+o_p(1)\right)\left(\widehat{\boldsymbol{\beta}}_{0S_n}-\boldsymbol{\beta}_{0S_n}+\left(\boldsymbol{B}_{\theta_0 S_n\times S_n}+\boldsymbol{V}_{S_n\times S_n}+o_p(1)\right)^{-1}\boldsymbol{c}_{S_n}\right)\\ &=n^{-1/2}\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0 S_n}^{\top}\left(\widetilde{\boldsymbol{Y}}_{\widehat{\theta}_0}-\widetilde{\boldsymbol{X}}_{\widehat{\theta}_0 S_n}\boldsymbol{\beta}_{0S_n}\right)+O_p\left(s_n^{1/2}n^{1/2}u_n^2\right).\end{aligned}\tag{4.95}$$

We should note that the first term on the right-hand side of the equality (4.95) matches the term A_{15} in (4.41) (when $\theta=\widehat{\theta}_0$ is considered in (4.41)) after multiplying it by $n^{-1/2}$ and removing $\boldsymbol{u}$. Note also that the orders in (4.70), (4.71), (4.73), (4.75) and (4.77) are still true if both the vector $\boldsymbol{u}$ is removed and p_n in r_n is changed by s_n (note that p_n comes from Lemma 4.17, where the maximum is taken over p_n elements; in the particular case of expressions (4.70), (4.71), (4.73), (4.75) and (4.77), the corresponding number of elements is s_n). Therefore, taking into account the decomposition (4.46) of A_{15}, denoting by γ_n the maximum of the orders in equations (4.70), (4.71), (4.73), (4.75) and (4.77) when p_n in r_n is changed by s_n, and multiplying each side of (4.95) by $\boldsymbol{A}_n\sigma_\varepsilon^{-1}\boldsymbol{B}_{\theta_0 S_n\times S_n}^{-1/2}$, we obtain that

$$\begin{aligned}&n^{1/2}\boldsymbol{A}_n\sigma_\varepsilon^{-1}\boldsymbol{B}_{\theta_0 S_n\times S_n}^{-1/2}\left(\boldsymbol{B}_{\theta_0 S_n\times S_n}+\boldsymbol{V}_{S_n\times S_n}\right)\left(\widehat{\boldsymbol{\beta}}_{0S_n}-\boldsymbol{\beta}_{0S_n}+\left(\boldsymbol{B}_{\theta_0 S_n\times S_n}+\boldsymbol{V}_{S_n\times S_n}\right)^{-1}\boldsymbol{c}_{S_n}\right)\\ &=n^{-1/2}\boldsymbol{A}_n\sigma_\varepsilon^{-1}\boldsymbol{B}_{\theta_0 S_n\times S_n}^{-1/2}\boldsymbol{\eta}_{\theta_0 S_n}^{\top}\boldsymbol{\varepsilon}+\boldsymbol{A}_n\boldsymbol{B}_{\theta_0 S_n\times S_n}^{-1/2}O_p\left(n^{-1/2}\gamma_n+s_n^{1/2}n^{1/2}u_n^2\right).\end{aligned}\tag{4.96}$$

Note that from assumptions in Theorem 4.5, $(n^{-1/2}\gamma_n + s_n^{1/2} n^{1/2} u_n^2) = o(1)$ holds. Now using that $\boldsymbol{A}_n \boldsymbol{A}_n^\top \to \boldsymbol{A}$ together with Assumption (4.23), we have that

$$||\boldsymbol{A}_n \boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2}||^2 \leq ||\boldsymbol{A}_n||^2 ||\boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2}||^2 = \Delta_{max}(\boldsymbol{A})(1+o(1)) \frac{1}{\Delta_{min}(\boldsymbol{B}_{\theta_0 S_n \times S_n})} = O(1).$$

Therefore, expression (4.96) can be simplified in the following way:

$$\begin{aligned} & n^{1/2} \boldsymbol{A}_n \sigma_\varepsilon^{-1} \boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2} (\boldsymbol{B}_{\theta_0 S_n \times S_n} + \boldsymbol{V}_{S_n \times S_n}) \times \\ & \times \left(\widehat{\boldsymbol{\beta}}_{0 S_n} - \boldsymbol{\beta}_{0 S_n} + (\boldsymbol{B}_{\theta_0 S_n \times S_n} + \boldsymbol{V}_{S_n \times S_n} + o_p(1))^{-1} \boldsymbol{c}_{S_n} \right) \\ = \; & n^{-1/2} \boldsymbol{A}_n \sigma_\varepsilon^{-1} \boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2} \boldsymbol{\eta}_{\theta_0 S_n}^\top \boldsymbol{\varepsilon} + o_p(1). \end{aligned}$$

As a consequence, the result will be proved if we show that

$$n^{-1/2} \boldsymbol{A}_n \sigma_\varepsilon^{-1} \boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2} \boldsymbol{\eta}_{\theta_0 S_n}^\top \boldsymbol{\varepsilon} = \sum_{i=1}^n \boldsymbol{Z}_{ni} \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{0}, \boldsymbol{A}), \tag{4.97}$$

where we use the notation $\boldsymbol{Z}_{ni} = n^{-1/2} \boldsymbol{A}_n \sigma_\varepsilon^{-1} \boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2} \boldsymbol{\eta}_{i, \theta_0 S_n} \varepsilon_i$.

For that, following exactly the same development used in Aneiros et al. [7], we obtain that the i.i.d. sequence of q-dimensional random vectors $\{\boldsymbol{Z}_{ni}\}$ satisfies the conditions of the Lindeberg–Feller central limit theorem. Specifically, firstly it is verified that

$$\begin{aligned} \sum_{i=1}^n Var(Z_{ni}) &= n^{-1} \sigma_\varepsilon^{-2} \boldsymbol{A}_n \boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2} \sum_{i=1}^n Var(\boldsymbol{\eta}_{i, \theta_0 S_n} \varepsilon_i) (\boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2})^\top \boldsymbol{A}_n^\top \\ &= \boldsymbol{A}_n \boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2} \boldsymbol{B}_{\theta_0 S_n \times S_n} \left(\boldsymbol{B}_{\theta_0 S_n \times S_n}^{-1/2} \right)^\top \boldsymbol{A}_n^\top = \boldsymbol{A}_n \boldsymbol{A}_n^\top \to \boldsymbol{A}. \end{aligned} \tag{4.98}$$

Secondly, to check Lindeberg's condition note that

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}\left(\|Z_{ni}\|^2 1_{\|Z_{ni}\| > \epsilon} \right) &= n \mathbb{E}\left(\|Z_{ni}\|^2 1_{\|Z_{ni}\| > \epsilon} \right) \\ &\leq n \mathbb{E}\left(\|Z_{ni}\|^4 \right)^{1/2} \mathbb{E}\left(1_{\|Z_{ni}\| > \epsilon} \right)^{1/2} \\ &= n \mathbb{E}\left(\|Z_{ni}\|^4 \right)^{1/2} \mathbb{P}\left(\|Z_{ni}\| > \epsilon \right)^{1/2}. \end{aligned} \tag{4.99}$$

Then, taking into account Assumption (4.22), it is verified that $\mathbb{E}(\varepsilon^4) < \infty$, and Assumption (4.21) gives us that $\max_{1\leq j\leq s_n} \mathbb{E}(\eta_{i,j\theta_0}^4) = O(1)$. Therefore, we obtain for fixed i $(i = 1, \ldots, n)$

$$\begin{aligned} \mathbb{E}\left(\|Z_{ni}\|^4\right) &= n^{-2}\sigma_\epsilon^{-4} \left\|\boldsymbol{A}_n \boldsymbol{B}_{\theta_0 S_n \times S_n}^{1/2}\right\|^4 \mathbb{E}\left((\boldsymbol{\eta}_{i,\theta_0 S_n}^\top \boldsymbol{\eta}_{i,\theta_0 S_n})^2\right) \mathbb{E}\left(\varepsilon_i^4\right) \\ &= O\left(\left(\frac{s_n}{n\Delta_{min}(\boldsymbol{B}_{\theta_0 S_n \times S_n}}\right)^2\right). \end{aligned} \tag{4.100}$$

Moreover, from Markov's inequality it is obtained

$$\mathbb{P}\left(\|Z_{ni}\| > \epsilon\right) \leq \frac{\mathbb{E}\left(\|Z_{ni}\|^2\right)}{\epsilon^2} = O\left(\frac{s_n}{n\Delta_{min}(\boldsymbol{B}_{\theta_0 S_n \times S_n})}\right). \tag{4.101}$$

Finally, from expressions (4.99)-(4.101) and taking into account that $s_n^3/n\Delta_{min}^3(\boldsymbol{B}_{\theta_0 S_n \times S_n}) = o(1)$ we obtain that

$$\sum_{i=1}^n \mathbb{E}\left(\|Z_{ni}\|^2 1_{\|Z_{ni}\|>\epsilon}\right) = O\left(\left(\frac{s_n}{n^{1/3}\Delta_{min}(\boldsymbol{B}_{\theta_0 S_n \times S_n})}\right)^{3/2}\right) = o(1).$$

Therefore, conditions on the Lindeberg-Feller central limit theorem are verified. Then, since assumptions (4.2) and (4.18) are verified, expression (4.98) gives us

$$\begin{aligned} &\mathbb{E}\left(\sum_{i=1}^n \boldsymbol{Z}_{ni}\right) = \boldsymbol{0}, \\ &\text{Var}\left(\sum_{i=1}^n \boldsymbol{Z}_{ni}\right) = \boldsymbol{A}_n^\top \boldsymbol{A}_n \to \boldsymbol{A}. \end{aligned} \tag{4.102}$$

Therefore, the result (4.97) is verified, which completes the proof. ■

4.6.4 Proof of Theorem 4.7

We have that

$$\begin{aligned}
|\widehat{r}_{\theta}(\chi) - r_{\theta_0}(\chi)| &= \left| \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) (r_{\theta_0}(\mathcal{X}_i) + \varepsilon_i) - r_{\theta_0}(\chi) \right. \\
&\quad \left. + \sum_{j \in S_n} \widehat{g}_{j,\theta}(\chi) \left(\beta_{0j} - \widehat{\beta}_{0j} \right) \right| \\
&\leq \left| \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) (r_{\theta_0}(\mathcal{X}_i) + \varepsilon_i) - r_{\theta_0}(\chi) \right| \\
&\quad + \left| \sum_{j \in S_n} (\widehat{g}_{j,\theta}(\chi) - g_{j,\theta_0}(\chi)) \left(\beta_{0j} - \widehat{\beta}_{0j} \right) \right| \\
&\quad + \left| \sum_{j \in S_n} g_{j,\theta_0}(\chi) \left(\beta_{0j} - \widehat{\beta}_{0j} \right) \right| \\
&\leq \left| \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) (r_{\theta_0}(\mathcal{X}_i) + \varepsilon_i) - r_{\theta_0}(\chi) \right| \\
&\quad + s_n^{1/2} \sup_{\chi \in \mathcal{C}, j \in S_n, \theta \in \Theta_n} |\widehat{g}_{j,\theta}(\chi) - g_{j,\theta_0}(\chi)| \left|\left| \widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0 \right|\right| \\
&\quad + s_n^{1/2} \sup_{\chi \in \mathcal{C}, j \in S_n, \theta \in \Theta_n} |g_{j,\theta_0}(\chi)| \left|\left| \widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0 \right|\right| . \qquad (4.103)
\end{aligned}$$

Applying to the decompositions considered in the proofs of Lemmas 4.14-4.17 (see Section 4.6.7), the techniques used in Ferraty et al. [49] to prove their Theorem 2, we obtain that

$$\sup_{\chi \in \mathcal{C}, \theta \in \Theta_n} \left| \sum_{i=1}^{n} w_{n,h,\theta}(\chi, \mathcal{X}_i) (r_{\theta_0}(\mathcal{X}_i) + \varepsilon_i) - r_{\theta_0}(\chi) \right| = O_p\left(h^{\alpha} + \sqrt{\frac{\psi_{\mathcal{C}}(1/n)}{n f(h)}} \right) \qquad (4.104)$$

(remember that, as a consequence of our assumptions on v_n, we have that $\psi_{\Theta_n}(1/n) = 0$). In addition, Lemma 4.17 gives us

$$\sup_{\chi \in \mathcal{C}, j \in S_n, \theta \in \Theta_n} |\widehat{g}_{j,\theta}(\chi) - g_{j,\theta_0}(\chi)| = O_p\left(h^{\alpha} + \sqrt{r_n} \right), \qquad (4.105)$$

where r_n was defined in (4.68).

Therefore, since $\sup_{\chi\in\mathcal{C},j\in S_n}|g_{j,\theta_0}(\chi)|=O(1)$, using Theorem 4.2 and expressions (4.103), (4.104) and (4.105), we can finally obtain

$$\begin{aligned}\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}|\widehat{r}_\theta(\chi)-r_{\theta_0}(\chi)| &= O_p\left(h^\alpha+\sqrt{\frac{\psi_{\mathcal{C}}(1/n)}{nf(h)}}\right)\\ &\quad +O_p\left(s_n^{1/2}u_n\left(h^\alpha+\sqrt{r_n}\right)\right)+O_p\left(s_n^{1/2}u_n\right)\\ &= O_p\left(h^\alpha+\sqrt{\frac{\psi_{\mathcal{C}}(1/n)}{nf(h)}}\right)+O_p\left(s_n^{1/2}u_n\right). \blacksquare\end{aligned}$$

4.6.5 Proof of Corollary 4.8

Trivial. ■

4.6.6 Proof of Corollary 4.9

From the first part and second one in Condition A in Corollary 4.9, one obtains that $\psi_{\mathcal{C}}(1/n)\approx\log n$ (see Example 4 in Ferraty et al. [49], page 338) and $f(h)\approx h$ (see Lemma 13.6 in Ferraty and Vieu [47]), respectively. Therefore, taking into account that $h\approx C(\log n/n)^{1/(2\alpha+1)}$, it is obtained that

$$h^\alpha+\sqrt{\frac{\psi_{\mathcal{C}}(1/n)}{nf(h)}}=O\left(\left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)}\right). \tag{4.106}$$

In addition, taking into account that $u_n=O(\sqrt{s_n}n^{-1/2})$ (see Condition D in Corollary 4.9), together with the fact that $s_n\approx cn^\gamma$ with $0<2\gamma\leq 1-2\alpha/(2\alpha+1)$, one obtains that

$$\sqrt{s_n}u_n\approx c'n^{\gamma-1/2}=O\left(\left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)}\right). \tag{4.107}$$

(4.106) and (4.107) conclude the proof. ■

4.6.7 Technical lemmas

This section shows some known lemmas used to prove the results in this chapter. In addition, novel lemmas, as well as their proofs, are presented. Their interest is not

restricted to the proof of our theorems, they could be useful in other contexts. In fact, some assumptions used in our novel lemmas are more general than the corresponding ones imposed in our theorems.

4.6.7.1 General assumptions

Let us present some additional assumptions to be used in some of the lemmas in Section 4.6.7.3.

Condition on the set of directions and the associated topology. The set of directions, Θ_n (see (4.8)), satisfies

$$\Theta_n \subset \bigcup_{j=1}^{N_{\Theta_n,\epsilon}} B(\theta_{\epsilon,j}, \epsilon), \text{ where } \epsilon = 1/n \tag{4.108}$$

and $N_{\Theta_n,\epsilon}$ is the minimal number of open balls in $(\mathcal{H}, d(\cdot,\cdot))$ of radius ϵ which are necessary to cover Θ_n.

Conditions on the entropies and the balls in (4.7). Let $\psi_{\Theta_n}(\epsilon)$ denotes the Kolmogorov entropy of $(\Theta_n, d(\cdot,\cdot))$ (that is, $\psi_{\Theta_n}(\epsilon) = \log(N_{\Theta_n,\epsilon})$). It is assumed that:

$$\begin{array}{rcl} \exists \beta > 1 \text{ such that } p_n \exp\left\{(1-\beta\log p_n)\left(\psi_{\mathcal{C}}\left(\frac{1}{n}\right)+\psi_{\Theta_n}\left(\frac{1}{n}\right)\right)\right\} & \to & 0 \\ \text{as } n & \to & \infty, \end{array} \tag{4.109}$$

and

$$\sup_{\theta\in\Theta_n} \max_{k\in\{1,\ldots,N^{\theta}_{\mathcal{C},1/n}\}} d(\chi^{\theta}_k, \chi^{\theta^*}_{k^*}) = O(1/n) \text{ (for notation, see (4.115))}. \tag{4.110}$$

Condition linking the entropies and the small-ball probabilities. There exists a constant $C_{13} > 0$ such that, for n large enough,

$$\psi_{\mathcal{C}}\left(\frac{1}{n}\right) + \psi_{\Theta_n}\left(\frac{1}{n}\right) \leq \frac{C_{13} n f(h)}{\alpha_n \log p_n}, \text{ where } \alpha_n \to \infty \text{ as } n \to \infty \tag{4.111}$$

(the function $f(\cdot)$ was defined in (4.12) and (4.13)).

Remark 4.11 *In the theorems presented in this chapter, the condition imposed on* v_n *(i.e.,* $ns_nv_n = O(hf(h))$*) implies that*

$$\Theta_n \subset B(\theta_0, 1/n). \tag{4.112}$$

Nevertheless, that does not necessarily happen in the novel lemmas proposed in Section 4.6.7.3 (they are applicable under more general scenarios). For this reason, the more general Assumption (4.108), as well as the new assumptions (4.109), (4.110) and (4.111), are introduced here (note that assumptions (4.10), (4.11) and (4.14), used in our theorems, are particular cases of (4.109), (4.110) and (4.111), respectively: it suffices to consider $N_{\Theta_n,\epsilon} = 1$ *and* $\theta_{\epsilon,1} = \theta_0$ *in (4.108)). Assumptions (4.108), (4.109) and (4.111) (together with (4.7) and (4.13) in Section 4.3.2) are common when one needs to obtain uniform orders over* $\mathcal{C}$ *and* Θ_n *(see, for instance, Wang et al. [110]). Finally, Condition (4.110) is really specific to the functional setting addressed here and, therefore, requires a deeper discussion. It is a technical assumption that links the topologies of* $(\mathcal{C}, d_\theta(\cdot,\cdot))$ *and* $(\Theta_n, d(\cdot,\cdot))$*, allowing to bound the difference* $d_\theta(\chi_k^\theta,\cdot) - d_{\theta^*}(\chi_{k^*}^{\theta^*},\cdot)$ *by means of bounds based on the topology of* $(\Theta_n, d(\cdot,\cdot))$ *(for details, see the proof of Lemma 4.14). In fact, Condition (4.110) could be changed by the more general (but maybe harder to interpret) one*

$$\sup_{\chi\in\mathcal{C}} \sup_{\theta\in\Theta_n} \max_{k\in\left\{1,\ldots,N_{\mathcal{C},1/n}^{\theta}\right\}} \left|d_\theta\left(\chi_k^\theta,\chi\right) - d_{\theta^*}\left(\chi_{k^*}^{\theta^*},\chi\right)\right| = O\left(1/n\right).$$

It is worth noting that Condition (4.110) is satisfied if, for instance, the following assumption holds:

$$d(\chi_k^\theta, \chi_{k^*}^{\theta^*}) \leq Cd(\theta,\theta^*), \text{ uniformly on } \theta\in\Theta_n \text{ and } k\in\left\{1,\ldots,N_{\mathcal{C},1/n}^{\theta}\right\}, \tag{4.113}$$

where C *denotes a positive constant. Actually, Condition (4.113) can be seen as a smoothness assumption: roughly speaking, it imposes "smooth changes" between the coverings (4.7) induced by the topologies of* $(\mathcal{C}, d_\theta(\cdot,\cdot))$ $(\theta\in\Theta)$ *when the indexes* θ *are close (to be more precise, see the definition of both* k^* *and* θ^* *in (4.115)).*

4.6.7.2 Additional notation

The following notation,

$$k^*_{(\theta,k,\epsilon)} = \arg \min_{k' \in \left\{1,\ldots,N^{\theta_{j_{(\theta,\epsilon)}}}_{\mathcal{C},\epsilon}\right\}} d(\chi^{\theta}_{\epsilon,k}, \chi^{\theta_{\epsilon,j_{(\theta,\epsilon)}}}_{\epsilon,k'}), \text{ where } j_{(\theta,\epsilon)} = \arg \min_{j \in \{1,\ldots,N_{\Theta_n,\epsilon}\}} d(\theta, \theta_{\epsilon,j}),$$

and

$$r^*_n = \frac{\log p_n \left(\psi_{\mathcal{C}}\left(1/n\right) + \psi_{\Theta_n}\left(1/n\right)\right)}{nf(h)}, \tag{4.114}$$

generalizes the previous notation $k^0_{(\theta,k,\epsilon)}$ (see (4.9)) and r_n (see (4.68)), respectively, to the more general setting considered in Section 4.6.7. In the sake of brevity, we will denote

$$\chi^{\theta}_k = \chi^{\theta}_{1/n,k},\ \theta^* = \theta_{1/n,j_{(\theta,1/n)}} \text{ and } k^* = k^*_{(\theta,k,1/n)}. \tag{4.115}$$

Finally, we introduce the statistics

$$\widehat{F}_\theta(\chi) = \frac{\sum_{i=1}^n K\left(d_\theta\left(\chi, \mathcal{X}_i\right)/h\right)}{n\mathbb{E}\left(K\left(d_\theta\left(\chi, \mathcal{X}\right)/h\right)\right)} \text{ and } \widehat{g}^*_{j,\theta}(\chi) = \widehat{F}_\theta(\chi)\widehat{g}_{j,\theta}(\chi)\ \ (j = 0, 1, \ldots, p_n),$$

which will be used in the proofs of some of our lemmas.

4.6.7.3 Results

Lemma 4.12 *(Lemma 3 in Aneiros-Pérez and Vieu [12]) Let $\{V_i\}_{i=1}^n$ be a zero-mean, stationary, independent and real process verifying that $\exists m > 4$ such that $\max_{1\leq i\leq n} \mathbb{E}|V_i|^m = O(1)$. Assume that $\{a_{ij}, i, j = 1, \ldots, n\}$ is a sequence of positive numbers such that $\max_{1\leq i,j\leq n} |a_{ij}| = O(a_n)$. Then,*

$$\max_{1\leq j\leq n} \left|\sum_{i=1}^n a_{ij}V_i\right| = O_p\left(a_n n^{1/2+1/m}\log n\right).$$

The conclusion of this lemma remains unchanged when a_{ij} are random variables satisfying the conditions earlier in probability.

Lemma 4.13 *(Lemma A.2 in Aneiros et al. [7]) Let $\{V_{ijk}\}_{i=1}^n$ $(1 \leq j \leq u_n, 1 \leq k \leq v_n)$ be independent random variables with zero mean and*

$\forall m \geq 2$, $E|V_{ijk}|^m \leq C_V(m!/2)$, *where* $0 < C_V < \infty$ *is a constant. Assume that* $\{a_{ijk}, 1 \leq i \leq n, 1 \leq j \leq u_n, 1 \leq k \leq v_n\}$ *is a set of positive numbers such that* $\max_{\substack{1\leq i\leq n\\ 1\leq j\leq u_n\\ 1\leq k\leq v_n}} |a_{ijk}| = O(a_n)$. *If* $u_n v_n n^{-\log n} \to 0$ *as* $n \to \infty$ *then:*

$$\max_{1\leq j\leq u_n} \max_{1\leq k\leq v_n} \left| \sum_{i=1}^{n} a_{ijk} V_{ijk} \right| = O_p\left(a_n n^{1/2} \log n\right).$$

The conclusion of this lemma remains unchanged when a_{ijk} *are random variables satisfying the conditions earlier in probability.*

Lemma 4.14 *Under assumptions (4.7), (4.12), (4.13), (4.15), (4.19) and (4.108)-(4.111), if in addition* $\sup_{\theta\in\Theta_n} \langle \theta, \theta \rangle^{1/2} = O(1)$ *and* $\mathcal{X}_i$ *are i.i.d, we have that there exists a positive constant,* C*, such that, for all* $\epsilon > 0$ *and* n *large enough,*

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n} \sup_{\chi\in\mathcal{C}} \left| \widehat{F}_\theta(\chi) - 1 \right| > \epsilon \sqrt{r_n^*} \right) \leq C\left(p_n^{-C\epsilon^2} + \left(N_{\Theta_n,1/n} N_{\mathcal{C},1/n} \right)^{1-C\epsilon^2 \log p_n} \right).$$

Proof of Lemma 4.14. To carry out the proof of this lemma, we will follow the same steps used in Ferraty et al. [49] to demonstrate their Lemma 8. Firstly, the following decomposition can be made:

$$\sup_{\theta\in\Theta_n} \sup_{\chi\in\mathcal{C}} \left| \widehat{F}_\theta(\chi) - 1 \right| \leq F_1 + F_2 + F_3, \tag{4.116}$$

where we have denoted

$$F_1 = \sup_{\theta\in\Theta_n} \sup_{\chi\in\mathcal{C}} \left| \widehat{F}_\theta(\chi) - \widehat{F}_\theta\left(\chi^\theta_{k_{(\theta,\chi,1/n)}} \right) \right|,$$

$$F_2 = \sup_{\theta\in\Theta_n} \sup_{\chi\in\mathcal{C}} \left| \widehat{F}_\theta\left(\chi^\theta_{k_{(\theta,\chi,1/n)}} \right) - \mathbb{E}\left(\widehat{F}_\theta\left(\chi^\theta_{k_{(\theta,\chi,1/n)}} \right) \right) \right|$$

and

$$F_3 = \sup_{\theta\in\Theta_n} \sup_{\chi\in\mathcal{C}} \left| \mathbb{E}\left(\widehat{F}_\theta\left(\chi^\theta_{k_{(\theta,\chi,1/n)}} \right) \right) - \mathbb{E}\left(\widehat{F}_\theta(\chi) \right) \right|,$$

with

$$k_{(\theta,\chi,1/n)} = \arg \min_{k\in\{1,\ldots,N^\theta_{\mathcal{C},1/n}\}} d_\theta(\chi, \chi^\theta_{1/n,k}) \tag{4.117}$$

(see also notation (4.115)).

Furthermore, note that using hypotheses (4.13) and (4.15), if $K(1) > C > 0$, it is verified that $\forall\theta \in \Theta_n$ and $\forall\chi \in \mathcal{C}$ there exist constants $0 < C < C' < \infty$ such that

$$Cf(h) \leq \mathbb{E}(K(d_\theta(\chi, \mathcal{X}_i)/h)) \leq C' f(h). \tag{4.118}$$

The same result it is obtained when $K(1) = 0$ with the combination of assumptions (4.13) and (4.12) (see Lemma 4.4 in Ferraty and Vieu [47]).

Study of the term F_1.

Starting with the term F_1, if we denote

$$\mathbb{I}_i = 1_{B_\theta(\chi,h)\cup B_\theta\left(\chi^\theta_{k_{\left(\theta,\chi,\frac{1}{n}\right)}},h\right)}(\mathcal{X}_i),$$

we can write

$$\begin{aligned} F_1 &\leq \sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{K\left(d_\theta(\chi,\mathcal{X}_i)/h\right)}{\mathbb{E}\left(K\left(d_\theta(\chi,\mathcal{X}_i)/h\right)\right)} - \frac{K\left(d_\theta\left(\chi^\theta_{k_{(\theta,\chi,1/n)}},\mathcal{X}_i\right)/h\right)}{\mathbb{E}\left(K\left(d_\theta(\chi^\theta_{k_{(\theta,\chi,1/n)}},\mathcal{X}_i)/h\right)\right)}\right| \\ &\leq \frac{C}{f(h)}\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\frac{1}{n}\sum_{i=1}^{n}\left|K\left(\frac{d_\theta\left(\chi,\mathcal{X}_i\right)}{h}\right) - K\left(\frac{d_\theta\left(\chi^\theta_{k_{(\theta,\chi,1/n)}},\mathcal{X}_i\right)}{h}\right)\right|\mathbb{I}_i, \end{aligned}$$

where in the last inequality we have used (4.118). Now we have to consider two situations derived from Assumption (4.15):

- Case $K(1) = 0$. K is Lipschitz continuous on $[0, 1]$. Therefore,

$$\left|K\left(\frac{d_\theta(\chi,\mathcal{X}_i)}{h}\right) - K\left(\frac{d_\theta\left(\chi^\theta_{k_{\left(\theta,\chi,\frac{1}{n}\right)}},\mathcal{X}_i\right)}{h}\right)\right| \leq \frac{C}{h}d_\theta\left(\chi,\chi^\theta_{k_{\left(\theta,\chi,\frac{1}{n}\right)}}\right).$$

Then,

$$F_1 \leq \frac{C}{n} \sup_{\theta\in\Theta_n} \sup_{\chi\in\mathcal{C}} \sum_{i=1}^{n} \frac{1}{nhf(h)} \mathbb{I}_i.$$

Let us denote

$$Z_i = \frac{1}{nhf(h)} \mathbb{I}_i \quad (i = 1, \ldots, n).$$

It is clear that

$$Z_i = O\left(\frac{1}{nhf(h)}\right).$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left(Z_i\right) &= \frac{1}{nhf(h)} \mathbb{P}\left(\left\{d_\theta\left(\chi, \mathcal{X}_i\right) < h\right\} \bigcup \left\{d_\theta\left(\chi_{k_{\left(\theta,\chi,\frac{1}{n}\right)}}^{\theta}, \mathcal{X}_i\right) < h\right\}\right) \\
&\leq \frac{1}{nhf(h)} \left(\phi_{\chi,\theta}(h) + \phi_{\chi_{k_{\left(\theta,\chi,\frac{1}{n}\right)}}^{\theta},\theta}(h)\right) \\
&\leq \frac{C}{nh},
\end{aligned}$$

where we have used Assumption (4.13) for obtaining the last inequality, and using this assumption again it is obtained that

$$\mathrm{Var}\left(Z_i\right) = \frac{1}{n^2h^2f(h)^2} \mathrm{Var}\left(\mathbb{I}_i\right) \leq \frac{C}{n^2h^2f(h)}.$$

A standard inequality for sums of bounded random variables (see Corollary A.9 in Ferraty and Vieu [47]) gives us that exists a positive constant, C, such that, for all $\epsilon > 0$ and n large enough

$$\mathbb{P}\left(F_1 > \epsilon\sqrt{r_n^*}\right) \leq \mathbb{P}\left(\sup_{\theta\in\Theta_n} \sup_{\chi\in\mathcal{C}} \frac{C}{n} \sum_{i=1}^{n} Z_i > \epsilon\sqrt{r_n^*}\right) \leq C p_n^{-C\epsilon^2}. \tag{4.119}$$

- Case $K(1) > C > 0$. Now K is Lipschitz on $[0, 1)$. In this case, we have to decompose F_1 into the following terms:

$$F_1 \leq C \sup_{\theta\in\Theta_n} \sup_{\chi\in\mathcal{C}} \left(F_{11} + F_{12} + F_{13}\right), \tag{4.120}$$

where:

$$F_{11} = \frac{1}{nf(h)} \sum_{i=1}^{n} \left| K\left(\frac{d_\theta\left(\chi, \mathcal{X}_i\right)}{h}\right) - K\left(\frac{d_\theta(\chi^\theta_{k_{(\theta,\chi,\frac{1}{n})}}, \mathcal{X}_i)}{h}\right) \right| \mathbb{I}_i^*,$$

$$F_{12} = \frac{1}{nf(h)} \sum_{i=1}^{n} K\left(\frac{d_\theta\left(\chi, \mathcal{X}_i\right)}{h}\right) 1_{B_\theta(\chi,h) \cap \overline{B_\theta(\chi^\theta_{k_{(\theta,\chi,1/n)}},h)}}\left(\mathcal{X}_i\right),$$

$$F_{13} = \frac{1}{nf(h)} \sum_{i=1}^{n} K\left(\frac{d_\theta(\chi^\theta_{k_{(\theta,\chi,\frac{1}{n})}}, \mathcal{X}_i)}{h}\right) 1_{\overline{B_\theta(\chi,h)} \cap B_\theta\left(\chi^\theta_{k_{\left(\theta,\chi,\frac{1}{n}\right)}},h\right)}\left(\mathcal{X}_i\right),$$

with

$$\mathbb{I}_i^* = 1_{B_\theta(\chi,h) \cap B_\theta\left(\chi^\theta_{k_{\left(\theta,\chi,\frac{1}{n}\right)}},h\right)}\left(\mathcal{X}_i\right).$$

In the case of F_{11}, one can carry out the same steps followed for the case $K(1) = 0$, obtaining the same result:

$$\mathbb{P}\left(F_{11} > \epsilon\sqrt{r_n^*}\right) \leq C p_n^{-C\epsilon^2}. \tag{4.121}$$

In the case of F_{12}, a similar reasoning to that made for the case $K(1) = 0$ allows us to write

$$F_{12} \leq \frac{C}{n} \sum_{i=1}^{n} W_i \quad \text{with} \quad W_i = \frac{1}{f(h)} 1_{B_\theta(\chi,h) \cap \overline{B_\theta(\chi^\theta_{k_{(\theta,\chi,1/n)}},h)}}\left(\mathcal{X}_i\right).$$

Using Assumption (4.12) and the inequality for sums of bounded random variables used before (see Corollary A.9 in Ferraty and Vieu [47]) one has

$$F_{12} = O\left(\frac{1}{nf(h)}\right) + O_{a.co.}\left(\sqrt{\frac{\log(p_n)}{n^2 f(h)^2}}\right). \tag{4.122}$$

The same rate can be stated for the term F_{13}. Then, putting together this results (4.121), (4.122), and using Assumption (4.14), we obtain that there

exists a positive constant, C, such that, for all $\epsilon > 0$ and n large enough

$$\mathbb{P}\left(F_1 > \epsilon\sqrt{r_n^*}\right) \leq C p_n^{-C\epsilon^2}. \tag{4.123}$$

Study of the term F_3.

Focusing now in the term F_3, it is clear that

$$F_3 \leq \mathbb{E}\left(\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\left|\widehat{F}_\theta(\chi) - \widehat{F}_\theta\left(\chi^\theta_{k_{(\theta,\chi,1/n)}}\right)\right|\right).$$

Therefore, following similar steps than in the case of F_1 we obtain that

$$F_3 = O\left(\sqrt{r_n^*}\right). \tag{4.124}$$

Study of the term F_2.

Focusing now on F_2, we have that

$$F_2 \leq F_{21} + F_{22} + F_{23}, \tag{4.125}$$

where we have denoted

$$F_{21} = \sup_{\theta\in\Theta_n}\max_{k\in\left\{1,\ldots,N^\theta_{\mathcal{C},1/n}\right\}}\left|\widehat{F}_\theta(\chi^\theta_k) - \widehat{F}_{\theta^*}\left(\chi^{\theta^*}_{k^*}\right)\right|,$$

$$F_{22} = \sup_{\theta\in\Theta_n}\max_{k\in\left\{1,\ldots,N^\theta_{\mathcal{C},1/n}\right\}}\left|\widehat{F}_{\theta^*}\left(\chi^{\theta^*}_{k^*}\right) - \mathbb{E}\left(\widehat{F}_{\theta^*}\left(\chi^{\theta^*}_{k^*}\right)\right)\right|$$

and

$$F_{23} = \sup_{\theta\in\Theta_n}\max_{k\in\left\{1,\ldots,N^\theta_{\mathcal{C},1/n}\right\}}\left|\mathbb{E}\left(\widehat{F}_{\theta^*}\left(\chi^{\theta^*}_{k^*}\right)\right) - \mathbb{E}\left(\widehat{F}_\theta\left(\chi^\theta_k\right)\right)\right|$$

(for notation, see (4.115)). First, we consider the terms F_{21} and F_{23}. Taking into

account that

$$\begin{aligned}\left|d_\theta\left(\chi_k^\theta,\mathcal{X}_i\right)-d_{\theta^*}\left(\chi_{k^*}^{\theta^*},\mathcal{X}_i\right)\right| &\leq \left|d_\theta\left(\chi_k^\theta,\chi_{k^*}^{\theta^*}\right)+d_\theta\left(\chi_{k^*}^{\theta^*},\mathcal{X}_i\right)-d_{\theta^*}\left(\chi_{k^*}^{\theta^*},\mathcal{X}_i\right)\right| \\ &\leq \left\langle\chi_k^\theta-\chi_{k^*}^{\theta^*},\chi_k^\theta-\chi_{k^*}^{\theta^*}\right\rangle^{1/2}\langle\theta,\theta\rangle^{1/2}+\left|\left\langle\theta-\theta^*,\chi_{k^*}^{\theta^*}-\mathcal{X}_i\right\rangle\right| \\ &\leq d\left(\chi_k^\theta,\chi_{k^*}^{\theta^*}\right)\langle\theta,\theta\rangle^{1/2}+d\left(\chi_{k^*}^{\theta^*},\mathcal{X}_i\right)d\left(\theta,\theta^*\right),\end{aligned}$$

and using assumptions (4.19) and (4.110), together with the condition $\sup_{\theta\in\Theta_n}\langle\theta,\theta\rangle^{1/2}=O(1)$, we obtain that

$$\sup_{\theta\in\Theta_n}\max_{k\in\left\{1,\ldots,N_{\mathcal{C},1/n}^\theta\right\}}\left|d_\theta\left(\chi_k^\theta,\mathcal{X}_i\right)-d_{\theta^*}\left(\chi_{k^*}^{\theta^*},\mathcal{X}_i\right)\right|=O\left(1/n\right).$$

Therefore, similar steps as those used to obtain (4.119) can be followed to get

$$\mathbb{P}\left(F_{21}>\epsilon\sqrt{r_n^*}\right)\leq Cp_n^{-C\epsilon^2}\ \text{ and }\ F_{23}=O\left(\sqrt{r_n^*}\right). \tag{4.126}$$

Finally, we study the term F_{22}. We have that

$$F_{22}=\max_{j\in\left\{1,\ldots,N_{\Theta_n,1/n}\right\}}\max_{k\in\left\{1,\ldots,N_{\mathcal{C},1/n}^{\theta_j}\right\}}\left|\widehat{F}_{\theta_j}\left(\chi_k^{\theta_j}\right)-\mathbb{E}\left(\widehat{F}_{\theta_j}\left(\chi_k^{\theta_j}\right)\right)\right|,$$

and

$$\begin{aligned}&\mathbb{P}\quad\left(F_{22}>\epsilon\sqrt{r_n^*}\right)\\ &\leq\quad N_{\Theta_n,\frac{1}{n}}N_{\mathcal{C},\frac{1}{n}}\max_{j\in\left\{1,\ldots,N_{\Theta_n,\frac{1}{n}}\right\}}\max_{k\in\left\{1,\ldots,N_{\mathcal{C},\frac{1}{n}}^{\theta_j}\right\}}\mathbb{P}\left(\left|\widehat{F}_{\theta_j}\left(\chi_k^{\theta_j}\right)-\mathbb{E}\left(\widehat{F}_{\theta_j}\left(\chi_k^{\theta_j}\right)\right)\right|>\epsilon\sqrt{r_n^*}\right).\end{aligned}$$

Let

$$Z_i=\frac{1}{\mathbb{E}\left(K\left(\frac{d_{\theta_j}\left(\chi_k^{\theta_j},\mathcal{X}\right)}{h}\right)\right)}\left|K\left(\frac{d_{\theta_j}(\chi_k^{\theta_j},\mathcal{X}_i)}{h}\right)-\mathbb{E}\left(K\left(\frac{d_{\theta_j}(\chi_k^{\theta_j},\mathcal{X}_i)}{h}\right)\right)\right|.$$

Using assumptions (4.13), (4.15) and expression (4.118), $Z_i=O(1/f(h))$ and

also $Var(Z_i) = O(1/f(h))$. Therefore, we can use Bernstein-type inequality (see, Corollary A.9 in Ferraty and Vieu [47]) to obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{F}_{\theta_j}\left(\chi_k^{\theta_j}\right) - \mathbb{E}\left(\widehat{F}_{\theta_j}\left(\chi_k^{\theta_j}\right)\right)\right| > \epsilon\sqrt{r_n^*}\right) \\
&= \mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} Z_i\right| > \epsilon\sqrt{r_n^*}\right) \\
&\leq 2\exp\left\{-C\epsilon^2 \log p_n \left(\psi_{\mathcal{C}}\left(\frac{1}{n}\right) + \psi_{\Theta_n}\left(\frac{1}{n}\right)\right)\right\} \\
&\leq C\left(N_{\Theta_n,\frac{1}{n}} N_{\mathcal{C},\frac{1}{n}}\right)^{-C\epsilon^2 \log p_n}.
\end{aligned}$$

Note that in the last inequality we have used that $\psi_{\mathcal{C}}\left(\frac{1}{n}\right) = \log(N_{\mathcal{C},\frac{1}{n}})$ and $\psi_{\Theta_n}\left(\frac{1}{n}\right) = \log(N_{\Theta_n,\frac{1}{n}})$. As a consequence:

$$\mathbb{P}\left(F_{22} > \epsilon\sqrt{r_n^*}\right) \leq C\left(N_{\Theta_n,\frac{1}{n}} N_{\mathcal{C},\frac{1}{n}}\right)^{1-C\epsilon^2 \log p_n}. \tag{4.127}$$

Then, taking into account (4.116), (4.119), (4.123) and (4.124) and putting together (4.125), (4.126) and (4.127), the proof is completed. ■

Lemma 4.15 *Under the assumptions of Lemma 4.14, if in addition assumptions (4.2), (4.17), (4.18) and (4.20) hold, then there exists a positive constant, C, such that, for all $\epsilon > 0$ and n large enough*

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\left|\widehat{g}^*_{j,\theta}(\chi) - \mathbb{E}\left(\widehat{g}^*_{j,\theta}(\chi)\right)\right| > \epsilon\sqrt{r_n^*}\right) \leq C\left(p_n^{-C\epsilon^2} + \left(N_{\Theta_n,1/n} N_{\mathcal{C},1/n}\right)^{1-C\epsilon^2 \log p_n}\right),$$

uniformly on $j = 0, 1, \ldots, p_n$.

Proof of Lemma 4.15. This proof can be easily obtained combining the techniques considered in the proof of Lemma 11 in Ferraty et al. [49] with the decompositions (adapted to the new setting) used in the proof of our Lemma 4.14. ■

Lemma 4.16 *Under assumptions (4.8), (4.13), (4.15), (4.16) and (4.19), if in addition v_n in (4.8) verifies $v_n = O(h)$, we have that*

$$\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\left|\mathbb{E}\left(\widehat{g}^*_{j,\theta}(\chi)\right) - g_{j,\theta_0}(\chi)\right| = O\left(h^{\alpha}\right),$$

uniformly on $j = 0, 1, \ldots, p_n$.

Proof of Lemma 4.16. Firstly, we note that, if $d_\theta(\mathcal{X}, \chi) < h$ holds, then, from the fact that $v_n = O(h)$ together with Assumption (4.19), we have that

$$\begin{aligned} d_{\theta_0}(\mathcal{X}, \chi) &\leq |d_{\theta_0}(\mathcal{X}, \chi) - d_\theta(\mathcal{X}, \chi)| + d_\theta(\mathcal{X}, \chi) = |\langle \mathcal{X} - \chi, \theta_0 - \theta \rangle| + d_\theta(\mathcal{X}, \chi) \\ &\leq \langle \mathcal{X} - \chi, \mathcal{X} - \chi \rangle^{1/2} \langle \theta_0 - \theta, \theta_0 - \theta \rangle^{1/2} + d_\theta(\mathcal{X}, \chi) \leq Ch. \end{aligned} \tag{4.128}$$

Inequalities (4.128) and (4.118) allow to follow the same steps as in the proof of Lemma 10 in Ferraty et al. [49]:

$$\begin{aligned} \left| \mathbb{E}\left(\widehat{g}^*_{j,\theta}(\chi)\right) - g_{j,\theta_0}(\chi) \right| &\leq \left| \frac{1}{\mathbb{E}\left(K(d_\theta(\chi, \mathcal{X})/h)\right)} \mathbb{E}\left(K\left(\frac{d_\theta(\chi, \mathcal{X}_i)}{h} \right) Z_{ij} \right) - g_{j,\theta_0}(\chi) \right| \\ &\leq \frac{1}{\mathbb{E}\left(K(d_\theta(\chi, \mathcal{X})/h)\right)} \mathbb{E}\left(K\left(\frac{d_\theta(\chi, \mathcal{X}_i)}{h} \right) \left| \mathbb{E}\left(Z_{ij} | \langle \theta_0, \mathcal{X} \rangle \right) - g_{j,\theta_0}(\chi) \right| \right) \\ &\leq \frac{1}{\mathbb{E}\left(K(d_\theta(\chi, \mathcal{X})/h)\right)} \mathbb{E}\left(K\left(\frac{d_\theta(\chi, \mathcal{X}_i)}{h} \right) \left| g_{j,\theta_0}(\mathcal{X}) - g_{j,\theta_0}(\chi) \right| \right). \end{aligned}$$

Now using assumptions (4.13) and (4.16), together with (4.128) and (4.118) we obtain

$$\begin{aligned} \left| \mathbb{E}\left(\widehat{g}^*_{j,\theta}(\chi)\right) - g_{j,\theta_0}(\chi) \right| &\leq \frac{C}{\mathbb{E}\left(K\left(\frac{d_\theta(\chi, \mathcal{X})}{h} \right) \right)} \mathbb{E}\left(K\left(\frac{d_\theta(\chi, \mathcal{X}_i)}{h} \right) \right) 1_{B_\theta(\chi, h)}(\mathcal{X}_i) d_{\theta_0}(\mathcal{X}_i, \chi)^\alpha \\ &\leq Ch^\alpha. \blacksquare \end{aligned}$$

Lemma 4.17 *Under assumptions (4.2), (4.7), (4.8), (4.12), (4.13), (4.15)-(4.20) and (4.108)-(4.111), if in addition* v_n *in (4.8) verifies* $v_n = O(h)$ *and* $p_n \to \infty$ *as* $n \to \infty$, *then*

$$\max_{0 \leq j \leq p_n} \sup_{\theta \in \Theta_n} \sup_{\chi \in \mathcal{C}} \{ |\widehat{g}_{j,\theta}(\chi) - g_{j,\theta_0}(\chi)| \} = O_p\left(h^\alpha + \sqrt{r^*_n} \right).$$

Proof of Lemma 4.17. It is verified that

$$\widehat{k}_{j,\theta}(\chi) = \left(\widehat{g}^*_{j,\theta}(\chi) - \mathbb{E}\left(\widehat{g}^*_{j,\theta}(\chi)\right)\right) + \left(\mathbb{E}(\widehat{g}^*_{j,\theta}(\chi)) - g_{j,\theta_0}(\chi)\right) + \left(1 - \widehat{F}_\theta(\chi)\right) g_{j,\theta_0}(\chi), \tag{4.129}$$

where we have denoted

$$\widehat{k}_{j,\theta}(\chi) = \widehat{F}_\theta(\chi)\left(\widehat{g}_{j,\theta}(\chi) - g_{j,\theta_0}(\chi)\right). \tag{4.130}$$

Therefore, from Lemmas 4.14-4.16 together with (4.129), we obtain that there exists a positive constant, C, such that, for all $\epsilon > 0$ and n large enough,

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\left|\widehat{k}_j(\chi)\right| > \epsilon\left(h^\alpha + \sqrt{r_n^*}\right)\right) \leq C\left(p_n^{-C\epsilon^2} + \left(N_{\Theta_n,1/n}N_{\mathcal{C},1/n}\right)^{1-C\epsilon^2\log p_n}\right), \tag{4.131}$$

uniformly on $j = 0, 1, \ldots, p_n$.

In addition, taking Lemma 4.14 into account together with the facts that $r_n^* \to 0$ and $p_n \to \infty$ as $n \to \infty$, we obtain that, for any $0 < \delta < 1$ and n large enough,

$$\begin{aligned}
\mathbb{P}\left(\inf_{\theta\in\Theta_n}\inf_{\chi\in\mathcal{C}}\widehat{F}_\theta(\chi) \geq \delta\right) &\geq 1 - \mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\left|\widehat{F}_\theta(\chi) - 1\right| > 1 - \delta\right) \\
&\geq 1 - \mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\left|\widehat{F}_\theta(\chi) - 1\right| > \delta\epsilon\sqrt{r_n^*}\right) \\
&\geq 1 - C\left(p_n^{-C\delta^2\epsilon^2} + \left(N_{\Theta_n,1/n}N_{\mathcal{C},1/n}\right)^{1-C\delta^2\epsilon^2\log p_n}\right) \geq 1/2.
\end{aligned} \tag{4.132}$$

Now, from (4.131) and (4.132) we have that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\frac{\left|\widehat{k}_j(\chi)\right|}{\widehat{F}_\theta(\chi)} > \epsilon(h^\alpha+\sqrt{r_n^*})\right)\\
\leq\ &\mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\frac{\left|\widehat{k}_j(\chi)\right|}{\widehat{F}_\theta(\chi)} > \epsilon(h^\alpha+\sqrt{r_n^*}) \mid \inf_{\theta\in\Theta_n}\inf_{\chi\in\mathcal{C}}\widehat{F}_\theta(\chi)\geq\delta\right)\\
&+\mathbb{P}\left(\inf_{\theta\in\Theta_n}\inf_{\chi\in\mathcal{C}}\widehat{F}_\theta(\chi)<\delta\right)\\
\leq\ &C\left(p_n^{-C\delta^2\epsilon^2}+\left(N_{\Theta_n,1/n}N_{\mathcal{C},1/n}\right)^{1-C\delta^2\epsilon^2\log p_n}\right). \qquad (4.133)
\end{aligned}$$

Finally, using (4.133) we obtain that,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\leq j\leq p_n}\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\frac{|\widehat{k}_j(\chi)|}{\widehat{F}_\theta(\chi)} > \epsilon(h^\alpha+\sqrt{r_n^*})\right)\\
\leq\ &\sum_{j=0}^{p_n}\mathbb{P}\left(\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\frac{|\widehat{k}_j(\chi)|}{\widehat{F}_\theta(\chi)} > \epsilon(h^\alpha+\sqrt{r_n^*})\right)\\
\leq\ &Cp_n\left(p_n^{-C\delta^2\epsilon^2}+\left(N_{\Theta_n,1/n}N_{\mathcal{C},1/n}\right)^{1-C\delta^2\epsilon^2\log p_n}\right). \qquad (4.134)
\end{aligned}$$

The proof is completed taking into account the notation (4.130) and choosing δ and ϵ in (4.134) such that $C\delta^2\epsilon^2=\beta$ (see (4.109)). ■

Lemma 4.18 *Under assumptions of Lemma 4.17, if in addition* $p_n/n^{\log n}\to 0$ *as* $n\to\infty$, *then*

$$\max_{0\leq j\leq p_n}\sup_{\theta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\{|\widehat{g}_{j,\theta}(\chi)-\widehat{g}_{j,\theta_0}(\chi)|\}=O_p\left(\frac{v_n}{hf(h)}\right).$$

Proof of Lemma 4.18. Let us denote

$$\widehat{F}_\theta^*(\chi)=\frac{\sum_{i=1}^n K(d_\theta(\mathcal{X}_i,\chi)/h)}{nf(h)} \text{ and } \widehat{g}_{j,\theta}^{**}(\chi)=\widehat{F}_\theta^*(\chi)\widehat{g}_{j,\theta}(\chi).$$

It is easy to obtain the decomposition

$$\begin{aligned}\widehat{g}_{j,\theta}(\chi)-\widehat{g}_{j,\theta_0}(\chi) &= \frac{1}{\widehat{F}^*_{\theta}(\chi)}\widehat{g}^{**}_{j,\theta}(\chi)-\frac{1}{\widehat{F}^*_{\theta_0}(\chi)}\widehat{g}^{**}_{j,\theta_0}(\chi)\\ &= \frac{1}{\widehat{F}^*_{\theta}(\chi)}\left(\widehat{g}^{**}_{j,\theta}(\chi)-\widehat{g}^{**}_{j,\theta_0}(\chi)\right)+\widehat{g}^{**}_{j,\theta_0}(\chi)\left(\frac{1}{\widehat{F}^*_{\theta}(\chi)}-\frac{1}{\widehat{F}^*_{\theta_0}(\chi)}\right).\end{aligned} \tag{4.135}$$

Now, we are going to analyze the terms in (4.135). Let us denote

$$Z_{i0}=Y_i \text{ and } Z_{ij}=X_{ij}\ (1\leq j\leq p_n).$$

$$\begin{aligned}\left|\widehat{g}^{**}_{j,\theta}(\chi)-\widehat{g}^{**}_{j,\theta_0}(\chi)\right| &\leq \frac{\sum_{i=1}^n \left|K(d_\theta(\mathcal{X}_i,\chi)/h)-K(d_{\theta_0}(\mathcal{X}_i,\chi)/h)\right||Z_{ij}|}{nf(h)}\\ &\leq \frac{\sum_{i=1}^n C\left|d_\theta(\mathcal{X}_i,\chi)-d_{\theta_0}(\mathcal{X}_i,\chi)\right||Z_{ij}|}{nhf(h)}\\ &\leq \frac{v_n}{hf(h)}\frac{C}{n}\sum_{i=1}^n|Z_{ij}|\\ &= O_p\left(\frac{v_n}{hf(h)}\right),\end{aligned} \tag{4.136}$$

uniformly over $0\leq j\leq p_n$, $\theta\in\Theta_n$ and $\chi\in\mathcal{C}$. The second inequality in (4.136) is a consequence of Assumption (4.15), while assumptions (4.8) and (4.19) give the third inequality. Finally, the equality comes from Assumption (4.20) together with Lemma 4.13 applied to the centred variables $\{|Z_{ij}|-\mathbb{E}(|Z_{ij}|\}_i)$.

In a similar way (considering $Z_{ij}=1$ in (4.136)), one obtains

$$\left|\widehat{F}^*_{\theta}(\chi)-\widehat{F}^*_{\theta_0}(\chi)\right|=O_p\left(\frac{v_n}{hf(h)}\right), \tag{4.137}$$

uniformly over $\theta\in\Theta_n$ and $\chi\in\mathcal{C}$.

Now, we focus on $\widehat{g}^{**}_{0,\theta_0}(\chi)$. Since expression (4.118) is verified, there exist positive

constants, C^* and C'^*, such that

$$C^* \widehat{F}_\theta(\chi) \leq \widehat{F}^*_\theta(\chi) \leq C^{*'} \widehat{F}_\theta(\chi). \tag{4.138}$$

On the one hand, from (4.138) together with Lemma 4.14 we obtain that

$$C^* \left(1 + o_p(1)\right) \leq \widehat{F}^*_\theta(\chi) \leq C^{*'} \left(1 + o_p(1)\right), \tag{4.139}$$

uniformly over $\theta \in \Theta_n$ and $\chi \in \mathcal{C}$. On the other hand, from the uniform convergence of $\widehat{g}_{j,\theta}(\chi)$ to $g_{j,\theta_0}(\chi)$ (see Lemma 4.17) together with the fact that

$$\max_{0\leq j\leq n} \max_{1\leq i\leq n} |g_{j,\theta_0}(\mathcal{X}_i)| = O(1)$$

(see Assumption (4.20)), we obtain that

$$\max_{0\leq j\leq p_n} \sup_{\theta\in\Theta^*_n} \max_{1\leq i\leq n} |\widehat{g}_{j,\theta}(\mathcal{X}_i)| = O_p(1). \tag{4.140}$$

As a consequence of (4.139) and (4.140), we have that

$$\widehat{g}^{**}_{j,\theta}(\chi) = \widehat{F}^*_\theta(\chi)\widehat{g}_{j,\theta}(\chi) = O_p(1), \tag{4.141}$$

uniformly over $0 \leq j \leq p_n$, $\theta \in \Theta_n$ and $\chi \in \mathcal{C}$.

Finally, (4.135), (4.136), (4.137), (4.139) and (4.141) give the result of the lemma. ■

Lemma 4.19 *(Lemma A.4 in Aneiros et al. [7]) Let us assume that* $\boldsymbol{\eta}^\top_{i,\theta_0}$ $(i = 1,\ldots,n)$ *are i.i.d. random vectors. If, in addition,* $\mathbb{E}\left(\eta^4_{\theta_0,1j}\right) < C$ *uniformly on* $1 \leq j \leq p_n$*, then*

$$\boldsymbol{u}^\top \left(\boldsymbol{\eta}^\top_{\theta_0}\boldsymbol{\eta}_{\theta_0} - n\boldsymbol{B}_{\theta_0}\right)\boldsymbol{u} = O_p\left(n^{1/2}p_n\right), \text{ uniformly over } \{\boldsymbol{u} \in \mathbb{R}^{p_n}, ||\boldsymbol{u}|| = M\}.$$

Lemma 4.20 *Let us assume that* $\boldsymbol{\eta}_{i,\theta_0}$ $(i = 1,\ldots,n)$ *are i.i.d. random vectors. Under assumptions (4.7), (4.8), (4.12), (4.13), (4.15), (4.16), (4.19)-(4.21) and (4.108)-(4.111) (*g_{0,θ_0} *and* Y *not included in assumptions (4.16) and (4.20), respectively), if in addition* v_n *in (4.8) verifies* $v_n = O(h)$*,* $p_n \to \infty$*,* $p_n = o(n^{1/2})$*,*

$nh^{4\alpha} = O(1)$ *and*

$$\log^2 p_n = O\left(n\left(\frac{f(h)}{\psi_{\mathcal{C}}(1/n) + \psi_{\Theta_n}(1/n)}\right)^2\right)$$

as $n \to \infty$*, then we have that*

$$\boldsymbol{u}^\top \left(\widetilde{\boldsymbol{X}}_\theta^\top \widetilde{\boldsymbol{X}}_\theta - n\boldsymbol{B}_{\theta_0}\right)\boldsymbol{u} = o_p(n), \text{ uniformly over } \{\boldsymbol{u} \in \mathbb{R}^{p_n}, ||\boldsymbol{u}|| = M\} \text{ and over } \theta \in \Theta_n.$$

Proof of Lemma 4.20. To prove this result, the outline used in proof of Lemma A.5 in Aneiros et al. [7] can be exactly followed, but now our Lemma 4.17 is needed to conclude instead of Lemma A.3 in Aneiros et al. [7]. ■

Lemma 4.21 *(Lemma A.6 in Aneiros et al. [7]) Let us assume that* $\boldsymbol{\eta}_{i,\theta_0 S_n}$ $(i = 1, \ldots, n)$ *are i.i.d. random vectors. If in addition* $s_n^2/n = o(1)$ *and* $\max_{1\le j\le s_n} \mathbb{E}(\eta_{1j,\theta_0}^4) = O(1)$*, then*

$$\left|\left|n^{-1}\boldsymbol{\eta}_{\theta_0 S_n}^\top \boldsymbol{\eta}_{\theta_0 S_n} - \boldsymbol{B}_{\theta_0 S_n \times S_n}\right|\right| = o_p(1),$$

Lemma 4.22 *Let us assume that* $\boldsymbol{\eta}_{i,\theta_0 S_n}$ $(i = 1, \ldots, n)$ *are i.i.d. random vectors. If, in addition, assumptions (4.7), (4.8), (4.12), (4.13), (4.15), (4.16), (4.19), (4.20) and (4.108)-(4.111) hold (but using* s_n *instead of* p_n*, and* g_{0,θ_0} *and* Y *not included in assumptions (4.16) and (4.20), respectively), and* v_n *in (4.8) verifies* $v_n = O(h)$*,* $p_n \to \infty$ *and*

$$\max\left\{h, s_n h^\alpha, s_n^2/n, s_n^2 \log s_n / \left(n\left(\frac{f(h)}{\psi_{\mathcal{C}}(1/n) + \psi_{\Theta_n}(1/n)}\right)\right)\right\} = o(1),$$

then

$$n^{-1}\widetilde{\boldsymbol{X}}_{\theta S_n}^\top \widetilde{\boldsymbol{X}}_{\theta S_n} = \boldsymbol{B}_{\theta_0 S_n \times S_n} + o_p(1), \text{ uniformly over } \theta \in \Theta_n.$$

Proof of Lemma 4.22. The scheme of proof of Lemma A.7 in Aneiros et al. [7] can be exactly followed, taking into account that Lemma A.3 and Lemma A.6 in Aneiros et al. [7] should be replaced by Lemma 4.17 and Lemma 4.21, respectively, of this chapter. ■

Lemma 4.23 *(Lemma A.8 in Aneiros et al. [7]) Let us assume that* $\left(\boldsymbol{\eta}_{i,\theta_0 S_n}^\top, \varepsilon_i\right)$ $(i = 1, \ldots, n)$ *are i.i.d random vectors with mean zero, and* $\{\boldsymbol{\eta}_{i,\theta_0 S_n}\}$ *and* $\{\varepsilon_i\}$ *are*

independent. If, in addition, $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) < C$, $\mathbb{E}(\eta_{1j,\theta_0}) < C$ *uniformly on* $1 \leq j \leq s_n$, *then*

$$\boldsymbol{\varepsilon}^\top \boldsymbol{\eta}_{\theta_0 S_n}^\top \boldsymbol{u} = O_p\left(n^{1/2} s_n^{1/2}\right), \text{ uniformly over } \{\boldsymbol{u} \in \mathbb{R}^{p_n}, ||\boldsymbol{u}|| = M\}.$$

Chapter 5

Contributions on the sparse bi-functional partial linear single-index model

5.1 Introduction

In this chapter, we are going to investigate the situation in which multiple functional predictors are included in the statistical sample. Therefore, we study a new model based on the mixture of partial linear, single-index and sparse ideas, the MFPLSIM briefly presented in Section 1.4.4. For the sake of making the study clearer and easier to follow, we are going to focus the presentation on the bi-functional case. The main idea is to model the effects of each functional covariate in a different way: one of the functional covariates ($\mathcal{X}$) enters in the model through a semiparametric single-index continuous structure; the other one (ζ) enters linearly in the model, but through the p_n-dimensional vector built from its discretized observations. Specifically, the MFPLSIM is given by the expression

$$Y = \sum_{j=1}^{p_n} \beta_{0j}\zeta(t_j) + r\left(\langle\theta_0, \mathcal{X}\rangle\right) + \varepsilon, \tag{5.1}$$

where Y is a real random response and $\mathcal{X}$ denotes a random element belonging to some separable Hilbert space $\mathcal{H}$ with inner product denoted by $\langle \cdot, \cdot \rangle$. The second functional predictor ζ is supposed to be a random curve defined on some interval $[a, b]$ which is observed at the points $a \leq t_1 < \cdots < t_{p_n} \leq b$. $(\beta_{01}, \ldots, \beta_{0p_n})^\top$ is a vector of unknown real coefficients and $r(\cdot)$ denotes a smooth unknown link function. In addition, θ_0 is an unknown functional direction in $\mathcal{H}$ and ε denotes the random error. For identifiability of model (5.1), it is needed to assume conditions (2.2) and (2.3).

As commented in Chapter 1, the main difference between the model dealt in this chapter, MFPLSIM, and that considered in Chapter 4, the SSFPLSIM (4.1), is the fact that, in the SSFPLSIM the covariates with linear effect do not come from a functional variable. In fact, the MFPLSIM has the nice feature to allow incorporating both continuous and point-wise effects of functional variables, involving interpretable parameters in both cases. Furthermore, one has to take into account that we have very big number, p_n, of linear covariates, while only a few of them could possess a real influence on the response. As in the previous chapter, we will denote the set of indices corresponding with the relevant variables as $S_n = \{j = 1, \ldots, p_n, \text{ such that } \beta_{0j} \neq 0\}$, and $s_n = \sharp S_n$ where the notation $\sharp A = card(A)$[1] was used. Therefore, this flexible model needs to be combined with an accurate variable selection method.

The problem is that the application of the standard PLS method presented in Chapter 4 to the MFPLSIM becomes dramatically infeasible. That is due to the huge computational time required by the PLS method to perform the variable selection even for moderate values of p_n. In addition, standard procedures, coming from an adaptation of the multivariate methodology to FDA, do not take into account the strong correlation structure present between linear covariates because of its functional origin (although there exist in the statistical literature some proposals to select covariates in linear models with features that can be ordered in some meaningful way: group LASSO, see Bakin [14]; fused LASSO, see Tibshirani et al. [105]; among others). Accordingly, we are going to develop two new algorithms for variable selection (in the linear component) and estimation of the MFPLSIM, which take advantage of the functional origin of these scalar variables included in the linear

[1]For the sake of brevity, through this chapter this notation will be used.

component of the model. In both algorithms, the MFPLSIM will be transformed in certain linear regression model in which the correlation between covariates is attenuated, and then some standard PLS procedure is applied. For that, one could consider some of the penalties briefly described in Section 1.4.1. As in the previous chapter, we will use the SCAD penalty (1.10), which enjoys the oracle property (see Fan and Li [38]).

Focusing now on the interest of the MFPLSIM in practice, a nice example can be developed for chemometrics. For this field of applied sciences, functional data analysis is traditionally of great interest (see Ferraty and Vieu [47] and references therein). This example we will based on Sugar data presented in Section 1.1. In Sugar data, for each sample of sugar, the absorbance spectra from 275 to 560 nm was measured in 0.5 nm intervals (therefore, $p_n = 571$) at excitation wavelengths 240 nm (ζ) and at excitation wavelengths 290 nm ($\mathcal{X}$). Samples of both curves can be seen in Figure 1.3. The ash content of each sugar sample, Y, was also determined and the practical question is how the value of ash content for a new sample can be predicted just by looking at its spectrometric curves. This is a typical regression problem with scalar response (the ash content Y). At this stage, there are two alternative ways to deal with the spectrometric data:

i) Consider data as p_n-dimensional vectors compound of the values observed at the discretized wavelengths.

ii) Consider data as curves obtained by smoothing the discretized observations.

Undoubtedly both approaches may have its own advantages and/or drawbacks, but it should be noted that both have to face the same dimensionality problem. Option i) leads to a multivariate regression analysis with number of variables ($2p_n = 1142$ variables) that is much larger than the sample size itself ($n = 268$). Moreover, there is a specific additional difficulty linked with the high correlation between variables. To deal with this discretized point of view, it is required to develop a suitable adaptation of the techniques used in Big Data Analysis, such as sparse modelling. Alternatively, Option ii) involves only two predictors (namely, the curves ζ and $\mathcal{X}$) but each of them are elements of an infinite-dimensional space. To deal with this continuous point of view, it is necessary to develop a suitable adaptation of regression techniques, which

should be insensitive to the dimensionality of the covariates, such as semiparametric modelling.

All in all, given the complexity of the sample, it is strongly recommended to analyse such spectrometric data in a way as flexible as possible. The methodology developed in this chapter achieves this goal by building a model that is a mixture of Options i) and ii) and providing statistical methods that combine both sparse and semiparametric ideas.

This chapter will be organized as follows. In Section 5.2 we will focus on the sparse feature of the MFPLSIM (5.1) and we will present two new variable selection algorithms. The first method is a fast algorithm which provides the variable selection and estimation of the MFPLSIM in a reasonable amount of time, even for very big values of p_n. The second variable selection procedure is a more refined method which adds a second step to the fast algorithm, allowing to complete and precise the set of relevant variables selected by the fast method (the second algorithm is an adaptation to model MFPLSIM of the PVS procedure presented in Aneiros and Vieu [4]). This In Section 5.3 a wide scope of asymptotics is obtained for giving mathematical support to both procedures. In Section 5.4, finite sample simulated experiments in two different scenarios will allow us to compare computational time and prediction accuracy of the fast algorithm and the PLS procedure and of the fast algorithm and the second proposed method, respectively. In the second scenario, the precision of impact point selection will be also quantified. Some conclusions will be derived, providing the scope of application of the proposed algorithms in practice. In Section 5.5 the real data application presented in Section 5.1 will be analysed by means of the proposed methods. That will be the opportunity to illustrate the triple interest of our methodology: high predictive power, interpretable outputs and reasonably low computational time. The proofs of theoretical results are reported in Section 5.6.

5.2 The algorithms

In this section we are going to present two new algorithms for variable selection in the specific setting of the MFPLSIM (5.1). Both methods take advantage of the

following fact: in the multivariate case, more variables in the linear part means, in general, more different external information about the response; as a contrast, when linear covariates have functional origin, with bigger p_n we obtain more precise information about the single continuous process that generates the curve ζ to be discretized.

5.2.1 The FASSMR algorithm

In practice, the progress in measurement technologies leads to a huge number of discretizations of the functional variable ζ. It is well-known that for any standard variable selection method (such as the PLS presented in Chapter 4, for instance), the larger p_n, the more computational time is required. Therefore, a crucial point for producing results in reasonable time is to provide algorithms saving as much as computational time as possible.

This fact leads us to propose the following *fast algorithm for sparse semiparametric multi-functional regression* (FASSMR). The main idea of this algorithm is to consider a reduced model, with only some (very few) linear covariates (but covering the entire discretization interval of ζ), and discarding directly the other linear covariates (since one could expect that they contain very similar information about the response). This idea is described below.

For introducing the variable selection algorithm, as usual, assume that we have a statistical sample of size n:

$$\{(\zeta_i, \mathcal{X}_i, Y_i)\}_{i=1}^n \quad \text{i.i.d. as } (\zeta, \mathcal{X}, Y). \tag{5.2}$$

verifying model MFPLSIM (5.1); that is

$$Y_i = \sum_{j=1}^{p_n} \beta_{0j}\zeta_i(t_j) + r\left(\langle\theta_0, \mathcal{X}_i\rangle\right) + \varepsilon_i, \quad (i = 1, \ldots, n).$$

We assume, without lost of generality, that the number p_n of linear covariates can be expressed as follows: $p_n = q_n w_n$ with q_n and w_n integers. The previous consideration allows us to present a subset of the initial p_n linear covariates, which will contain only w_n equally spaced discretized observations of ζ covering the whole interval $[a, b]$.

This subset will be the following:

$$\mathcal{R}_n^{\mathbf{1}} = \left\{ \zeta\left(t_k^{\mathbf{1}}\right), \quad k = 1, \ldots, w_n \right\}, \tag{5.3}$$

where $t_k^{\mathbf{1}} = t_{[(2k-1)q_n/2]}$ and $[z]$ denotes the smallest integer not less than $z \in \mathbb{R}$.

It is noteworthy that correlation between consecutive variables within $\mathcal{R}_n^{\mathbf{1}}$ is much less important than in the whole set of p_n initial linear covariates. Therefore, it is expected that the standard PLS method will work better applied to variables in $\mathcal{R}_n^{\mathbf{1}}$ than applied to the full set of p_n linear covariates. Moreover, we hope that computational time will be greatly reduced if we use moderate values for w_n. As a consequence, the standard PLS variable selection procedure will be carried out among variables in $\mathcal{R}_n^{\mathbf{1}}$.

In this way, we are going to consider the following reduced model, which involves only linear covariates belonging to $\mathcal{R}_n^{\mathbf{1}}$:

$$Y_i = \sum_{k=1}^{w_n} \beta_{0k}^{\mathbf{1}} \zeta_i(t_k^{\mathbf{1}}) + r^{\mathbf{1}}\left(\left\langle \theta_0^{\mathbf{1}}, \mathcal{X}_i \right\rangle\right) + \varepsilon_i^{\mathbf{1}}. \tag{5.4}$$

For this model, the set of relevant indices and its estimation can be denoted by

$$\mathcal{S}_n^{\mathbf{1}} = \{k = 1, \ldots, w_n, \quad \beta_{0k}^{\mathbf{1}} \neq 0\},$$

$$\widehat{\mathcal{S}}_n^{\mathbf{1}} = \{k = 1, \ldots, w_n, \quad \widehat{\beta}_{0k}^{\mathbf{1}} \neq 0\},$$

with $s_n^{\mathbf{1}} = \sharp(\mathcal{S}_n^{\mathbf{1}})$ and $\widehat{s}_n^{\mathbf{1}} = \sharp\left(\widehat{\mathcal{S}}_n^{\mathbf{1}}\right)$. In addition, it is assumed that

$$\exists c, \quad \forall n, \quad \inf_n \min_{k \in \mathcal{S}_n^{\mathbf{1}}} \left|\beta_{0k}^{\mathbf{1}}\right| > c > 0. \tag{5.5}$$

Then, the variable selection task can be developed following the next steps:

1. The first idea is to transform model (5.4) into a linear model, by extracting from Y_i and $\zeta_i(t_k^{\mathbf{1}})$ $(k = 1, \ldots, w_n)$ the effect of the functional variable $\mathcal{X}_i$ when is projected on the direction $\theta_0^{\mathbf{1}}$. Specifically,

$$Y_i - \mathbb{E}\left(Y_i | \left\langle \theta_0^{\mathbf{1}}, \mathcal{X}_i \right\rangle\right) = \sum_{k=1}^{w_n} \beta_{0k}^{\mathbf{1}} \left(\zeta_i(t_k^{\mathbf{1}}) - \mathbb{E}\left(\zeta_i(t_k^{\mathbf{1}}) | \left\langle \theta_0^{\mathbf{1}}, \mathcal{X}_i \right\rangle\right)\right) + \varepsilon_i^{\mathbf{1}}. \tag{5.6}$$

Since in expression (5.6) the conditional expectations are unknown, they may be estimated by means of regression. Nadaraya-Watson kernel estimators are used for this task. Then, we are going to consider the following $n \times n$-matrix of local weights:

$$\boldsymbol{W}_{h,\theta} = \left(w_{n,h,\theta}(\mathcal{X}_i, \mathcal{X}_\ell)\right)_{i,\ell=1,\ldots,n},$$

where $w_{n,h,\theta}(\chi, \mathcal{X}_i)$ was defined in (2.11). As a result, we will obtain the following transformed variables, for each $\theta \in \mathcal{H}$:

$$\widetilde{\boldsymbol{Y}}_\theta = (\boldsymbol{I} - \boldsymbol{W}_{h,\theta})\boldsymbol{Y}, \quad \widetilde{\boldsymbol{\zeta}}^{\mathbf{1}}_\theta = (\boldsymbol{I} - \boldsymbol{W}_{h,\theta})\boldsymbol{\zeta}^{\mathbf{1}},$$

where $\boldsymbol{\zeta}^{\mathbf{1}}$ is the $n \times w_n$ matrix $(\zeta_i(t^{\mathbf{1}}_k), 1 \leq i \leq n, 1 \leq k \leq w_n)$, and $\boldsymbol{Y}$ the is the vector of responses $(Y_1, \ldots, Y_n)^\top$.

2. The standard PLS variable selection procedure is applied among the set $\mathcal{R}^{\mathbf{1}}_n$. Specifically, the penalized profile least squares function is minimized over the pair $(\boldsymbol{\beta}^{\mathbf{1}}, \theta^{\mathbf{1}})$ with $\boldsymbol{\beta}^{\mathbf{1}} \in \mathbb{R}^{w_n}$ and $\theta^{\mathbf{1}} \in \Theta^{\mathbf{1}}_n \subset \mathcal{H}$:

$$\mathcal{Q}^{\mathbf{1}}\left(\boldsymbol{\beta}^{\mathbf{1}}, \theta^{\mathbf{1}}\right) = \frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_{\theta^{\mathbf{1}}} - \widetilde{\boldsymbol{\zeta}}^{\mathbf{1}}_{\theta^{\mathbf{1}}}\boldsymbol{\beta}^{\mathbf{1}}\right)^\top \left(\widetilde{\boldsymbol{Y}}_{\theta^{\mathbf{1}}} - \widetilde{\boldsymbol{\zeta}}^{\mathbf{1}}_{\theta^{\mathbf{1}}}\boldsymbol{\beta}^{\mathbf{1}}\right) + n\sum_{k=1}^{w_n} \mathcal{P}_{\lambda_{k_n}}\left(|\beta^{\mathbf{1}}_k|\right), \tag{5.7}$$

being $\mathcal{P}_{\lambda_{k_n}}(\cdot)$ the SCAD penalty function defined in (1.10).

3. We denote by $(\widehat{\boldsymbol{\beta}}^{\mathbf{1}}_0, \widehat{\theta}^{\mathbf{1}}_0)$ a local minimizer of the criterion $\mathcal{Q}^{\mathbf{1}}(\cdot,\cdot)$, where $\widehat{\boldsymbol{\beta}}^{\mathbf{1}}_0 = (\widehat{\beta}^{\mathbf{1}}_{01}, \ldots, \widehat{\beta}^{\mathbf{1}}_{0w_n})^\top$. Then, $\zeta(t^{\mathbf{1}}_k)$ is selected in $\mathcal{R}^{\mathbf{1}}_n$ if, and only if, $\widehat{\beta}^{\mathbf{1}}_{0k} \neq 0$.

Remark 5.1 *As expected, to obtain asymptotic results related to the presented variable selection algorithm (FASSMR), two kinds of assumptions should be considered. On the one hand, specific assumptions to treat with covariates with linear effect coming from the discretization of a curve (functional nature of the linear covariates). On the other hand, general assumptions to deal with the standard PLS procedure. Both kinds of assumptions will be shown in Section 5.3 (see forthcoming conditions (5.18)-(5.21) and (5.22), (5.23), respectively). We should emphasize that assumptions related to the standard PLS procedure (conditions (5.22) and (5.23)) will be formulated in a rather general form. In that way, different sets of hypotheses could*

give rise to these assumptions. For instance, in Chapter 4, we can find conditions under which (5.22) and (5.23) hold. In addition, in Section 4.3 (i) the existence of a local minimizer, $(\widehat{\boldsymbol{\beta}}_0^{\mathbf{1}}, \widehat{\theta}_0^{\mathbf{1}})$, of $\mathcal{Q}^{\mathbf{1}}(\cdot,\cdot)$ is ensured, (ii) the corresponding convergence rates are obtained, and (iii) the subset of eligible directions, $\Theta_n^{\mathbf{1}}$, is characterized theoretically (practical considerations about $\Theta_n^{\mathbf{1}}$ are included in Section 5.4.1.2). In addition, specific requirements for a general penalty function $\mathcal{P}_\lambda(\cdot)$ are included. These assumptions are satisfied by the SCAD penalty used throughout this dissertation. Finally, note that to treat with partial linear single-index models, it is frequent to impose some additional assumption to ensure identifiability. Such assumption links the two kinds of covariates in the model (ζ and $\mathcal{X}$, in the case of our MFPLSIM), and it prevents the possibility that covariates with different types of effect (linear or semiparametric) are equal (see, for instance, condition (vi) in Liang et al. [75] and Condition (4.23) in the previous chapter for cases of scalar and functional covariates, respectively). In this chapter, Condition (4.23) is implicitly assumed (as far as we know, the investigation in Chapter 4 is the only work in the statistical literature dealing with penalized variable selection in sparse semi-functional partial linear single-index regression).

5.2.1.1 The outputs of the FASSMR algorithm

Once the variable selection procedure is carried out, the parameters of the model can be estimated. Then, coming back to model (5.1) and considering the whole set of initial of p_n linear covariates, a variable $\zeta(t_j) \in \{\zeta(t_1), \ldots, \zeta(t_{p_n})\}$ is selected if, and only if, it belongs to $\mathcal{R}_n^{\mathbf{1}}$ and its estimated coefficient, which can be denoted by $\widehat{\beta}_{0k_j}^{\mathbf{1}}$, is non null. Therefore, the following estimated set of relevant variables is obtained:

$$\widehat{S}_n = \left\{ j = 1, \ldots, p_n, \quad \text{such that } t_j = t_{k_j}^{\mathbf{1}} \text{ with } \zeta(t_{k_j}^{\mathbf{1}}) \in \mathcal{R}_n^{\mathbf{1}} \text{ and } \widehat{\beta}_{0k_j}^{\mathbf{1}} \neq 0 \right\}.$$

In addition, a natural choice for the estimates of the linear coefficients and of θ_0 is to use the estimates obtained from the variable selection procedure. That is,

$$\widehat{\beta}_{0j} \quad = \quad \begin{cases} \widehat{\beta}_{0k_j}^{\mathbf{1}} & \text{if } j \in \widehat{\mathcal{S}}_n, \\ 0 & \text{otherwise,} \end{cases}$$

$$\widehat{\theta}_0 \;=\; \widehat{\theta}_0^{\mathbf{1}}.$$

Finally, denoting by $\widehat{\boldsymbol{\beta}}_0$ the vector of estimated parameters, an estimator of the function $r_{\theta_0}(\cdot) \equiv r(\langle \theta_0, \chi \rangle)$ can be obtained by smoothing the residuals of the parametric fit:

$$\widehat{r}_{\widehat{\theta}_0}(\chi) \equiv \widehat{r}\left(\left\langle \widehat{\theta}_0, \chi \right\rangle\right) = \frac{\sum_{i=1}^{n}\left(Y_i - \boldsymbol{\zeta}_i^{\top}\widehat{\boldsymbol{\beta}}_0\right) K\left(d_{\widehat{\theta}_0}\left(\chi, \mathcal{X}_i\right)/h\right)}{\sum_{i=1}^{n} K\left(d_{\widehat{\theta}_0}\left(\chi, \mathcal{X}_i\right)/h\right)}, \tag{5.8}$$

where we have denoted $\boldsymbol{\zeta}_i = (\zeta_i(t_1), \ldots, \zeta_i(t_{p_n}))^{\top}$. Note that the estimation of $r_{\theta_0}(\cdot)$ is the obtained for $r^{\mathbf{1}}_{\theta_0^{\mathbf{1}}}(\cdot)$. In other words, $\widehat{r}_{\widehat{\theta}_0}(\chi) = \widehat{r}^{\mathbf{1}}_{\widehat{\theta}_0^{\mathbf{1}}}(\chi)$.

Remark 5.2 *Once we have presented our FASSMR algorithm, we will make some comments about the design points (t_j, $j = 1, \ldots, p_n$) over which the curve ζ is discretized and about the theoretical or computational complexity of the algorithm. Focusing on the design, note that, for the sake of simplicity, it was assumed an equispaced grid. Actually, that is not restrictive in practice. In fact, if data are unbalanced one can (as a first stage) smooth each observed curve, and then compute the smoothed curves at some regularly spaced points to create a new (balanced) curves dataset. Anyway, it should be noted that our results hold if the assumption of equispaced grid is changed by a grid $a \leq t_1 < \cdots < t_{p_n} \leq b$ supposed to be regular in the sense: $\exists c_1, c_2$ such that $\forall j = 1, \ldots, p_n - 1,\ 0 < c_1 p_n^{-1} < t_{j+1} - t_j < c_2 p_n^{-1} < \infty$. Focusing on the theoretical or computational complexity of the algorithm, we should take into account: (i) the construction of the linear models for the application of the variable selection procedure (that is, the estimate of the conditional expectations in (5.6)), and (ii) the application of the variable selection procedure to such linear models. For a fixed value $\theta \in \Theta_n^{\mathbf{1}}$ (see the definition of $\Theta_n^{\mathbf{1}}$ in Remark 5.1), and given tuning parameters h, w_n and λ, the computational complexity for (i) is $O(n^2 w_n)$, while for (ii) the computational complexity of the more computationally efficient algorithm we know is $O(n w_n)$ (see Shi et al. [102]). Therefore, the computational complexity of the proposed FASSMR algorithm is $O(n^2 w_n \sharp \Theta_n^{\mathbf{1}})$. Moreover, for the standard PLS procedure (see Chapter 4) such complexity is $O(n^2 p_n \sharp \Theta_n)$ (Θ_n is the set of eligible directions θ for estimating the full model (5.1); usually, $\Theta_n = \Theta_n^{\mathbf{1}}$). Then, it is expected that, in practice, when $w_n << p_n$, our algorithm will be much*

faster than the standard one (especially when $w_n << n << p_n$). Finally, note that the factor n^2 in the orders above is more a consequence of model complexity than of algorithm complexity (specifically, it is due to the presence of the nonparametric component $r(\cdot)$ in the MFPLSIM; that is, if $r(\cdot)$ were known, n^2 should be replaced by n).

5.2.2 IASSMR: A refined variable selection algorithm

From the previous section we can derive that the FASSMR provides an important computational time saving if we compare it with the direct application of standard procedures (like the PLS method). However, since the algorithm is based on directly discarding variables, the price for this big improvement in efficiency is that the set of relevant variables could not be exactly obtained in many contexts.

Therefore, the natural question is whether we could propose an additional algorithm to solve this problem. The new method should take into account the functional origin of the linear covariates and should be able to select a more precise set of impact points, but without destroying the main features (in particular, its fast implementation) of the FASSMR.

Following these principles, we present in this section the *improved algorithm for sparse semiparametric regression* (IASSMR). Roughly speaking, the idea of the IASSMR is to add a second stage, which takes into account the q_n variables in the neighbourhood of the selected in the first stage by the FASSMR. Then, a second variable selection procedure is applied among this new set of variables. This idea is described below.

For developing the IASSMR, the sample (5.2) is split into two independent subsamples asymptotically of the same size $n_1 \sim n_2 \sim n/2$. One of them will be used in the first stage of the method and the other, in the second stage:

$$\mathcal{E}^{\mathbf{1}} = \{(\zeta_i, \mathcal{X}_i, Y_i), \quad i = 1, \ldots, n_1\},$$

$$\mathcal{E}^{\mathbf{2}} = \{(\zeta_i, \mathcal{X}_i, Y_i), \quad i = n_1 + 1, \ldots, n_1 + n_2 = n\}.$$

From now on, the superscript $\boldsymbol{s}$ with $\boldsymbol{s} = \mathbf{1}, \mathbf{2}$ indicates the stage of the method in which the sample, function, variable or parameter is involved.

First stage. The FASSMR is applied, but now using only the subsample $\mathcal{E}^{\mathbf{1}}$:

1. The variable selection procedure is started among variables belonging to $\mathcal{R}_n^{\mathbf{1}}$, see (5.3). The MFPLSIM is transformed into a linear model as in (5.6). Note that since we only use $\mathcal{E}^{\mathbf{1}}$, we obtain a $n_1 \times n_1$-matrix of local weights $\boldsymbol{W}_{h,\theta}^{\mathbf{1}} = \left(w_{n_1,h,\theta}(\mathcal{X}_i, \mathcal{X}_\ell)\right)_{i,\ell=1,\ldots,n_1}$, where $w_{n_1,h,\theta}(\chi, \mathcal{X}_i)$ was defined in (2.11). In addition, for each $\theta \in \mathcal{H}$, we denote $\widetilde{\boldsymbol{Y}}_\theta^{\mathbf{1}} = \left(\boldsymbol{I} - \boldsymbol{W}_{h,\theta}^{\mathbf{1}}\right)\boldsymbol{Y}^{\mathbf{1}}$, with $\boldsymbol{Y}^{\mathbf{1}} = (Y_1, \ldots, Y_{n_1})^\top$ and $\widetilde{\boldsymbol{\zeta}}_\theta^{\mathbf{1}} = \left(\boldsymbol{I} - \boldsymbol{W}_{h,\theta}^{\mathbf{1}}\right)\boldsymbol{\zeta}^{\mathbf{1}}$, where, abusing of notation, we denote by $\boldsymbol{\zeta}^{\mathbf{1}}$ the $n_1 \times w_n$ matrix $(\zeta_i(t_k^{\mathbf{1}}), 1 \leq i \leq n_1, 1 \leq k \leq w_n)$.

2. The standard PLS variable selection procedure is applied within the set $\mathcal{R}_n^{\mathbf{1}}$ by minimizing the penalized least squares criterion over the pair $(\boldsymbol{\beta}^{\mathbf{1}}, \theta^{\mathbf{1}})$, with $\boldsymbol{\beta}^{\mathbf{1}} \in \mathbb{R}^{w_n}$ and $\theta^{\mathbf{1}} \in \Theta_n^{\mathbf{1}}$:

$$\mathcal{Q}^{\mathbf{1}}\left(\boldsymbol{\beta}^{\mathbf{1}}, \theta^{\mathbf{1}}\right) = \frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_{\theta^{\mathbf{1}}}^{\mathbf{1}} - \widetilde{\boldsymbol{\zeta}}_{\theta^{\mathbf{1}}}^{\mathbf{1}}\boldsymbol{\beta}^{\mathbf{1}}\right)^\top \left(\widetilde{\boldsymbol{Y}}_{\theta^{\mathbf{1}}}^{\mathbf{1}} - \widetilde{\boldsymbol{\zeta}}_{\theta^{\mathbf{1}}}^{\mathbf{1}}\boldsymbol{\beta}^{\mathbf{1}}\right) + n_1 \sum_{k=1}^{w_n} \mathcal{P}_{\lambda_{kn}}\left(|\beta_k^{\mathbf{1}}|\right). \tag{5.9}$$

3. We obtain $(\widehat{\boldsymbol{\beta}}_0^{\mathbf{1}}, \widehat{\theta}_0^{\mathbf{1}})$ by minimizing (5.9); then, $\zeta(t_k^{\mathbf{1}})$ is selected in $\mathcal{R}_n^{\mathbf{1}}$ if, and only if, $\widehat{\beta}_{0k}^{\mathbf{1}} \neq 0$.

Second stage. Variables in the neighbourhood of the ones selected in the first stage are included. Then the PLS procedure is carried out again. For that, we consider only the subsample $\mathcal{E}^{\mathbf{2}}$. Specifically:

1. A new set of variables is considered:

$$\mathcal{R}_n^{\mathbf{2}} = \bigcup_{\left\{k, \widehat{\beta}_{0k}^{\mathbf{1}} \neq 0\right\}} \left\{\zeta(t_{(k-1)q_n+1}), \ldots, \zeta(t_{kq_n})\right\}.$$

   Denoting by $r_n = \sharp(\mathcal{R}_n^{\mathbf{2}})$, we can rename the variables in $\mathcal{R}_n^{\mathbf{2}}$ as follows:

$$\mathcal{R}_n^{\mathbf{2}} = \left\{\zeta(t_1^{\mathbf{2}}), \ldots, \zeta(t_{r_n}^{\mathbf{2}})\right\},$$

and consider the following model

$$Y_i = \sum_{k=1}^{r_n} \beta_{0k}^{\mathbf{2}} \zeta_i(t_k^{\mathbf{2}}) + r^{\mathbf{2}}\left(\left\langle \theta_0^{\mathbf{2}}, \mathcal{X}_i \right\rangle\right) + \varepsilon_i^{\mathbf{2}}. \tag{5.10}$$

2. As in the first stage, model (5.10) is transformed into a linear model in the same way as in (5.6):

$$Y_i - \mathbb{E}\left(Y_i | \left\langle \theta_0^{\mathbf{2}}, \mathcal{X}_i \right\rangle\right) = \sum_{k=1}^{r_n} \beta_{0k}^{\mathbf{2}} \left(\zeta_i(t_k^{\mathbf{2}}) - \mathbb{E}\left(\zeta_i(t_k^{\mathbf{2}}) | \left\langle \theta_0^{\mathbf{2}}, \mathcal{X}_i \right\rangle\right)\right) + \varepsilon_i^{\mathbf{2}}, \tag{5.11}$$

but now we use the subsample $\mathcal{E}^{\mathbf{2}}$ for obtaining the estimator of the conditional expectations. Therefore, $\boldsymbol{W}_{h,\theta}^{\mathbf{2}}$ is obtained analogously to $\boldsymbol{W}_{h,\theta}^{\mathbf{1}}$ but employing $\mathcal{E}^{\mathbf{2}}$ instead of $\mathcal{E}^{\mathbf{1}}$. As in the previous stage, for each $\theta \in \mathcal{H}$, $\widetilde{\boldsymbol{Y}}_\theta^{\mathbf{2}} = \left(\boldsymbol{I} - \boldsymbol{W}_{h,\theta}^{\mathbf{2}}\right)\boldsymbol{Y}^{\mathbf{2}}$ with $\boldsymbol{Y}^{\mathbf{2}} = (Y_{n_1+1}, \ldots, Y_n)^\top$ and $\widetilde{\boldsymbol{\zeta}}_\theta^{\mathbf{2}} = \left(\boldsymbol{I} - \boldsymbol{W}_{h,\theta}^{\mathbf{2}}\right)\boldsymbol{\zeta}^{\mathbf{2}}$ with $\boldsymbol{\zeta}^{\mathbf{2}}$ the $n_2 \times r_n$ matrix $\left(\zeta_i(t_k^{\mathbf{2}}), n_1 + 1 \leq i \leq n, 1 \leq k \leq r_n\right)$, and $\boldsymbol{\beta}^{\mathbf{2}} = (\beta_1^{\mathbf{2}}, \ldots, \beta_{r_n}^{\mathbf{2}})^\top$.

3. The PLS procedure is applied again, but now within the set $\mathcal{R}_n^{\mathbf{2}}$ by minimizing the profile least squares function over the pair $(\boldsymbol{\beta}^{\mathbf{2}}, \theta^{\mathbf{2}})$, with $\boldsymbol{\beta}^{\mathbf{2}} \in \mathbb{R}^{r_n}$ and $\theta^{\mathbf{2}} \in \Theta_n^{\mathbf{2}} \subset \mathcal{H}$:

$$\mathcal{Q}^{\mathbf{2}}\left(\boldsymbol{\beta}^{\mathbf{2}}, \theta^{\mathbf{2}}\right) = \frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_{\theta^{\mathbf{2}}} - \widetilde{\boldsymbol{\zeta}}_{\theta^{\mathbf{2}}}^{\mathbf{2}}\boldsymbol{\beta}^{\mathbf{2}}\right)^\top \left(\widetilde{\boldsymbol{Y}}_{\theta^{\mathbf{2}}} - \widetilde{\boldsymbol{\zeta}}_{\theta^{\mathbf{2}}}^{\mathbf{2}}\boldsymbol{\beta}^{\mathbf{2}}\right) + n_2 \sum_{k=1}^{r_n} \mathcal{P}_{\lambda_{k_n}}\left(|\beta_k^{\mathbf{2}}|\right). \tag{5.12}$$

4. The minimizer of the criterion $\mathcal{Q}^{\mathbf{2}}(\cdot,\cdot)$ is denoted by $\left(\widehat{\boldsymbol{\beta}}_0^{\mathbf{2}}, \widehat{\theta}_0^{\mathbf{2}}\right)$. At the end of the second stage, $\zeta(t_k^{\mathbf{2}})$ is selected in $\mathcal{R}_n^{\mathbf{2}}$ if the associated coefficient, $\widehat{\beta}_{0k}^{\mathbf{2}}$, is non-null.

Remark 5.3 *Theoretical considerations for subsets $\Theta_n^{\mathbf{2}}$ and $\Theta_n^{\mathbf{1}}$ and local-minimizer existence in the IASSMR are the same as those given in the Remark 5.1 for the FASSMR.*

5.2.2.1 The outputs of IASSMR algorithm

At the end of this two-stage procedure, a variable $\zeta(t_j) \in \{\zeta(t_1), \ldots, \zeta(t_{p_n})\}$ is selected if and only if belongs to $\mathcal{R}_n^{\mathbf{2}}$ and its estimated coefficient in the second stage, said $\widehat{\beta}_{0k_j}^{\mathbf{2}}$, is non-null. Therefore, the following estimated set of relevant variables is obtained:

$$\widehat{S}_n = \left\{ j = 1, \ldots, p_n, \text{ such that } t_j = t_{k_j}^{\mathbf{2}}, \text{ with } \zeta(t_{k_j}^{\mathbf{2}}) \in \mathcal{R}_n^{\mathbf{2}} \text{ and } \widehat{\beta}_{0k_j}^{\mathbf{2}} \neq 0 \right\}. \quad (5.13)$$

In this case, a natural way to obtain estimates of the linear coefficients and of the direction θ_0 is to use the estimates from the second stage of the algorithm. That is:

$$\widehat{\beta}_{0j} \;=\; \begin{cases} \widehat{\beta}_{0k_j}^{\mathbf{2}} & \text{if } j \in \widehat{S}_n, \\ 0 & \text{otherwise,} \end{cases} \quad (5.14)$$

$$\widehat{\theta}_0 \;=\; \widehat{\theta}_0^{\mathbf{2}}. \quad (5.15)$$

If we denote by $\widehat{\boldsymbol{\beta}}_0$ the vector of estimated linear coefficients, an estimator of the function $r_{\theta_0}(\cdot) \equiv r(\langle \theta_0, \chi \rangle)$ can be obtained by smoothing the residuals of the linear component as in (5.8), but now with $\widehat{\beta}_{0j}$ and $\widehat{\theta}_0$ obtained as in (5.14) and (5.15), respectively. In other words, $\widehat{r}_{\widehat{\theta}_0}(\chi) = \widehat{r}_{\widehat{\theta}_0^{\mathbf{2}}}^{\mathbf{2}}(\chi)$.

Remark 5.4 *Once we have presented our IASSMR algorithm, we will make some comments about both the considered subsamples ($\mathcal{E}^{\mathbf{1}}$ and $\mathcal{E}^{\mathbf{2}}$ for first and second stages, respectively) and the theoretical or computational complexity of the algorithm. Focusing on $\mathcal{E}^{\mathbf{1}}$ and $\mathcal{E}^{\mathbf{2}}$, note that since they are different (and therefore independent) the bias of selection is avoided and proofs of our asymptotic results are greatly facilitated. In addition, although in our general presentation we choose $n_1 \sim cn$ and $n_2 \sim n - n_1$ for $c = 1/2$ (maybe the natural choice), it could be considered any value $0 < c < 1$ (this fact does not affect the asymptotic properties while, in some scenarios as those where the sample size (n) is too small, it could be convenient to consider $c \neq 1/2$). Focusing now on the theoretical or computational complexity of the algorithm, we should take into account: (i) the construction of the linear model to be treated in the*

first stage, (ii) the application of the variable selection procedure to such linear model, (iii) the construction of the linear model to deal with in the second stage, and (iv) the application of the variable selection procedure to such linear model. For a fixed value $\theta \in \Theta_n$ *(in the sake of clarity, we consider* $\Theta_n = \Theta_n^{\mathbf{1}} = \Theta_n^{\mathbf{2}}$*), and given tuning parameters* h*,* w_n *and* λ*, the computational complexities related to the first stage ((i) and (ii)) are* $O(n^2 w_n)$ *and* $O(n w_n)$*, respectively. For the second stage ((iii) and (iv)), the computational complexities depend on* r_n *(the number of covariates in the linear model of the second stage), which is a random variable. It can be seen in Aneiros and Vieu [4] that, under suitable conditions, it is verified that* $r_n = O(s_n)$ *with probability 1 (w.p.1). Then, the computational complexities related to (iii) and (iv) are* $O(n^2 s_n)$ *and* $O(n s_n)$ *w.p.1, respectively. To sum up, the computational complexity of the proposed IASSMR algorithm is* $O(n^2 w_n \sharp\Theta_n) + O(n^2 s_n \sharp\Theta_n)$ *w.p.1. Therefore, it is expected that, in practice, in the usual case where* $\max\{s_n, w_n\} << p_n$*, the IASSMR algorithm will be much faster than the standard one (especially in situations where* $\max\{s_n, w_n\} << n << p_n$*) but slower than the FASSMR algorithm (especially in situations where* $s_n >> w_n$*); for the computational complexities of the IASSMR algorithm and the standard PLS procedure, see Remark 5.2.*

5.3 Asymptotic theory

5.3.1 Asymptotics for the FASSMR algorithm

For presenting theoretical results related to the variable selection performed by the FASSMR, we have to assume some technical conditions:

Conditions on the non-null parameters We assume standard hypotheses such as

$$\sharp S_n = s_n = o(p_n), \tag{5.16}$$

or

$$\exists c, \quad \forall n, \quad \sum_{j \in S_n} |\beta_{0j}| < c < \infty. \tag{5.17}$$

Conditions on the curve ζ. The curve ζ is observed in a grid such that

$$\exists c_1, c_2, \quad \forall j = 1, \ldots, p_n - 1, \quad 0 < \frac{c_1}{p_n} < t_{j+1} - t_j < \frac{c_2}{p_n} < \infty. \tag{5.18}$$

In addition,

$$\zeta \text{ is Lipschitz continuous on its support,} \tag{5.19}$$

and bounded away from zero; that is:

$$\exists c, \quad \forall t \in [a, b], \quad |\zeta(t)| \geq c > 0. \tag{5.20}$$

Conditions on the coefficients of the model. Let us assume that

$$\exists c, \quad \forall j = 1, \ldots, q_n, \quad \forall k = 1, \ldots, w_n, \quad \beta_{0j+(k-1)q_n} \neq 0 \Longrightarrow \left| \sum_{j=1}^{q_n} \beta_{0j+(k-1)q_n} \right| > c > 0. \tag{5.21}$$

Conditions on the standard variable selection method. Let us consider the SFPLSIM,

$$Y = \sum_{k=1}^{w_n} \alpha_{0j} X_j + g\left(\langle \delta_0, \mathcal{X} \rangle\right) + \varepsilon, \tag{5.22}$$

where X_j are random real covariates, δ_0 is an unknown functional direction and if we denote by $S_n^* = \{k = 1, \ldots, w_n,\ \alpha_{0j} \neq 0\}$ and $\sharp(S_n^*) = s_n^*$, it is verified that $s_n^* = o(w_n)$. The standard SCAD-PLS procedure leads to estimates $\widehat{\alpha}_{0j}$ of α_{0j} satisfying the following property:

$$\mathbb{P}\left(\{k = 1, \ldots, w_n;\ \alpha_{0k} = 0\} = \{k = 1, \ldots, w_n;\ \widehat{\alpha}_{0k} = 0\}\right) \to 1, \quad \text{when } n \to \infty. \tag{5.23}$$

Remark 5.5 *Note that suitable conditions under which (5.23) holds can be seen in Section 4.3. On the other hand, Assumption (5.21) is specific of the functional setting addressed here in this chapter (scalar variables with functional origin). Assumption (5.21) was first introduced in Aneiros and Vieu [4]; discussion and examples under*

which this condition is satisfied can be seen in Aneiros and Vieu [4, 5].

Finally, for introducing the theoretical result, for each $j = 1, \ldots, p_n$, we denote by k_j the unique integer $k \in \{1, \ldots, w_n\}$ such that $j \in \{(k-1)q_n + 1, \ldots, kq_n\}$. The following result establishes the relationship between the variable selection procedures related to the MFPLSIM (5.1) and the reduced model (5.4), in the following sense: if the j^{th} $(j = 1, \ldots, p_n)$ variable is relevant in the MFPLSIM (5.1), for n big enough, the corresponding k_j^{th} neighbouring variable in model (5.4) will be estimated as non-null; conversely, if the k^{th} variable is estimated as non-null in model (5.4), there will exist a neighbouring variable in the MFPLSIM (5.1) which will be relevant.

Proposition 5.6 *Under conditions (5.1), (5.2), (5.4), (5.5), (5.16)-(5.23), assuming that $w_n \to \infty$ when $n \to \infty$, it is verified that:*

1. $\mathbb{P}\left(\forall j \in S_n,\ \widehat{\beta}^{\mathbf{1}}_{0k_j} \neq 0\right) \longrightarrow 1$ *when* $n \to \infty$.
2. $\mathbb{P}\left(\forall k \in \widehat{S}^{\mathbf{1}}_n,\ \exists j \in \{1, \ldots, q_n\} \text{ such that } \beta_{0j+(k-1)q_n} \neq 0\right) \longrightarrow 1$ *when* $n \to \infty$.

5.3.2 Asymptotics for the IASSMR algorithm

For introducing asymptotic results related to the estimators from the IASSMR, let us add some hypotheses to those needed for the FASSMR:

Conditions on the coefficients of the model.

$$\forall k = 1, \ldots, w_n,\ \exists 0 < a_k < \infty,\ \sum_{j=1}^{q_n} \beta_{0j+(k-1)q_n} \neq 0 \Longrightarrow \sharp S^k \sim a_k q_n \ \text{ as } \ n \to \infty, \tag{5.24}$$

where, for any $k = 1, \ldots, w_n$,

$$S^k = \{j = 1, \ldots, p_n, \text{ such that } j = 1 + (k-1)q_n, \ldots, kq_n \text{ and } \beta_{0j} \neq 0\}.$$

Conditions on the standard variable selection method. Let us consider the SFPLSIM

$$Y = \sum_{j \in \mathcal{P}_n} \alpha_{0j} X_j + g\left(\langle \delta_0, \mathcal{X} \rangle\right) + \varepsilon, \tag{5.25}$$

where $\mathcal{P}_n \subset \{1,\ldots,p_n\}$ with $\sharp\mathcal{P}_n = O(w_n)$ or $\sharp\mathcal{P}_n = O(s_n)$. The standard SCAD-PLS procedure leads to estimates $\widehat{\alpha}_{0j}$ of α_{0j} and $\widehat{\delta}_0$ of δ_0 satisfying properties:

$$\mathbb{P}\left(\{j \in \mathcal{P}_n;\ \alpha_{0j} = 0\} = \{j \in \mathcal{P}_n;\ \widehat{\alpha}_{0j} = 0\}\right) \longrightarrow 1,\ \text{as } n \to \infty, \tag{5.26}$$

$$\exists \gamma \geq 0 \text{ such that } ||\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_0|| = O_p\left(n^{-1/2}\left(\sharp\mathcal{P}_n\right)^{\gamma}\right) \tag{5.27}$$

$$\text{and } \exists d : \mathbb{R} \to (0,\infty) \text{ such that } \left|\left|\widehat{\delta}_0 - \delta_0\right|\right| = O_p\left(n^{-1}d(h)\left(\sharp\mathcal{P}_n\right)^{\gamma-3/2}\right), \tag{5.28}$$

where we denote by $\boldsymbol{\alpha}_0 = (\alpha_{0j},\ j \in \mathcal{P}_n)^\top$ and $\widehat{\boldsymbol{\alpha}}_0 = (\widehat{\alpha}_{0j},\ j \in \mathcal{P}_n)^\top$.

Conditions on the semiparametric estimate. Let us consider the following semiparametric models:

$$Y = g_0\left(\langle \delta_0, \mathcal{X}\rangle\right) + \varepsilon,$$

$$X_j = g_j\left(\langle \delta_0, \mathcal{X}\rangle\right) + \eta_j, \quad j = 1,\ldots,p_n,$$

and denote $g_{j,\delta_0}(\chi) \equiv g_j\left(\langle \delta_0, \chi\rangle\right)$ with $j = 0,\ldots,p_n$. Let $\widehat{g}_{j,\delta_0}(\chi)$ be the corresponding semiparametric estimate for $g_{j,\delta_0}(\chi)$, with $j = 0,\ldots,p_n$, obtained from the models above by using the same kind of weights used in the IASSMR, and $\delta \in \Theta_n \subset \mathcal{H}$. The following assumptions will be needed

$$\sup_{\delta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\left\{|\widehat{g}_{0,\delta}(\chi) - g_{0,\delta_0}(\chi)|\right\} = O_p\left(a_n\right), \tag{5.29}$$

$$\max_{j\in S_n}\sup_{\delta\in\Theta_n}\sup_{\chi\in\mathcal{C}}\left\{|\widehat{g}_{j,\delta}(\chi) - g_{j,\delta_0}(\chi)|\right\} = O_p\left(b_n\right), \tag{5.30}$$

$$\max_{j\in S_n}\sup_{\chi\in\mathcal{C}}\left\{|g_{j,\delta_0}(\chi)|\right\} = O\left(1\right). \tag{5.31}$$

Remark 5.7 *On the one hand, Condition (5.24) is specific of the framework of scalar variables with functional origin (for details, discussion and examples under which this condition is satisfied, see Aneiros and Vieu [4, 5]). On the other hand, in Chapter 4 we have pointed conditions under which (5.26)–(5.31) hold (see Section 4.3), including the characterization of the function $d(\cdot)$ and the functional subset Θ_n. In the same way, in Lemma 4.17 are specified rates a_n and b_n.*

Next theorem presents some asymptotic results related to our proposed estimators obtained from the IASSMR algorithm. For expressions of $\widehat{S}_n$, $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\theta}_0$, see (5.13), (5.14) and (5.15), respectively.

Theorem 5.8 *Under conditions (5.1), (5.2), (5.5), (5.16)-(5.21), (5.24)-(5.28), and if $w_n \to \infty$ as $n \to \infty$, it is obtained*

$$\left|\left|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right|\right| = O_p\left(n^{-1/2} s_n^{\gamma}\right), \tag{5.32}$$

$$\left|\left|\widehat{\theta}_0 - \theta_0\right|\right| = O_p\left(n^{-1} d(h) s_n^{\gamma-3/2}\right), \tag{5.33}$$

and

$$\mathbb{P}\left(\widehat{S}_n = S_n\right) \longrightarrow 1, \quad n \to \infty. \tag{5.34}$$

Finally, using the estimation of the linear coefficients obtained in (5.14), for each $\theta \in \mathcal{H}$ define

$$\widehat{r}_\theta(\chi) \equiv \widehat{r}\left(\langle \theta, \chi \rangle\right) = \frac{\sum_{i=1}^n \left(Y_i - \boldsymbol{\zeta}_i^\top \widehat{\boldsymbol{\beta}}_0\right) K\left(d_\theta\left(\chi, \mathcal{X}_i\right)/h\right)}{\sum_{i=1}^n K\left(d_\theta\left(\chi, \mathcal{X}_i\right)/h\right)}, \ \forall \chi \in \mathcal{H}.$$

Theorem 5.9 *Under assumptions of Theorem 5.8, if in addition conditions (5.29), (5.30) and (5.31) are satisfied, $h \to 0$ and $b_n \to 0$ as $n \to \infty$, then*

$$\sup_{\theta \in \Theta_n} \sup_{\chi \in \mathcal{C}} \left\{\left|\widehat{r}_\theta(\chi) - r_{\theta_0}(\chi)\right|\right\} = O_p\left(a_n\right) + O_p\left(n^{-1/2} s_n^{\gamma+1/2}\right). \tag{5.35}$$

Corollary 5.10 *Under assumptions of Theorem 5.9, if in addition $\Theta_n^{\mathbf{2}} \subset \Theta_n$ and $\widehat{\theta}_0$ is the estimator of θ_0 obtained in (5.15), we have that*

$$\sup_{\chi \in \mathcal{C}} \left\{\left|\widehat{r}_{\widehat{\theta}_0}(\chi) - r_{\theta_0}(\chi)\right|\right\} = O_p\left(a_n\right) + O_p\left(n^{-1/2} s_n^{\gamma+1/2}\right). \tag{5.36}$$

5.3.2.1 The grouped impact point case

In FDA, due to the continuity of the curve ζ, we could expect that the impact points are grouped in some situations; that is, the significant variables are very close on the

discretization. Let us denote the set of true impact points as

$$T_n = \{t_j,\ j = 1, \ldots, p_n,\ \beta_{0j} \neq 0\}, \tag{5.37}$$

and its estimation as

$$\widehat{T}_n = \{t_j,\ j = 1, \ldots, p_n,\ \widehat{\beta}_{0j} \neq 0\}, \tag{5.38}$$

In the situation of Grouped-Impact-Point MFPLSIM (GIP-MFPLSIM), it could make sense the introduction of the following condition:

Conditions on the grouping of the impact points. There exist some intervals $I_1, \ldots, I_{J_n}$ such that $I_i \cap I_j = \emptyset$ and such that $T_n \subset I_n$ where $I_n = \cup_{j=1}^{J_n} I_j$ and

$$\mathbb{P}\left(T_n = I_n\right) \longrightarrow 1 \text{ when } n \to \infty. \tag{5.39}$$

Corollary 5.11 *Under the same conditions of Theorem 5.8, if, in addition, Assumption (5.39) holds, then*

$$\mathbb{P}\left(\widehat{T}_n = I_n\right) \longrightarrow 1 \quad \textit{when} \ \ n \to \infty. \tag{5.40}$$

5.4 Simulation study

In this section we are going to present two different scenarios of simulation to illustrate the behaviour in practice of the proposed algorithms, FASSMR and IASSMR.

5.4.1 First scenario

The aim of this first scenario is to show that the FASSMR algorithm achieves its main goal: it provides a good performance in comparison with the standard PLS procedure with much lower computational cost. In Section 5.4.1.1 we will introduce the model on which the simulation is based. To ensure high degree of generality, the model involves a mixture of smooth functional covariates together with very rough ones (Brownian motions). Then, we will discuss in Section 5.4.1.2 some practical issues

linked with the choice of the parameters of the method (with special attention to the key parameter w_n). Finally, results are reported along Section 5.4.1.3. Here, the computational time and the quality of estimation have been quantified for both the FASSMR and the standard PLS procedure. Section 5.4.1.4 provides a brief summary of conclusions and motivates the second scenario presented in Section 5.4.2.

5.4.1.1 The design

For different values of the sample size, $n \in \{100, 200, 300\}$, and different number of linear covariates, $p_n \in \{101, 201, 501, 1001, 10001\}$, we generated observations i.i.d. $\mathcal{D} = \{(\zeta_i, \mathcal{X}_i, Y_i)\}_{i=1}^{n+100}$ from the model

$$Y_i = \sum_{j=1}^{p_n} \beta_{0j}\zeta_i(t_j) + r\left(\langle\theta_0, \mathcal{X}_i\rangle\right) + \varepsilon_i \quad (i = 1, \ldots, n + 100), \tag{5.41}$$

where:

- t_j, with $j = 1, \ldots, p_n$, are equispaced points in $[0, 1]$, with $t_1 = 0$ and $t_{p_n} = 1$.

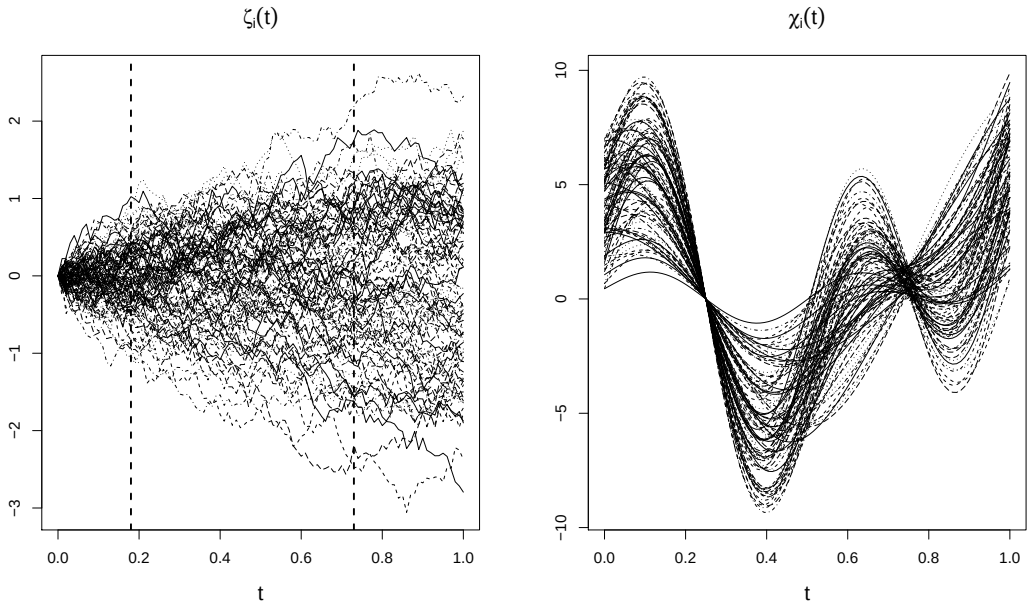
- ζ_i is a standard Brownian motion. We will consider only two non-null coefficients $\beta_{0j_1} = 2$ and $\beta_{0j_2} = -3$, being $t_{j1} = 0.18$ and $t_{j2} = 0.73$ the impact points (Figure 5.1 (a) shows 100 sample paths of standard Brownian motion with influential points marked on dotted vertical lines).

- Curves involved in the non-linear component were generated from:

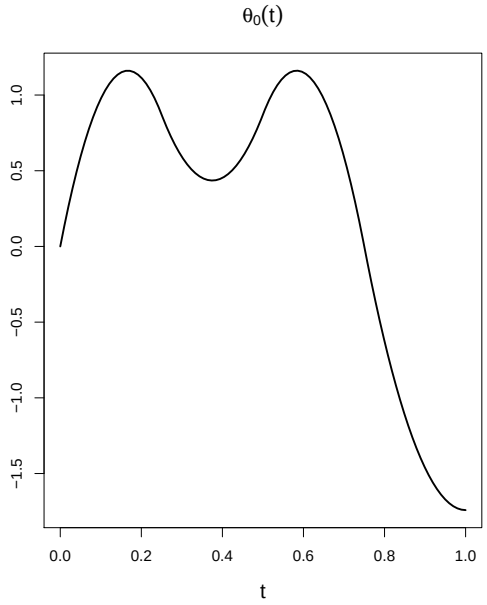
$$\mathcal{X}_i(t) = a_i \cos(2\pi t) + b_i \sin(4\pi t) + 2c_i(t - 0.25)(t - 0.5) \quad \forall t \in [0, 1], \tag{5.42}$$

  where the random variables a_i, b_i and c_i $(i = 1, \ldots, n + 100)$ are independent (both between and within vectors $(a_i, b_i, c_i)^\top$) and uniformly distributed on the interval $[0, 6]$. These curves were discretized on the same grid of 100 equispaced points in $[0, 1]$ (the representation of 100 of these curves can be seen in Figure 5.1 (b)).

Figure 5.1: Graphical representation of some components of model (5.41). In (a) dotted vertical lines mark the impact points at instants $t_{j1} = 0.18$ and $t_{j2} = 0.73$.



(a) Sample paths of standard Brownian. **(b)** Sample curves obtained from (5.42).



(c) θ_0.

- The true direction of projection was generated using the procedure described in Section 2.4. Values $l = 3$ and $m_n = 3$ were considered (note that the process of optimization involved in the estimation of the MFPLSIM requires intensive computation, which forces us to select a manageable number of interior knots) and the vector of coefficients of θ_0 in expression (2.36) was

$$(\alpha_1, \ldots, \alpha_{d_n})^\top = (0, 1.741539, 0, 1.741539, -1.741539, -1.741539)^\top. \quad (5.43)$$

  Graphical representation of the theoretical θ_0 can be seen in Figure 5.1 (c).

- The inner product and the link function considered were $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$ and $r(\langle\theta_0, \chi\rangle) = \langle\theta_0, \chi\rangle^3$, respectively.

- The i.i.d. random errors, ε_i $(i = 1, \ldots, n+100)$, were simulated from a normal distribution with zero mean and standard deviation equals to 0.1 times the standard deviation of the regression function $\sum_{j=1}^{p_n} \beta_{0j}\zeta_i(t_j) + r\left(\langle\theta_0, \mathcal{X}_i\rangle\right)$.

A total of $M = 100$ independent samples (i.e. $M = 100$ independent copies of $\mathcal{D}$) were generated from model (5.41). Each set $\mathcal{D}$ was split into two samples: a training sample

$$\mathcal{D}_{train} = \{(\zeta_i, \mathcal{X}_i, Y_i)\}_{i=1}^{n}, \quad (5.44)$$

and a testing sample,

$$\mathcal{D}_{test} = \{(\zeta_i, \mathcal{X}_i, Y_i)\}_{i=n+1}^{n+100}. \quad (5.45)$$

The training sample was used to make the estimation of all the parameters involved in (5.41). The testing sample was used to measure the quality of the corresponding predictions (i.e., the performance of the procedures) through the MSEP (2.40) with $n_{test} = 100$. For each sample, the FASSMR and the PLS procedures were applied.

5.4.1.2 Practical considerations

In practice, it is necessary to select some parameters to make the estimation using the FASSMR. The same problems are presented for estimating by means of the standard PLS method, except the choice of the division parameter $w = w_n$. That is specific to our new algorithm. Other important parameters to be chosen are the bandwidth

h, involved in semiparametric estimation, and the tuning penalization parameter λ_k, used in the variable selection procedure. Regarding to λ_k, to reduce the quantity of tuning parameters to be selected for each sample, we consider penalty parameters of the specific form $\lambda_k = \lambda\widehat{\sigma}_{\beta_{0k,OLS}}$, with $k = 1, \ldots, w$, where $\beta_{0k,OLS}$ denotes the OLS estimation of β_{0k} in the reduced model associated to (5.41) for each w, and $\widehat{\sigma}_{\beta_{0k,OLS}}$ is the estimated standard deviation. The selection of those mentioned parameters, as well as other related issues, can be performed as described below.

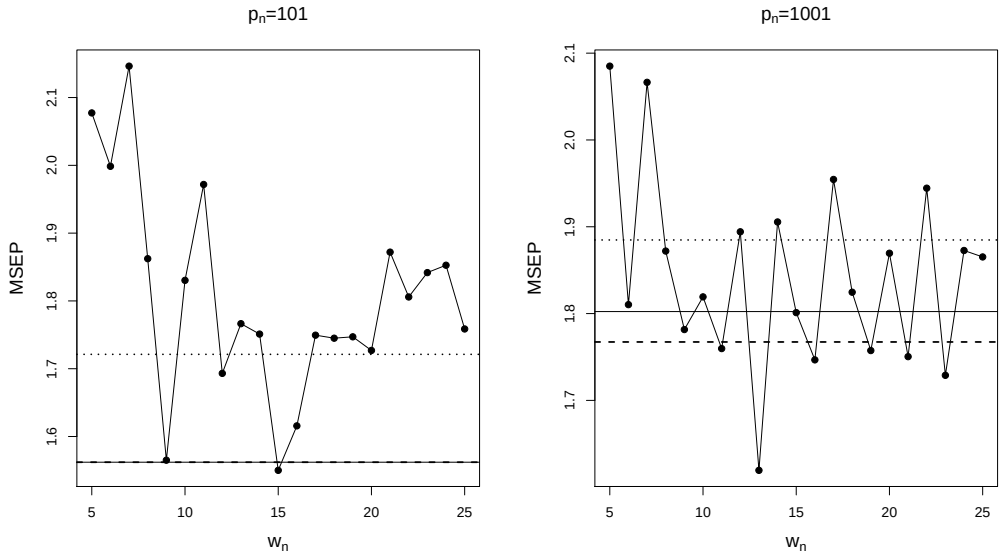
Firstly, the kernel K used was the Epanechnikov's one throughout this chapter (note that the choice of the kernel has low impact on the estimates). In addition, since we are concerned with computational time saving, the parameters h, λ and w will be selected by means of the BIC procedure (which has lower computational cost than the cross-validation selectors). Specifically, the BIC value corresponding to $(\widehat{\boldsymbol{\beta}}^{\mathbf{1}}_{0,h,\lambda,w}, \widehat{\theta}^{\mathbf{1}}_{0,h,\lambda,w})$ (the estimate of the parameter $(\boldsymbol{\beta}^{\mathbf{1}}_0, \theta^{\mathbf{1}}_0)$ in the linear model (5.6)) was computed from the routine `select` of the R package *grpreg*.

To select the penalty parameter λ in practice, it is usual to search it in a grid, $\{\lambda_{min}, \ldots\}$, where λ_{min} denotes de minimum value. A sensitivity analysis of the FASSMR (and the PLS) to the value λ_{min} was implemented. For that, for each value λ_{min} considered, a grid of 100 values, $\{\lambda_{min}, \ldots\}$, was provided to the program. Then $\widehat{\lambda}(\lambda_{min})$ was selected in such grid by means of the BIC and the corresponding MSEP$(\widehat{\lambda}(\lambda_{min}))$ was computed. Panel (c) in Figure 5.2 shows that the FASSMR is really affected by the value λ_{min}, while for the PLS method small values should be discarded.

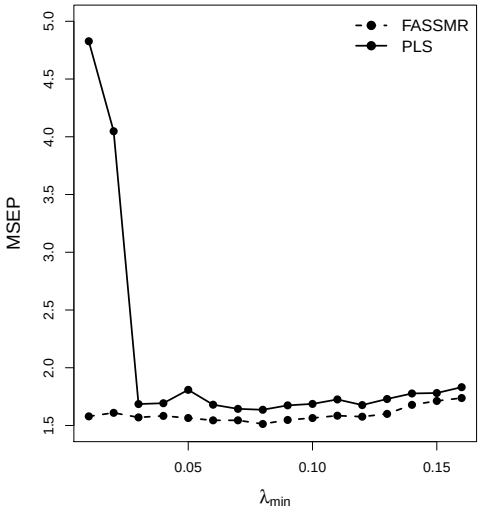
To select the splitting parameter w, the main task is to choose the eligible values for w before applying the BIC. Figure 5.2 shows the mean of MSEP over each value of $w \in W = \{5, 6, \ldots, 25\}$ for $M = 10$ samples of size $n = 100$, using $p_n = 101$ (panel (a)) and $p_n = 1001$ (panel (b)). In addition, it reports the MSEP from the FASSMR when w is selected using the BIC in W (see the solid horizontal line) or in W^* (see the dashed horizontal line), where we have denoted

$$W^* = \{10, 15, 20\}. \tag{5.46}$$

Figure 5.2: Panel (a) and (b): Mean of MSEP for each value of $w_n \in W$ for $M = 10$ samples of size $n = 100$ of (5.41) considering θ_0 known. Solid horizontal line is the mean of MSEP for the optimal value of $w_n \in W$ selected by the BIC. Dashed line is the mean of MSEP obtained for the optimal value of $w_n \in W^*$ selected by the BIC. Dotted line is the mean of the MSEP obtained through the PLS method for the same $M = 10$ samples. Panel (c): MSEP when the tuning parameter λ ($\widehat{\lambda}(\lambda_{min})$) is selected by minimizing the BIC over a grid starting in λ_{min} (θ_0 known).



(a) Case $p_n = 101$ **(b)** Case $p_n = 1001$



(c) MSEP over λ_{min}

Finally, the MSEP obtained from the PLS procedure is also shown (see the dotted horizontal line). The following conclusions can be derived from Figure 5.2 (panels (a) and (b)):

- The FASSMR is very sensitive to the value w, especially when w is small. This should be expected taking into account that the FASSMR is applied to an artificial (or reduced) model which, for small values of w, could be very different from the true model.

- The FASSMR improves PLS results in terms of MSEP for several values of w between 5 and 25. Focusing now on the performance of the FASSMR when w is selected in W by means of the BIC, we can conclude that the BIC is a suitable method (the MSEP provided is reasonable (see the solid horizontal line) and clearly improves the one obtained with the standard PLS procedure (see the dotted horizontal line)).

- In the sake of reducing computational time, we could consider to select w in W^* (instead of in W) by means of the BIC. The main reason for this (in addition to reduce the computational time) is that there is no loss in terms of MSEP using W^* (compare dashed and solid horizontal lines; results are even a bit better in case $p_n = 1001$ than using W as set of eligible values for w).

In conclusion, from now on, the set of eligible values for w will be W^*, and the selection will be made by means of the BIC procedure.

There is a minor question to be tuned and which is linked with the fact that, in many practical situations, the condition $p_n = w_n q_n$ fails. We will use the solution proposed in Aneiros and Vieu [5], based on consider not fixed $q_n = q_{n,k}$ values $k = 1, \ldots, w_n$, when p_n/w_n is not an integer number. Specifically:

$$q_{n,k} = \begin{cases} [p_n/w_n] + 1 & k \in \{1, \ldots, p_n - w_n[p_n/w_n]\}, \\ [p_n/w_n] & k \in \{p_n - w_n[p_n/w_n] + 1, \ldots, w_n\}, \end{cases} \tag{5.47}$$

where $[z]$ denotes the integer part of $z \in \mathbb{R}$.

Finally, in order to carry out the estimation of θ_0, a suitable set of eligible directions, $\Theta_n^{\mathbf{1}}$, should be considered. Accordingly, we follow again the procedure described in Section 2.4.

5.4.1.3 Results

Computational times (in seconds) required for estimating one sample using the FASSMR and the PLS method are collected in Table 5.1. Some comments can be made about the obtained times. On the one hand, PLS method is completely inefficient for big values of p_n. On the other hand, computational time saving is evident by means of the FASSMR, which allows obtaining results in acceptable time even for very big values of p_n.

Table 5.1: Computational time in seconds needed for making the estimation of one sample of model (4.31), using the PLS procedure and the FASSMR. Different values of n and p_n were considered. In the case of the FASSMR, the eligible values for w_n are those belonging to W^*. The results were obtained with a computer with the following features: Intel Core i7-7700HQ CPU, 8 GB RAM, 1 TB HDD, 256 GB SSD.

n	Method	$p_n = 101$	$p_n = 201$	$p_n = 501$	$p_n = 1001$	$p_n = 10001$
100	PLS	727.55	1324.2	2571.52	4959.70	43137.25
	FASSMR	374.28	362.67	367.23	365.14	357.17
200	PLS	1089.18	2625.83	7211.37	14823.14	153540.27
	FASSMR	1058.52	1034.00	1032.03	1008.02	702.67
300	PLS	3341.82	8091.13	20537.98	33868.11	224890.17
	FASSMR	3184.35	3412.45	2361.60	3123.95	2448.00

In Table 5.2 means of MSEP (using $M = 100$ samples) in different scenarios are computed for both methods. As can be derived from Table 5.2, using the FASSMR there is no loss in terms of MSEP. Note that in Table 5.2 results are obtained using 100 samples; then, we do not consider big values of n and p_n due to the huge computational time needed by the PLS procedure even for estimating only one sample (see again Table 5.1).

Table 5.2: For $M = 100$ samples of model (4.31), mean of MSEP for the PLS procedure and the FASSMR.

		$p_n = 101$		$p_n = 201$		$p_n = 501$	
n	Method	Mean	SD	Mean	SD	Mean	SD
100	PLS	1.2572	0.9546	1.3213	1.0078	1.3126	0.9688
	FASSMR	1.1579	1.0335	1.2694	1.0675	1.2025	0.9507
200	PLS	0.7662	0.4630	0.7617	0.4209	0.7194	0.4606
	FASSMR	0.6984	0.4274	0.8049	0.5000	0.7357	0.4860

5.4.1.4 Conclusions

The FASSMR allows to obtain the variable selection and estimation of model (5.1) in a reasonable amount of time, even for very big values of p_n. As can be derived from the simulation study, the developed algorithm clearly exceeds the standard PLS procedure in computational efficiency without loss in prediction power. Moreover, if we apply the Diebold-Mariano test for comparing the forecast accuracy of the two methods in each scenario of Table 5.2, we can obtain further conclusions: in some scenarios, there are not significant differences between the PLS and the FASSMR and in the scenarios where differences are significant, the FASSMR provides better prediction power.

However, the price to pay for this big computational time improvement is that the set of relevant variables could not be exactly obtained in many cases. That is supported by the asymptotic analysis derived from Proposition 5.6. In some real data applications, this lack of precision can be an inconvenient. In addition, situations of grouped impact points can be common, in which case the FASSMR could not provide the full set of influential variables.

5.4.2 Second scenario

In this section, Monte Carlo studies were carried out to compare the finite sample behaviour of the FASSMR and the IASSMR in two different frameworks: a first

design with spaced impact points and a second one with grouped impact points (GIP).

In Section 5.4.2.1 we will introduce the first simulation model, which has spaced impact points. We will briefly discuss some practical issues in Section 5.4.2.2. The results for this first design are reported throughout Section 5.4.2.3, where the computational time, the quality of the estimation and the precision of the impact point selection have been quantified for both the FASSMR and the IASSMR. In Section 5.4.2.4 we will present the second simulation model, which has grouped impact points. The results of this second design will be addressed in Section 5.4.2.5. The same features than in Section 5.4.2.3 were measured to compare both algorithms. Finally, Section 5.4.2.6 provides a brief summary of conclusions and the scope of application of each algorithm.

5.4.2.1 First design: spaced impact points

Observations i.i.d. $\mathcal{D} = \{(\zeta_i, \mathcal{X}_i, Y_i)\}_{i=1}^{n+100}$ were generated from the model

$$Y_i = \sum_{j=1}^{p_n} \beta_{0j}\zeta_i(t_j) + r\left(\langle\theta_0, \mathcal{X}_i\rangle\right) + \varepsilon_i, \quad (i = 1, \ldots, n+100) \tag{5.48}$$

where:

- Curves involved in the non-linear part, $\mathcal{X}_i = \mathcal{X}_{a_i,b_i,c_i}$ were generated from expression (5.42), but now the random variables a_i, b_i and c_i (which are independent both between and within vectors $(a_i, b_i, c_i)^\top$) are uniformly distributed on the interval $[0, 5]$. These curves were discretized on the same grid of 100 equispaced points in $[0, 1]$.

- t_j with $j = 1, \ldots, p_n$ are equispaced points on $[0, 1]$, with $t_1 = 0$ and $t_{p_n} = 1$.

- Curves involved in the linear component were generated from the expression

$$\zeta_i(t_j) = c_i t_j + d_i \tag{5.49}$$

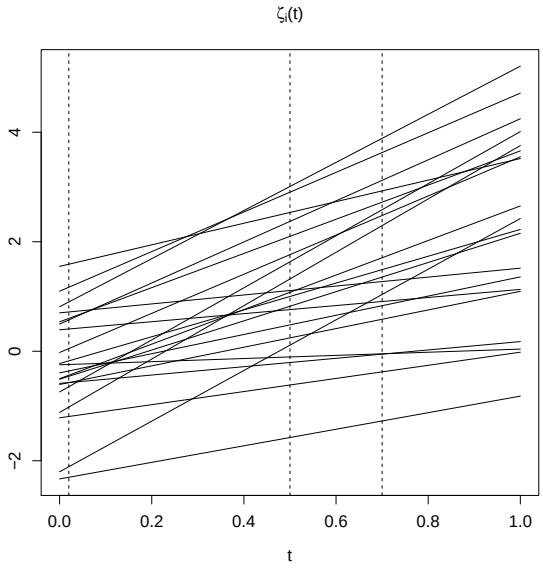
  where d_i is normally distributed with mean 0 and standard deviation equals to 1 and c_i was defined in the first item. As a consequence, there will exist some

dependence between $\mathcal{X}$ and ζ. In addition, we will consider only three non-null coefficients: $\beta_{0j_1} = 4$, $\beta_{0j_2} = 3$ and $\beta_{0j_3} = -3.2$, being $t_{j1} = 0.02$, $t_{j2} = 0.50$ and $t_{j3} = 0.70$ the impact points (left panel in Figure 5.3 shows 100 curves ζ_i with influential points marked in dotted vertical lines).

- The true direction (θ_0), the inner product ($\langle \cdot, \cdot \rangle$), the link function ($r(\cdot)$) and the random errors (ε_i) were generated as in Section 5.4.1.1.

$M = 100$ independent samples were generated from (5.48), which will be divided in $\mathcal{D}_{train}$ (see (5.44)) and $\mathcal{D}_{test}$ (see (5.45)). Values $p_n \in \{101, 201, 501, 1001, 10001\}$ will be considered. In this case, instead of fixing the sample size to be equal for the two methods, we are going to fix the sample size of the first step (the only step in the FASSMR), $n_1 = 100$, and we will vary the sample size of the second step, n_2 (we will consider $n_2 \in \{0, 100, 200\}$; case $n_2 = 0$ corresponds with the FASSMR). To perform the estimation using each method, we will follow the technical considerations collected in Subsection 5.4.1.2. Note that, as in the FASSMR, in the IASSMR we will use W^* (see (5.46)) as set of eligible values for w_n. In addition, the set of eligible directions, $\Theta_n^{\mathbf{2}}$, was generated in the same way as $\Theta_n^{\mathbf{1}}$ (see Section 2.4).

Figure 5.3: Sample of 20 lines generated from (5.49), together with impact points (dotted vertical lines) at instants $t_{j1} = 0.02$, $t_{j2} = 0.50$ and $t_{j3} = 0.70$.



Note that in this simulation study we want to compare the practical behaviour of the IASSMR and the FASSMR in computational efficiency and MSEP (2.40), but also in precision of the impact point selection. That is, we want to quantify the accuracy of the set $\widehat{S}_n$ obtained using each procedure. However, the continuous origin of the linear covariates makes difficult to difference the effect of points which are very close in the discretization. For that, comparing $\widehat{S}_n$ with S_n (for instance, by means of classical measures such as false discovery rate, specificity and sensibility) can be inappropriate, but it can make sense considering the following sets

$$I_n = [0.00, 0.05] \cup [0.47, 0.53] \cup [0.67, 0.73],$$
$$\overline{I}_n = (0.05, 0.47) \cup (0.53, 0.67) \cup (0.73, 1],$$

and classifying as well-chosen all those selected points belonging to I_n and as wrongly-chosen those belonging to $\overline{I}_n$. That is, we are going to quantify Right= $\sharp\{I_n \cap \widehat{T}_n\}$ and Wrong= $\sharp\{\overline{I}_n \cap \widehat{T}_n\}$ for the IASSMR and the FASSMR, where $\widehat{T}_n$ was defined in (5.38).

5.4.2.2 Practical considerations

In practice, as in the case of the FASSMR, various tuning parameters have to be selected for performing the estimation associated to the IASSMR. Here, we focus on the selection of the parameters h, λ and w because, since two stages are considered in the IASSMR, some clarifications on the BIC procedure are needed. The goal is to select such parameters in a way that the final estimator in the second stage (equivalently, in the second model M2 (5.11)) achieves the minimum value for the BIC. Specifically, since the covariates in M2 depend on the covariates selected in the first model M1 (the model in stage 1), we first select, for each w, the covariates in M1 using the BIC procedure to choose the corresponding parameters $h_w^{\mathbf{1}}$ and $\lambda_w^{\mathbf{1}}$. Then, once M2=M2$_w$ is constructed, the BIC procedure is applied again to choose the parameters $h_w^{\mathbf{2}}$ and $\lambda_w^{\mathbf{2}}$ corresponding to the estimators of $(\boldsymbol{\beta}_0^{\mathbf{2}}, \theta_0^{\mathbf{2}})$ in M2$_w$. Finally, if we denote $\text{BIC}_w^{\mathbf{2}} = \text{BIC}(h_w^{\mathbf{2}}, \lambda_w^{\mathbf{2}})$ (the BIC value corresponding to such estimators), the selected parameters are $w_{opt}^{\mathbf{2}}$, $h_{w_{opt}^{\mathbf{2}}}^{\mathbf{2}}$ and $\lambda_{w_{opt}^{\mathbf{2}}}^{\mathbf{2}}$, where $w_{opt}^{\mathbf{2}} = \arg\min \text{BIC}_w^{\mathbf{2}}$.

5.4.2.3 First design: results

Table 5.3: For $M = 100$ samples from the MFPLSIM (5.48), time in seconds needed to make the estimation of one sample of size $n = n_1 + n_2$ ($n_1 = 100$). The results were obtained with a computer with the following features: Intel Core i7-7700HQ CPU, 8 GB RAM, 1 TB HDD, 256 GB SSD.

n_2	Method	$p_n = 101$	$p_n = 201$	$p_n = 501$	$p_n = 1001$	$p_n = 10001$
0	FASSMR	203.00	202.75	202.64	203.20	210.21
100	IASSMR	486.94	678.60	1243.92	2332.59	21278.42
200	IASSMR	840.93	1153.92	2435.96	6636.79	43554.89

Table 5.4: For $M = 100$ samples from the MFPLSIM (5.48), mean of MSEP for the FASSMR and the IASSMR using $n_1 = 100$ ($n = n_1 + n_2$).

		$p_n = 101$		$p_n = 201$		$p_n = 501$	
n_2	Method	Mean	SD	Mean	SD	Mean	SD
0	FASSMR	0.5877	0.3666	0.6122	0.4498	0.6164	0.4636
100	IASSMR	0.3888	0.1563	0.3965	0.1689	0.3800	0.1543
200	IASSMR	0.3174	0.1049	0.3161	0.1045	0.3166	0.1040

Tables 5.3-5.5 show the effect of adding a second step to the FASSMR in terms of computational efficiency, MSEP and precision in the impact point selection, respectively, as well as the influence of the sample size of this second step. As expected, the second stage increases the total time required for estimation and this increase is greater the larger the size of the discretization (because of the construction of $\mathcal{R}_n^{\mathbf{2}}$). However, from Tables 5.4 and 5.5 it can be derived that MSEP and precision of impact point selection are clearly improved with a second stage (note that we do not consider big values for p_n because of the big computational time needed for estimating one sample in the IASSMR case). Furthermore, if we analyse the effect

of the increase of p_n in Table 5.4, we can see that the IASSMR is less affected than the FASSMR. In fact, in the IASSMR case there is no deterioration of the results.

Table 5.5: For $M = 100$ samples from the MFPLSIM (5.48), mean of the number of variables rightly-chosen ($\sharp\{I_n \cap \widehat{T}_n\}$) and wrongly-chosen ($\sharp\{\overline{I}_n \cap \widehat{T}_n\}$) by the FASSMR and the IASSMR procedures using $n_1 = 100$ ($n = n_1 + n_2$).

		$p_n = 101$		$p_n = 201$		$p_n = 501$	
n_2	Method	Right	Wrong	Right	Wrong	Right	Wrong
0	FASSMR	1.24	1.60	1.28	1.36	1.29	1.50
100	IASSMR	1.31	1.21	1.27	1.15	1.34	0.90
200	IASSMR	1.36	1.16	1.30	1.18	1.33	0.88

5.4.2.4 Second design: grouped impact points

As in Section 5.4.1, observations i.i.d. $\mathcal{D} = \{(\zeta_i, \mathcal{X}_i, Y_i)\}_{i=1}^{n+100}$ were generated using $n \in \{100, 200, 300\}$ and $p_n \in \{101, 201, 501, 1001, 10001\}$, but now, coming from the following modification of model (5.41):

$$Y_i = \sum_{j=1}^{p_n} \beta_{0j}\zeta_i(t_j) + r\left(\langle \theta_0, \mathcal{X}_i \rangle\right) + \varepsilon_i, \quad (i = 1, \ldots, n + 100) \tag{5.50}$$

where, in this case:

- Ten non-null coefficients will be considered, which will be in correspondence with the following impact points:

$$\begin{cases} \beta_{0j_1} = 1.0, & t_{j_1} = 0.15, \\ \beta_{0j_2} = 1.2, & t_{j_2} = 0.16, \\ \beta_{0j_3} = 1.0, & t_{j_3} = 0.17, \\ \beta_{0j_4} = 1.2, & t_{j_4} = 0.18, \\ \beta_{0j_5} = 1.0, & t_{j_5} = 0.19. \end{cases} \quad \begin{cases} \beta_{0j_6} = 1.0, & t_{j_6} = 0.70, \\ \beta_{0j_7} = 1.2, & t_{j_7} = 0.71, \\ \beta_{0j_8} = -1.2, & t_{j_8} = 0.72, \\ \beta_{0j_9} = -1.2, & t_{j_9} = 0.73, \\ \beta_{0j_{10}} = -1.2, & t_{j_{10}} = 0.74. \end{cases} \tag{5.51}$$

- Curves involved in the non-linear part were generated from (5.42), but the random variables a_i, b_i and c_i $(i = 1, \ldots, n)$ are independent and uniformly distributed on the interval $[0, 5]$.

As a consequence of (5.51), we obtain a Grouped-Impact-Point MFPLSIM, GIP-MFPLSIM (in fact, relevant variables are consecutive in the case $p_n = 101$). $M = 100$ independent samples were generated from the GIP-MFPLSIM (5.50), which will be divided into $\mathcal{D}_{train}$ (see (5.44)) and $\mathcal{D}_{test}$ (see (5.45)). Subsequently, the FASSMR and IASSMR procedures were applied following the same scheme and considerations as in the first scenario, but now we are going to consider the same sample size for the two procedures (which is closer to what happens in applications to real data). Then, in the case of the IASSMR the sample is divided into two parts, one for the first stage and one for the second stage. In this application we will consider $n_1 = n_2 = n/2$.

As before, we will compare both methods in computational time for estimating one sample, MSEP (2.40) and precision of the impact point selection. Focusing in this last point, we will consider the following sets

$$I_n = [0.14, 0.20] \cup [0.69, 0.75],$$
$$\overline{I}_n = [0, 0.14) \cup (0.20, 0.69) \cup (0.75, 1].$$

and we are going to quantify Right=$\sharp\{I_n \cap \widehat{T}_n\}$ and Wrong=$\sharp\{\overline{I}_n \cap \widehat{T}_n\}$ for the IASSMR and the FASSMR.

5.4.2.5 Second design: results

First of all, Table 5.6 shows computational times for estimating one sample using both IASSMR and FASSMR. Looking at Table 5.6, we can see that the computational time needed by the IASSMR is affected by p_n (as it is also derived from Table 5.3), while the FASSMR is only affected by increasing n. It is noteworthy remark that for moderate sample size ($n = 200, 300$) and small p_n, computational time needed by the IASSMR is similar or even smaller than that needed by the FASSMR. This is due to the division of the sample in the IASSMR two-stage procedure.

Tables 5.7 and 5.8 allow us to analyse and compare the accuracy of the predictions and the variable selection, respectively, performed by the FASSMR and the IASSMR.

Note that we do not consider big values for p_n because of the big computational time needed for estimating one sample in the IASSMR case. Some general observations can be derived from those tables:

i) The performance of both procedures improves with increasing sample size (n). The effect of increasing number of linear covariates (p_n) is more difficult to analyse. Both procedures are adversely affected with increasing p_n if we compare the case $p_n = 101$ with the cases $p_n = 201$ and $p_n = 501$. However, from $p_n = 201$ to $p_n = 501$ there is no deterioration on results. In fact, results for $p_n = 501$ are even better in some cases.

ii) For small sample size ($n = 100$), the FASSMR outperforms the results obtained by the IASSMR in terms of MSEP for all considered values of p_n. On the other hand, the number of variables well selected is bigger in the IASSMR case, but it is also bigger the number of wrongly selected variables.

iii) For moderate sample size ($n = 200$ and $n = 300$), the IASSMR provides better results than the FASSMR for all considered values of p_n.

Table 5.6: Computational time in seconds needed for making the estimation of one sample from the GIP-MFPLSIM (5.50), using the IASSMR and the FASSMR, for different values of n and p_n. The eligible values for w_n are those belonging to W^* (see (5.46)). The results were obtained with a computer with the following features: Intel Core i7-7700HQ CPU, 8 GB RAM, 1 TB HDD, 256 GB SSD.

n	Method	$p_n = 101$	$p_n = 201$	$p_n = 501$	$p_n = 1001$	$p_n = 10001$
100	FASSMR	405.53	479.6	436.22	260.44	251.28
	IASSMR	653.26	1156.32	3016.60	4877.91	24860.64
200	FASSMR	983.36	822.14	805.41	580.97	558.92
	IASSMR	931.11	1070.92	3047.36	5450.66	31641.58
300	FASSMR	2241.66	2080.84	1979.23	2062.11	2337.90
	IASSMR	1684.22	1950.67	2290.78	9041.99	71789.74

Table 5.7: For $M = 100$ samples from the GIP-MFPLSIM (5.50), mean of MSEP obtained from the FASSMR and the IASSMR procedures.

		$p_n = 101$		$p_n = 201$		$p_n = 501$	
n	Method	Mean	SD	Mean	SD	Mean	SD
100	FASSMR	0.5827	0.3208	0.6908	0.3890	0.6803	0.3743
	IASSMR	1.0988	2.0134	3.3929	5.6060	2.7154	4.9089
200	FASSMR	0.4076	0.1484	0.4579	0.1443	0.4510	0.1625
	IASSMR	0.3954	0.2038	0.4097	0.2154	0.4255	0.2522
300	FASSMR	0.3573	0.1217	0.4127	0.1296	0.3857	0.1074
	IASSMR	0.2916	0.1208	0.3142	0.1326	0.3018	0.1166

Table 5.8: For $M = 100$ samples from the GIP-MFPLSIM (5.50), mean of the number of variables rightly-chosen ($\sharp\{I_n \cap \widehat{T}_n\}$) and wrongly-chosen ($\sharp\{\overline{I}_n \cap \widehat{T}_n\}$) by the FASSMR and the IASSMR procedures.

		$p_n = 101$		$p_n = 201$		$p_n = 501$	
n	Method	Right	Wrong	Right	Wrong	Right	Wrong
100	FASSMR	1.95	2.45	1.78	2.81	1.76	3.12
	IASSMR	4.11	3.52	4.26	8.34	4.69	8.85
200	FASSMR	1.99	2.06	1.90	2.61	1.88	2.54
	IASSMR	4.46	1.12	4.55	1.59	4.68	2.08
300	FASSMR	2.00	2.08	1.92	2.23	1.87	2.21
	IASSMR	4.82	0.50	4.84	0.79	5.17	0.98

It should be noted that fact ii) is a consequence of dividing the sample of size 100 into two subsamples, of size 50, to perform the two-stage procedure associated to the IASSMR. This sample size seems insufficient to get a good estimation of θ_0.

In the case of enough sample size (observation iii), the second stage in the IASSMR makes it possible to recover some information that was lost during the first stage. In this situation, results provided by the IASSMR are less affected by the discretization size and by w_n, surpassing those obtained with the FASSMR both in MSEP and in number of variables well and wrongly selected.

5.4.2.6 Conclusions

The simulation study performed in the GIP-MFPLSIM illustrates the utility of refining the FASSMR in order to obtain a more sophisticated algorithm. In particular, as expected and as it was highlighted theoretically (see result (5.40)), the IASSMR overcomes the drawbacks of the FASSMR in terms of impact points selection. This improvement goes with high predictive performance and reasonable computational costs. Furthermore, comparisons between the practical behaviour of the FASSMR and IASSMR (which were performed in three ways: MSEP, accuracy of variable selection and computational time) give us a guidelines about which algorithm should be used in each practical situation. Basically, the main recommendations can be summarized as follows:

- For small n and big p_n we should use the FASSMR.
- For big/moderate n and small/moderate p_n, it is advisable to use the IASSMR.
- For big both n and p_n, the FASSMR will provide us a first approximation. Results of the IASSMR will probably give us more precision in the set of selected variables but, of course, with higher computational costs. However, we should note that computational time required by the IASSMR will be much lower than that needed by the standard PLS method.

5.5 Application to real data

The aim of this section is to show the usefulness of the presented methodology through its application to solve a real problem: the prediction of the ash content in a sugar sample, having its absorbance spectra at two different excitation wavelengths.

Although the ash content can be determined by chemical analyses, the use of functional regression to predict this variable will provide a high economical advantage. Then, this section is devoted to analyse Sugar data (see Section 1.1) by using the MFPLSIM and the algorithms described just before.

5.5.1 The data

Sugar data was presented in Section 1.1 and briefly described in Section 5.1. As mentioned, we have 268 samples from $(Y, \zeta, \mathcal{X})$, where Y is a scalar random variable (ash content) and ζ and $\mathcal{X}$ are functional random variables (absorbance spectra from 275 to 560 nm at excitation wavelengths 240 nm and 290 nm, respectively; both variables were observed on $p_n = 571$ equally spaced wavelengths in the interval $[275, 560]$). Although the number of available samples was 268, two samples were discarded in this application as extreme outliers. Therefore, our dataset consists in 266 samples $\mathcal{D} = \{(\zeta_i, \mathcal{X}_i, Y_i)\}_{i=1}^{266}$, and our goal is to predict Y by means of ζ and $\mathcal{X}$. More details, can be seen in Sections 1.1 and 5.1, and for graphics of the curves, see Figure 1.3.

In order to evaluate models and estimation methods that will be proposed, the dataset, $\mathcal{D}$, will be split into two subsamples: a training sample $\mathcal{D}_{train} = \{(\zeta_i, \mathcal{X}_i, Y_i)\}_{i=1}^{216}$ and a testing sample $\mathcal{D}_{test} = \{(\zeta_i, \mathcal{X}_i, Y_i)\}_{i=217}^{266}$. Therefore, the estimation task is made by means of $\mathcal{D}_{train}$, while $\mathcal{D}_{test}$ is used to measure the quality of predictions. For that, we will use again the MSEP (2.40) with $n = 216$ and $n_{test} = 50$.

5.5.2 Results

To get a first idea of the effect of each functional variable in the response, we have carried out a preliminary study. Firstly, we have modelled data through two uni-functional models: a FLM (1.1) and a FSIM (2.1), and in both cases, we have constructed models with each functional variable. Secondly, we have modelled data through a bi-functional model, which combines in an additive way a functional linear component with a functional single-index one. In this model, called functional partial linear single-index model (FPLSIM), both covariates enter with continuous effect in

the response. We have built two FPLSIM, each one with a different variable in each component of the model.

The FLM was estimated using functional principal component analysis (PCA) (by means of the `fregre.pc` function in the *fda.usc* R package, see Febrero-Bande and Oviedo de la Fuente [42]). The FSIM was estimated using kernel-based estimators (see (2.10)) together with Epanecknikov kernel, selecting h by means of the BIC criterion. θ was estimated by means of the procedure described in Section 2.4, using $l = 3$ and selecting m_n, see expression (2.36), by means of the cross-validation procedure (as in Chapter 2). The FPLSIM was estimated following the procedure described in Lian [74], using the previous described tools to perform the estimation of the functional linear and functional single-index regressions involved.

Models and results of MSEP can be seen in Tables 5.9 and 5.10. From the results obtained we can derive that a linear effect will be convenient for ζ, since the lowest MSEP in Table 5.9 and also in Table 5.10 is obtained using such effect for ζ. Comparing Tables 5.9 and 5.10 we can see that the addition of the variable $\mathcal{X}$ semiparametrically provides a slight benefit on the prediction power.

Table 5.9: Uni-functional regression models and values of the criterion error.

	Model	MSEP
FLM	$Y = \gamma_0 + \int_{275}^{560} \mathcal{X}(t)\gamma(t)dt + \varepsilon$	4.5878
	$Y = \gamma_0 + \int_{275}^{560} \zeta(t)\gamma(t)dt + \varepsilon$	2.2072
FSIM	$Y = r\left(\langle\theta, \mathcal{X}\rangle\right) + \varepsilon$	3.6981
	$Y = r\left(\langle\theta, \zeta\rangle\right) + \varepsilon$	2.6802

Table 5.10: Bi-functional regression models (both variables enter with continuous effect) and values of the criterion error.

	Model	MSEP
FPLSIM	$Y = \int_{275}^{560} \mathcal{X}(t)\gamma(t)dt + r\left(\langle\theta, \zeta\rangle\right) + \varepsilon$	2.8749
	$Y = \int_{275}^{560} \zeta(t)\gamma(t)dt + r\left(\langle\theta, \mathcal{X}\rangle\right) + \varepsilon$	2.1854

Therefore, it makes sense to consider the following model

$$Y_i = \sum_{j=1}^{571} \beta_{0j}\zeta_i(t_j) + r\left(\langle\theta_0, \mathcal{X}_i\rangle\right) + \varepsilon_i, \tag{5.52}$$

and to analyse the results obtained if we combine this model with the variable selection tools presented throughout this chapter.

To estimate model (5.52), we apply the standard PLS method, the FASSMR and the IASSMR. For this task, in the three cases, technical considerations collected in Section 5.4.1.2 were used. Value $l = 3$ was considered in (2.36), while m_n was selected by means of the BIC criterion.

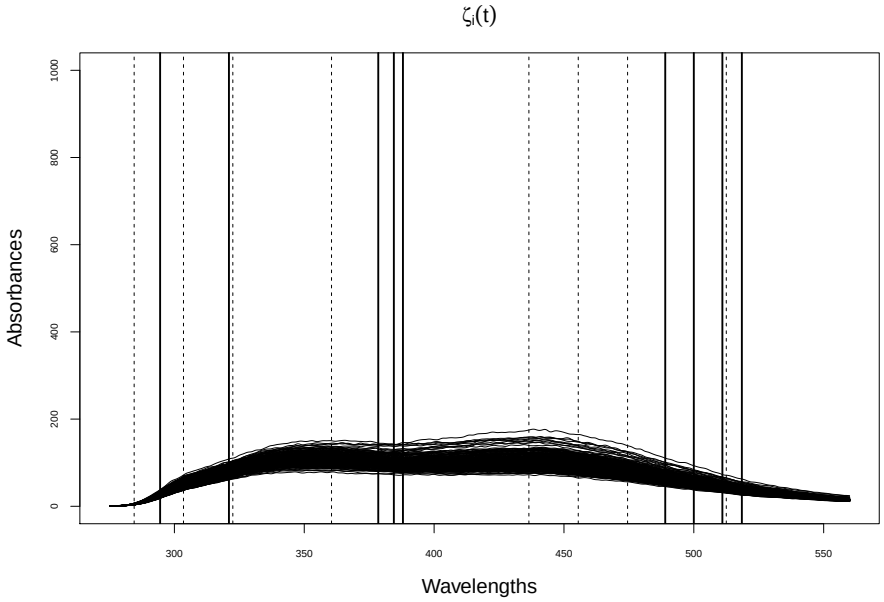
Table 5.11: Results obtained applying each variable selection method to model (5.52).

Method	MSEP	$\widehat{s}_n$	$\widehat{m}_n$	$\widehat{w}_n$	p_t
PLS	6.0375	53	4	-	9.8689
FASSMR	3.0329	8	2	15	1
IASSMR	2.0064	9	2	15	4.3399

Several comments can be made about Table 5.11, where the numerical results are presented. The column p_t contains the proportion of time needed by the three methods to return the final results compared to the fastest algorithm of the three (the FASSMR). On the one hand, the PLS method offers the most complex model: a total of 53 linear covariates and a more complicated expression for the estimated direction $\widehat{\theta}_0$ (4 regularly spaced interior knots are needed for its B-spline representation). Furthermore, this complexity is accompanied by the worst result in MSEP and the worst calculation time. On the other hand, the FASSMR clearly improves the PLS results in terms of complexity of the model (we get a simpler model) and MSEP; but the best result in MSEP is obtained by the IASSMR. This fact is related to the set $\widehat{S}_n$ obtained with this algorithm: the second stage in the IASSMR specifies and completes the set of relevant variables provided by the FASSMR. Figure 5.4 illustrates this fact and shows us that some GIP structure may be present around

385 nm. In addition, it illustrates the fact that none of the relevant variables selected by the FASSMR is also selected by the IASSMR (note that this is not incoherent since the goal of the IASSMR is to refine the selection made by the FASSMR). We should also note that the IASSMR is faster than the PLS procedure.

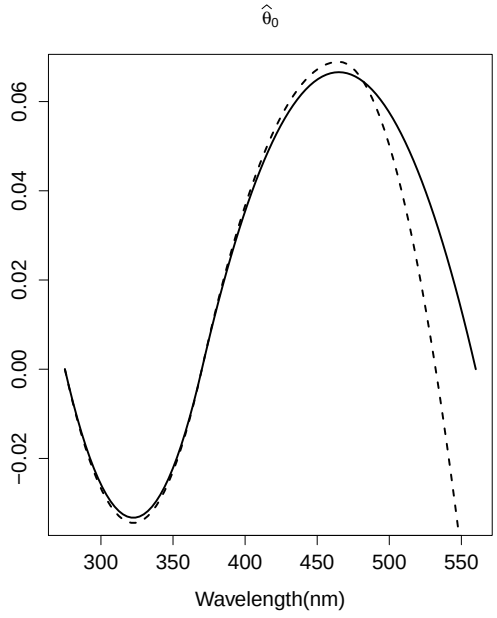
Figure 5.4: Absorbance curves at excitation wavelengths 240 nm (ζ) with impact points selected using the FASSMR (dashed vertical lines) and the IASSMR (solid vertical lines).



Finally, if we compare results in Tables 5.10 and 5.11 we will see that the MFPLSIM (combined with the IASSMR) offers better prediction power than the FPLSIM. This fact supports the idea that the MFPLSIM allows to include point-wise effects of the functional variable which can not be reached using the continuous curve.

Finally, as can be seen in Figure 5.5, the estimated direction using both introduced algorithms has a quite similar shape: in both cases, it presents a bump around 325 nm and a peak around 475 nm, which could be important indicators of the effect of $\mathcal{X}$ on the ash content of sugar.

Figure 5.5: $\widehat{\theta}_0$ using IASSMR (solid line) and FASSMR (dashed line).



5.5.3 Conclusions

The Sugar data application illustrates both the utility of the MFPLSIM in modelling real problems and the good performance of the presented algorithms in the estimation of this model. On the one hand, the MFPLSIM has two great advantages: it allows the inclusion of more than one functional covariate in the model and, in addition, these covariates enter involving interpretable parameters ($\boldsymbol{\beta}_0$ and θ_0). On the other hand, its semiparametric feature, combined with a good estimation tool, provides low prediction errors. Furthermore, the algorithms developed for variable selection and estimation of the MFPLSIM show a good behaviour compared to the standard PLS method both in MSEP and computational time. The FASSMR provides a preliminary quick result while the IASSMR gives refined estimates. In particular, the combination of MFPLSIM and IASSMR seems a potent tool, since in this application reaches the best result in MSEP.

5.6 Appendix Chapter 5: Proofs

5.6.1 Proof of Proposition 5.6

Note that assertion 1 of Proposition 5.6 can be proved ensuring that

$$\begin{aligned} \mathbb{P}\left(\exists j \in S_n \text{ such that } \widehat{\beta}^{\mathbf{1}}_{0k_j} = 0\right) \quad = \\ \mathbb{P}\left(\exists j = 1, \ldots, p_n,\ \beta_{0j} \neq 0 \text{ such that } \widehat{\beta}^{\mathbf{1}}_{0k_j} = 0\right) \quad \longrightarrow \quad 0 \text{ when } n \to \infty. \end{aligned} \tag{5.53}$$

For that, we should note that we are under the assumptions of Lemma 2 in Aneiros and Vieu [4]. Using the first assertion of that lemma we obtain:

$$\begin{aligned} \mathbb{P}\left(\exists j = 1, \ldots, p_n,\ \beta_{0j} \neq 0 \text{ such that } \widehat{\beta}^{\mathbf{1}}_{0k_j} = 0\right) \leq \\ \mathbb{P}\left(\exists k = 1, \ldots, w_n,\ \beta_{0k} \neq 0 \text{ such that } \widehat{\beta}^{\mathbf{1}}_{0k} = 0\right). \end{aligned} \tag{5.54}$$

Now using Assumption (5.23), the right hand term of expression (5.54) tends to 0 as n tends to ∞. Then, (5.53) is proved and, as a consequence, the assertion 1 of Proposition 5.6.

Following an analogous reasoning, assertion 2 of Proposition 5.6 can be proved ensuring that

$$\begin{aligned} \mathbb{P}\left(\exists k \in \widehat{S}^{\mathbf{1}}_n \text{ such that } \forall j \in \{1, \ldots, q_n\},\ \beta_{0j+(k-1)q_n} = 0\right) \quad = \\ \mathbb{P}\left(\exists k = 1, \ldots, w_n,\ \widehat{\beta}^{\mathbf{1}}_{0k} \neq 0 \text{ such that } \forall j \in \{1, \ldots, q_n\},\ \beta_{0j+(k-1)q_n} = 0\right) \quad \longrightarrow \quad 0 \\ \text{when } n \to \infty. \end{aligned} \tag{5.55}$$

Using the Assumption (5.23), we obtain:

$$\begin{aligned} \mathbb{P}\left(\exists k = 1, \ldots, w_n,\ \widehat{\beta}^{\mathbf{1}}_{0k} \neq 0 \text{ such that } \forall j \in \{1, \ldots, q_n\},\ \beta_{0j+(k-1)q_n} = 0\right) \leq \\ \mathbb{P}\left(\exists k = 1, \ldots, w_n,\ \beta^{\mathbf{1}}_{0k} \neq 0 \text{ such that } \forall j \in \{1, \ldots, q_n\},\ \beta_{0j+(k-1)q_n} = 0\right) + o(1). \end{aligned} \tag{5.56}$$

Now applying the second assertion in Lemma 2 in Aneiros and Vieu [4], the right

hand term in (5.56) tends to 0 as n tends to ∞. Therefore, (5.55) is proved and then assertion 2 of Proposition 5.6. ■

5.6.2 Proof of Theorem 5.8

This theorem provides three results. The proof of each of them is presented below.

5.6.2.1 Proof of (5.32)

If we define

$$\mathcal{R}_n^{\mathbf{1}*} = \left\{ j = 1, \ldots, p_n, \text{ such that } \zeta(t_j) \in \mathcal{R}_n^{\mathbf{1}} \right\},$$
$$\mathcal{R}_n^{\mathbf{2}*} = \left\{ j = 1, \ldots, p_n, \text{ such that } \zeta(t_j) \in \mathcal{R}_n^{\mathbf{2}} \right\},$$

one can write:

$$\left|\left|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right|\right|^2 = \sum_{j \in \mathcal{R}_n^{\mathbf{2}*}} \left(\widehat{\beta}_{0j}^{\mathbf{2}} - \beta_{0j}^{\mathbf{2}}\right)^2 + \sum_{j \notin \mathcal{R}_n^{\mathbf{2}*}, j \in S_n} (\widehat{\beta}_{0j} - \beta_{0j})^2. \tag{5.57}$$

On the one hand, considering $\mathcal{P}_n = \mathcal{R}_n^{\mathbf{2}*}$ in (5.27) and taking into account that $\sharp \mathcal{R}_n^{\mathbf{2}*} = O(s_n)$, using the reasoning to achieve (A.8) in Aneiros and Vieu [4] we obtain

$$\sum_{j \in \mathcal{R}_n^{\mathbf{2}*}} \left(\widehat{\beta}_{0j}^{\mathbf{2}} - \beta_{0j}^{\mathbf{2}}\right)^2 = O_p\left(n^{-1} s_n^{2\gamma}\right). \tag{5.58}$$

On the other hand, considering $\mathcal{P}_n = \mathcal{R}_n^{\mathbf{1}*}$ in (5.27) and using both Assumption (5.26) and the first assertion in Proposition 5.6, we obtain

$$\sum_{j \notin \mathcal{R}_n^{\mathbf{2}*}, j \in S_n} \left(\widehat{\beta}_{0j} - \beta_{0j}\right)^2 = O_p\left(n^{-1} s_n^{2\gamma}\right) \tag{5.59}$$

(for specific details, see proof of (A.12) in Aneiros and Vieu [4]). Therefore, result (5.32) is obtained by combining (5.57) with (5.58) and (5.59). ■

5.6.2.2 Proof of (5.33)

It is verified that

$$\left|\left|\widehat{\theta}_0 - \theta_0\right|\right| = \left|\left|\widehat{\theta}_0^{\mathbf{2}} - \theta_0\right|\right| \leq \left|\left|\widehat{\theta}_0^{\mathbf{2}} - \theta_0^{\mathbf{2}}\right|\right| + \left|\left|\theta_0^{\mathbf{2}} - \theta_0\right|\right|. \tag{5.60}$$

On the one hand, considering $\mathcal{P}_n = \mathcal{R}_n^{\mathbf{2}*}$ in (5.28) and using that $\sharp\mathcal{R}_n^{\mathbf{2}*} = O(s_n)$ (see (A.9) and (A.10) in Aneiros and Vieu [4]), it is obtained that

$$\left|\left|\widehat{\theta}_0^{\mathbf{2}} - \theta_0^{\mathbf{2}}\right|\right| = O_p\left(n^{-1}d(h)s_n^{\gamma-3/2}\right). \tag{5.61}$$

On the other hand, note that $\theta_0^{\mathbf{2}}$ depends on the variable selection of the first stage. Then, we can ensure for all $\eta > 0$ that

$$\begin{aligned}
\mathbb{P}\left(\left|\left|\theta_0^{\mathbf{2}} - \theta_0\right|\right| \geq \eta n^{-1}d(h)s_n^{\gamma-3/2}\right) &\leq \mathbb{P}\left(\left|\left|\theta_0^{\mathbf{2}} - \theta_0\right|\right| \geq 0\right) \\
&= \mathbb{P}\left(\exists j \in S_n \text{ and } j \notin \mathcal{R}_n^{\mathbf{2}*}\right) \\
&\leq \mathbb{P}\left(\exists j = 1, \ldots, p_n,\ \beta_{0j} \neq 0 \text{ and } \widehat{\beta}_{0k_j}^{\mathbf{1}} = 0\right).
\end{aligned}$$

Therefore, using the first assertion in Proposition 5.6, we obtain for all $\eta > 0$

$$\mathbb{P}\left(\left|\left|\theta_0^{\mathbf{2}} - \theta_0\right|\right| \geq \eta n^{-1}d(h)s_n^{\gamma-3/2}\right) \to 0 \text{ as } n \to \infty. \tag{5.62}$$

Then, the desired result comes from the combination of (5.61) and (5.62) in (5.60). ■

5.6.2.3 Proof of (5.34)

As in Aneiros and Vieu [4], we can make the following decomposition:

$$\begin{aligned}
\mathbb{P}\left(\widehat{S}_n \neq S_n\right) &\leq \mathbb{P}\left(\exists j \in \mathcal{R}_n^{\mathbf{2}*},\ \beta_{0j}^{\mathbf{2}} \neq 0 \text{ and } \widehat{\beta}_{0j}^{\mathbf{2}} = 0\right) + \mathbb{P}\left(\exists j \in \widehat{\mathcal{S}}_n^{\mathbf{2}},\ \beta_{0j}^{\mathbf{2}} = 0\right) \\
&+ \mathbb{P}\left(\exists j = 1, \ldots, p_n,\ \beta_{0j} \neq 0 \text{ and } \widehat{\beta}_{0k_j}^{\mathbf{1}} = 0\right),
\end{aligned} \tag{5.63}$$

where we have denoted $\widehat{\mathcal{S}}_n^{\mathbf{2}} = \{j \in \mathcal{R}_n^{\mathbf{2}*},\ \widehat{\beta}_{0j}^{\mathbf{2}} \neq 0\}$.

On the one hand, considering $\mathcal{P}_n = \mathcal{R}_n^{\mathbf{2}*}$ in (5.26), the first two terms in the right

hand side of (5.63) tend to zero as $n \to \infty$.

On the other hand, applying the first assertion in Proposition 5.6, the third term in the right hand side of (5.63) tends to zero as $n \to \infty$.

As a consequence, $\mathbb{P}\left(\widehat{S}_n \neq S_n\right) \to 0$ as $n \to \infty$, and we obtain (5.34). ■

5.6.3 Proof of Theorem 5.9

It is easy to obtain that:

$$|\widehat{r}_\theta(\chi) - r_{\theta_0}(\chi)| \leq |\widehat{g}_{0\theta}(\chi) - g_{0\theta_0}(\chi)| + \left(\sharp\left(S_n \cup \widehat{S}_n\right)\right)^{1/2} \times \left(\sup_{u\in\mathcal{C}, j\in S_n\cup\widehat{S}_n, \theta\in\Theta_n} |\widehat{g}_{j\theta}(u) - g_{j\theta_0}(u)| \left|\left|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right|\right| + \sup_{u\in\mathcal{C}, j\in S_n\cup\widehat{S}_n} |g_{j\theta_0}(u)| \left|\left|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right|\right|\right) \tag{5.64}$$

On the one hand, using Condition (5.16) and (5.34) we obtain

$$\sharp\left(S_n \cup \widehat{S}_n\right) = O_p(s_n). \tag{5.65}$$

On the other hand, from (5.31) and (5.34) it is obtained

$$\sup_{u\in\mathcal{C}, j\in S_n\cup\widehat{S}_n} |g_{j\theta_0}(u)| = O_p(1). \tag{5.66}$$

As a consequence, using (5.64)-(5.66) and result (5.32), together with conditions (5.29), (5.30) and assumptions $b_n \to 0$ and $h \to 0$ as $n \to \infty$, the desired result (5.35) is obtained. ■

5.6.4 Proof of Corollary 5.10

It is a direct consequence of (5.34). ■

Chapter 6

Conclusions and future work

In this dissertation several regression models with a functional single-index component have been studied. This analysis has shown the great potential of semiparametric modelling when it is combined with appropriated estimation tools. The good behaviour of the proposed models and methodologies for estimation (in the FSIM and the SFPLSIM) or variable selection and estimation (in the SSFPLSIM and the MFPLSIM) was ensured both theoretically (formulating and proving asymptotic results) and in practical applications (by means of finite-sample Monte Carlo studies and real data applications). Precisely, in applications to real data, the superiority of the proposed models and procedures over other approaches was demonstrated: better predictive power (thanks to their flexibility and accuracy of the estimation tools) and estimation involving interpretable parameters. These facts allow us to prognosticate that this methodology can be very useful in applied areas.

However, from the applied point of view, the use of the proposed methodology requires the computational implementation of the procedures in a statistical software. This process could be a hard task (see, for instance, practical recommendations in Section 2.4). Precisely, in order to facilitate the practical use of the proposed methods, a package will be built in the statistical software R Core Team [93] containing methodology proposed on estimation and/or variable selection in FSIM, SFPLSIM, SSFPLSIM and MFPLSIM. This package will be available in a few months and will contain functions that will directly allow the adjustment of the studied models using the proposed methodology.

Finally, it is natural to think about continuing the work developed in this thesis from a methodological point of view. Our aspirations go in two directions:

- On the one hand, going much deeper into the SFPLSIM and the SSFPLSIM, including measurement error models, varying regression coefficients or responses missing at random (for related papers with scalar variables, see Zhao and Huang [117], Feng and Xue [44] and Lai and Wang [72], respectively; for related models with functional variables, see Zhu et al. [118], Luo et al. [80] and Ling et al. [78], respectively); and also considering selection of impact points in these three new versions.

- On the other hand, developing new methodology for semiparametric flexible models such as the *multiple-index model* (Ma [81] or Bouraine et al. [18]):

$$Y = r\left(\langle\theta_{0,1}, \mathcal{X}_1\rangle + \cdots + \langle\theta_{0,p}, \mathcal{X}_p\rangle\right) + \varepsilon, \tag{6.1}$$

  or the *functional projection pursuit* one (Chen et al. [27]; Ferraty et al. [50]):

$$Y = r_1\left(\langle\theta_{0,1}, \mathcal{X}\rangle\right) + \cdots + r_p\left(\langle\theta_{0,p}, \mathcal{X}\rangle\right) + \varepsilon. \tag{6.2}$$

  For example, Chapters 2 and 3 have highlighted the good behaviour of the proposed automatic and location-adaptive procedure, based on kNN ideas, in models that contain a functional single-index component. Precisely, location-adaptive estimates (such as kNN) and fully automatic procedures (such as cross-validation), as far as we know, have not been developed yet for general models (6.1) and (6.2). Our guess is that the uniform ideas developed in Chapters 2 and 3 could pave the way for that challenging purpose.

Asymptotic results will be obtained for the aforementioned procedures related to new versions of the SFPLSIM and the SSFPLSIM or the new studied models. Simulation studies and applications to real data will illustrate both the finite sample size behaviour and the usefulness of our proposals. We hope, in the medium term, to have developed some of the investigations that we have just mentioned.

Appendix A

Resumo en galego

Na actualidade, os avances tecnolóxicos na recollida e almacenamento de datos fan cada vez máis frecuente obter observacións de variables medidas nun continuo. Como consecuencia obtéñense medicións en forma de curvas, imaxes ou incluso estruturas máis complexas, en lugar de medidas escalares ou vectores como se obtiñan tradicionalmente. Daquela, en moitas ciencias aplicadas (como a medicina, quimiometría, biometría, econometría...) o estudo de fenómenos reais produce observacións de variables funcionais, é dicir, datos funcionais.

Dende o punto de vista estatístico, unha variable $\mathcal{X}$ considérase funcional se toma valores nun espazo de dimensión infinita (o espazo funcional). Daquela, un conxunto de datos funcional está composto por observacións de n variables funcionais $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ identicamente distribuídas a $\mathcal{X}$. Neste caso, "os átomos" do conxunto de datos son funcións aleatorias e os conxuntos de datos conteñen mostras desas funcións aleatorias. As variables funcionais teñen unha importante característica distintiva: teñen dimensión infinita, en contraste cos tipos usuais de datos atopados na Estatística. Debido a isto, os métodos estatísticos usados no contexto non-funcional (finito-dimensional) fallan cando traballamos con variables funcionais. De feito, o uso directo das técnicas tradicionais obrigaríanos a traballar coas observacións discretizadas das variables funcionais, o cal tería, polo menos, tres importantes desvantaxes: a presenza de correlacións fortes entre as variables resultantes, o desaproveitamento da orixe funcional da variable ou o problema da dimensión (o ratio entre o tamaño de mostra e o número de variables). Polo tanto, foi necesario

desenvolver nova metodoloxía específica para tratar os datos funcionais.

O termo Análise de Datos Funcionais (en inglés *Functional Data Analysis*, usualmente referido a través das súas siglas, FDA) foi acuñado por Ramsay [94] e Ramsay and Dalzell [95] para referirse a todas aquelas ferramentas estatísticas deseñadas para tratar con datos funcionais. Non obstante, ao comezo a produción científica na área foi esporádica. A popularización da FDA veu a finais dos anos noventa, a medida que os datos funcionais empezaron a aparecer de maneira abundante nas aplicacións e a medida que xorden varias monografías revisando unha selección de tópicos relacionados coa FDA (por exemplo, Bosq [17], Ramsay and Silverman [96], Ramsay and Silverman [97] ou Ferraty and Vieu [47]). Nas dúas últimas décadas, a FDA converteuse nunha das disciplinas principais da Estatística e existe unha ampla literatura cubrindo diferentes áreas (técnicas de redución da dimensión, correlación e regresión, clasificación supervisada e non supervisada...), pero existen aínda moitos retos metodolóxicos para analizar os datos funcionais debido a súa dimensión infinita (véxase, por exemplo, Aneiros et al. [9] para ter unha idea xeral dos tópicos nos que se traballa actualmente na disciplina).

Precisamente a dimensión foi unha das primeiras preocupacións da literatura en FDA. Os investigadores déronse conta de que transformar a mostra de datos funcionais en elementos dun espazo de dimensión finita permitía un tratamento estatístico máis simple e unha interpretación máis doada na práctica. Estes feitos leváronos ao desenvolvemento de técnicas de redución da dimensión como a análise de compoñentes principais funcionais (véxase Dauxois et al. [30], Silverman [103], Boente and Fraiman [16] ou Li and Hsing [73]) mínimos cadrados parciais (véxase Preda and Saporta [91], Krämer et al. [70], Delaigle and Hall [32] ou Aguilera et al. [2] no contexto da regresión, Preda et al. [92] no marco da clasificación supervisada e Reiss and Ogden [99] ou Febrero-Bande et al. [43] para unha comparación entre compoñentes principais funcionais e mínimos cadrados parciais) ou a selección de variables no contexto da regresión (para a extensión de ideas procedentes do marco multivariante, tales como Tibshirani [104] ou Fan and Lv [40], véxase Aneiros and Vieu [4] ou Aneiros and Vieu [5]).

Outro dos tópicos fundamentais na FDA é a regresión. A regresión é unha ferramenta usualmente empregada con dous obxectivos principais: por unha banda,

modelar a dependencia entre unha variable de interese (a variable resposta) e outras variables (as variables explicativas ou covariables) que usualmente son máis fáciles de obter ou medir; por outra banda, usar o modelo proposto para predicir o valor da resposta usando novos valores das variables explicativas. Os problemas de regresión foron amplamente estudados para variables reais ou multivariantes e a medida que os datos funcionais se fixeron frecuentes, os investigadores interesáronse en relacionar as variables funcionais con outras variables de interese (funcionais ou non). Como consecuencia, existe unha extensa literatura en regresión funcional (véxase Greven and Scheipl [57] para unha presentación xeral). Centrándonos no caso de resposta escalar e covariables funcionais, esta literatura céntrase ou ben en modelos paramétricos (véxase o Capítulo 11 de Hsing and Eubank [63]) ou ben en modelos non-paramétricos (popularizados por Ferraty and Vieu [47]; véxase Geenens [53] ou Ling and Vieu [76] para revisións recentes). Non obstante, a regresión semi-paramétrica é aínda un campo moi pouco desenvolvido na FDA (véxase Goia and Vieu [54] para unha revisión recente). O contexto semi-paramétrico constitúe un excelente punto medio entre a metodoloxía paramétrica e a non-paramétrica, superando a estas dúas en moitos sentidos, xa que permite flexibilidade (a diferencia dos modelos paramétricos) e tamén interpretabilidade e redución da dimensión no contexto funcional (a diferenza do modelado non-paramétrico). Estas propiedades da regresión semi-paramétrica resultan fundamentais no contexto funcional, como sinalan varios estudos recentes (véxase Cuevas [29], Goia and Vieu [55], Vieu [108] ou Aneiros et al. [8]), e convértena nunha ferramenta transversal ás técnicas de redución da dimensión.

Dado que o campo da regresión semi-paramétrica funcional con resposta escalar está moi pouco desenvolvido e os avances nesta área teñen un enorme interese na actualidade, nesta tese estudamos varios modelos de regresión semi-paramétricos que permiten incluír unha ou varias variables funcionais. Os avances que aportamos centráronse na tarefa de estimación destes modelos, establecendo propiedades teóricas dos estimadores derivados e analizando o seu comportamento na práctica con mostras finitas (tanto por medio de datos simulados como por medio de datos reais). A tese está dividida en seis capítulos: o Capítulo 1 contén unha introdución á regresión funcional; os Capítulos 2, 3, 4 e 5, recollen as novas contribucións metodolóxicas

aportadas, estando cada un deles adicado a un modelo de regresión funcional semi-paramétrico; finalmente, o Capítulo 6 contén unhas breves conclusións, así como o traballo que se prevé realizar no futuro próximo. A continuación detállase o contido de cada un destes capítulos.

Capítulo 1: Cara a regresión semi-paramétrica funcional

Neste capítulo proporciónase unha breve introdución do contexto estatístico no que se sitúa esta tese. Nel presentamos os modelos de regresión semi-paramétricos que imos estudar. Xunto con eles, imos describir tamén outras formas de modelado propostas na literatura, as cales constitúen un piar para os modelos estudados e servirannos para comparalos con eles nas aplicacións a datos reais. A presentación de modelos comeza cos máis simples para finalmente lidar coas estruturas máis complexas coas que traballamos nesta tese. Deste xeito, na Sección 1.2 introducimos a regresión con resposta escalar e unicamente unha variable explicativa funcional. Na Sección 1.3 presentaremos modelos que combinan de maneira aditiva covariables escalares con efecto linear cunha covariable funcional con efecto non-linear. Finalmente, na Sección 1.4 faremos unha introdución á regresión *sparse* (termo inglés que indica que do conxunto de todas as variables explicativas, unicamente algunhas, moi poucas, teñen influencia na resposta, polo que moitos dos coeficientes asociados van valer cero), centrándonos nos modelos que imos estudar e naqueles cos que os imos comparar.

Capítulo 2: Contribucións no modelo de índice único funcional

Neste capítulo desenvólvese un amplo estudo dun modelo de regresión semi-paramétrico: o modelo de índice único funcional (coñecido como FSIM, iniciais do nome do modelo en inglés *functional single-index model*). As investigacións desenvolvidas neste capítulo céntranse na estimación deste modelo e están publicadas no artigo Novo, Aneiros, and Vieu [87] da revista *Journal of Nonparametric Statistics*.

Na Sección 2.2 estúdase un novo procedemento automático, con ventá dependente da variable explicativa funcional, para estimar a regresión no modelo FSIM. Este procedemento de estimación está baseado no método de k-veciños-máis-próximos (coñecido como kNN, iniciais de *k-Nearest-Neighbours*). Na Sección 2.3 realízase un estudo asintótico da estimación da regresión por medio do estatístico kNN que

inclúe resultados de consistencia uniforme sobre todos os parámetros involucrados. Aínda que o noso obxectivo era estudar o procedemento kNN, para establecer estes resultados obtivemos, e usamos como ferramentas preliminares, novos resultados análogos para o estimador núcleo de tipo Nadaraya-Watson (o máis usado na literatura e na práctica). Os resultados obtidos xeneralizan os proporcionados Kara-Zaitri et al. [66] e Kara-Zaitri et al. [67] no caso do modelo funcional non-paramétrico. Unha das principais características das taxas de converxencia que nós acadamos é que son similares ás obtidas nos problemas unidimensionais, dando evidencias da propiedade de redución da dimensión que proporciona a metodoloxía estudada. Ademais, unha consecuencia importante destes resultados asintóticos é que dan validez teórica aos selectores de todos os parámetros involucrados na estimación obtidos de xeito automático a partir da mostra. Isto fai a ambos procedementos, núcleo e kNN, directamente utilizables na práctica.

A Sección 2.4 contén algúns consellos para abordar certos problemas prácticos relacionados coa metodoloxía presentada. Tales suxestións apóianse na Sección 2.5 por medio dun estudo de simulación que, ademais, compara o comportamento práctico dos procedementos núcleo e kNN. Neste estudo de simulación pode verse que o método kNN supera amplamente o método núcleo en eficiencia preditiva baixo heteroxeneidade. Na Sección 2.6 ilustramos a metodoloxía presentada por medio dun conxunto de datos reais de referencia, os datos do Tecator. Neste caso tamén o método kNN ofrece mellores resultados ca o método núcleo. Ademais, mostramos que o carácter semi-paramétrico do FSIM non só proporciona un bo poder preditivo, senón que tamén permite obter resultados facilmente interpretables e representables.

Finalmente, na Sección 2.7 recóllense as probas dos resultados teóricos presentados na Sección 2.3.

Capítulo 3: Contribucións no modelo semi-funcional parcialmente linear de índice único

Neste capítulo estudamos a estimación do modelo de índice único parcialmente linear semi-funcional (coñecido como SFPLSIM, iniciais do nome do modelo en inglés *semi-functional partial linear single-index model*). Unha das características fundamentais deste modelo é que permite incluír como preditores un vector multivariante e unha

variable funcional. Ademais, o vector multivariante entra de xeito parcialmente linear, mentres que o preditor funcional entra por medio dunha estrutura de índice único, o que convirte a este modelo nun modelo semi-paramétrico, polo que herdará todas as boas propiedades mencionadas para este tipo de modelado. Os resultados obtidos neste capítulo están publicados no artigo Novo, Aneiros, and Vieu [89] da revista *Statistics and Probability Letters*.

Na Sección 3.2 estudamos o método kNN para realizar a estimación da compoñente funcional de índice único do SFPLSIM. Na Sección 3.3 estendemos os resultados teóricos obtidos no Capítulo 2 para este novo modelo, tanto para a estimación por medio do procedemento kNN como por medio do procedemento tipo núcleo.

Os resultados asintóticos obtidos acompañáronse de simulacións (Sección 3.4), as cales poñen de manifesto as vantaxes do procedemento kNN fronte a estimación tipo núcleo. Finalmente, na Sección 3.5 analizamos os datos do Tecator de novo, e cun estudo comparativo mostramos tamén que o modelado semi-paramétrico supera outras formas de modelado existentes.

A Sección 3.6 recolle as probas dos resultados teóricos presentados na Sección 3.3.

Capítulo 4: Contribucións no modelo de índice único semi-funcional parcialmente linear sparse

Neste capítulo realizamos un amplo estudo relativo á estimación do modelo de índice único parcialmente linear semi-funcional sparse (coñecido como SSFPLSIM, siglas do nome en inglés *sparse semi-functional partial linear single-index model*). Neste modelo semi-paramétrico os preditores son unha mestura de variable funcional, que entra no modelo por medio dunha estrutura de índice único, e un vector de alta dimensión, que entra no modelo de xeito linear. Desta maneira, o modelo SSFPLSIM proposto é unha xeneralización do modelo SFPLSIM estudado no Capítulo 3 ao caso de ter un número diverxente de covariables na compoñente linear (que tende a infinito a medida que o tamaño de mostra tende a infinito) e que soamente algunhas delas, moi poucas, inflúan na resposta (é un modelo de regresión *sparse*). O noso obxectivo é ser capaces de proporcionar unha redución da dimensión, por medio do modelado semi-paramétrico e a selección de variables na compoñente linear e,

ao mesmo tempo, conseguir unha metodoloxía flexible con bo poder preditivo. As investigacións desenvolvidas neste capítulo foron publicadas no artigo Novo, Aneiros, and Vieu [88] da revista TEST.

Na Sección 4.2 proponse un procedemento de selección de variables e simultánea estimación da compoñente linear do modelo baseado en mínimos cadrados penalizados. Tamén se propuxo un estimador para a compoñente funcional de índice único (baseado na estimación tipo núcleo). Na Sección 4.3 obtivemos unha ampla variedade de resultados asintóticos relativos ao procedemento presentado: dende taxas de converxencia dos estimadores ata o comportamento asintótico do procedemento de selección de variables. Debemos salientar que as taxas de converxencia obtidas para o estimador da compoñente linear son as mesmas que obtiveron nun escenario menos complexo ca o noso Aneiros et al. [7] (e as mesmas que as acadadas en Fan and Lv [40] no contexto do modelo linear). Ademais, demostramos que o procedemento de selección de variables proposto satisfai a *propiedade do oráculo* (véxase Fan and Li [38]) e que a compoñente funcional de índice único do modelo se estima coa mesma taxa que no caso de que a variable fose unidimensional (confirmándose así a propiedade de redución da dimensión).

Na Sección 4.4 analizamos o comportamento práctico da metodoloxía proposta mediante mostras simuladas, mentres que na Sección 4.5 mostramos a súa utilidade no modelado de datos reais mediante unha aplicación ao conxunto de datos do Tecator. Por medio desta aplicación comprobamos as grandes vantaxes que aporta a metodoloxía presentada neste capítulo fronte a outras alternativas existentes na literatura.

Finalmente, a Sección 4.6 contén as probas dos resultados teóricos presentados na Sección 4.3.

Capítulo 5: Contribucións no modelo de índice único bi-funcional parcialmente linear sparse

Neste capítulo estúdase de xeito detallado a estimación do modelo de índice único parcialmente linear multi-funcional (MFPLSIM, siglas do nome en inglés *sparse multi-functional partial linear single-index model*). Este modelo permite incorporar a influencia na variable resposta escalar de dúas ou máis variables aleatorias funcionais,

aínda que por simplicidade imos centrar a exposición no caso de ter unicamente dúas variables funcionais (é dicir, no caso bi-funcional). Unha das variables funcionais inclúese no modelo por medio dunha estrutura de índice único e a outra linearmente, pero mediante o vector de alta dimensión formado polas súas observacións discretizadas. É dicir, este modelo permite incorporar tanto efectos continuos como efectos puntuais das variables funcionais involucrando parámetros interpretables en ambos casos. Ademais, debemos notar que o modelo MFPLSIM é unha adaptación do modelo SSFPLSIM estudado no Capítulo 4 ao caso no que as covariables na compoñente linear teñen orixe funcional. O Capítulo 5 ten por obxectivo lidar coa característica sparse do modelo MFPLSIM. Como neste caso as covariables presentes na parte linear proveñen da discretización dunha curva, van presentar fortes correlacións entre elas e, ademais, é de esperar que o número covariables con efecto linear resultante sexa moi grande (moito maior que o tamaño de mostra). Nestas condicións a aplicación directa do método de selección de variables presentado no Capítulo 4 (ou doutros métodos de selección de variables tradicionais) vólvese completamente inviable: por unha banda, necesitaríase un enorme tempo de cálculo para realizar a selección de variables, incluso para valores moderados do tamaño de discretización; por outra banda, este procedemento non ten en conta a orixe funcional das variables e, ademais, poderíase ver afectado negativamente pola presenza desas fortes correlacións existentes entre elas. Por tales razóns é preciso desenvolver novos métodos de selección de variables no MFPLSIM. As investigacións contidas neste capítulo forman parte do artigo Novo, Aneiros, and Vieu [90], o cal foi enviado a unha revista para a súa publicación.

Na Sección 5.2 preséntanse dous novos algoritmos para seleccionar variables (na compoñente linear) e estimar o modelo MFPLSIM. Ambos procedementos aproveitan a orixe funcional das covariables da parte linear.

Na Sección 5.3 realizamos varios estudos de simulación que mostran o ámbito de aplicación dos dous métodos: o primeiro algoritmo proporciona unha solución, sen perda de poder de poder de predición, ao grande tempo computacional que precisan os métodos como o presentado no Capítulo 4 para estimar o modelo MFPLSIM. Ademais, ao non precisar a división da mostra, proporciona mellores resultados que o segundo algoritmo no caso de mostras de tamaño pequeno. O segundo algoritmo,

aínda que precisa en xeral máis tempo de computación que o primeiro, mellora a eficiencia preditiva deste no caso de contarmos cun tamaño de mostra moderado, e tamén realiza unha selección de variables relevantes máis precisa. Na Sección 5.5, a aplicación a datos reais dunha planta de azucre ilustraranos o triplo interese da metodoloxía presentada, a cal proporciona grande poder de predición xunto con resultados interpretables e un tempo de computación razoablemente baixo.

Na Sección 5.6 recóllense as probas dos resultados teóricos expostos na Sección 5.3.

Capítulo 6: Conclusións e traballo futuro

Neste capítulo proporciónase un breve resumo das conclusións ás que chegamos na tese: despois de todo o traballo realizado, puidemos ver o gran potencial do modelado semi-paramétrico ao combinalo coas ferramentas de estimación (nos modelos FSIM e SFPLSIM) ou de estimación e selección de variables (nos modelos SSFPLSIM e MFPLSIM) axeitadas.

Tamén indicamos neste capítulo algunhas ideas de traballo futuro. Unha delas é a realización dun paquete no software estatístico R Core Team [93] que axude á utilización na práctica de toda a metodoloxía presentada, dado que a implementación da mesma require coñecementos de programación xunto con familiaridade coas ferramentas estatísticas empregadas. Ademais, comentamos as liñas nas que temos previsto continuar coas contribucións no modelado semi-paramétrico, tanto no que se refire a afondar nos modelos estudados, como no que incumbe á proposta de novos modelos.

Bibliography

[1] A. M. Aguilera, M. Escabias, C. Preda, and G. Saporta. Using basis expansions for estimating functional PLS regression: Applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2):289–305, 2010. URL http://www.sciencedirect.com/science/article/pii/S0169743910001747.

[2] A. M. Aguilera, M. C. Aguilera-Morillo, and C. Preda. Penalized versions of functional PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 154:80–92, 2016. URL http://www.sciencedirect.com/science/article/pii/S0169743916300491.

[3] A. Ait-Saïdi, F. Ferraty, R. Kassa, and P. Vieu. Cross-validated estimations in the single-functional index model. *Statistics*, 42(6):475–494, 2008. URL https://doi.org/10.1080/02331880801980377.

[4] G. Aneiros and P. Vieu. Variable selection in infinite-dimensional problems. *Statistics & Probability Letters*, 94:12–20, 2014. URL http://www.sciencedirect.com/science/article/pii/S0167715214002363.

[5] G. Aneiros and P. Vieu. Partial linear modelling with multi-functional covariates. *Computational Statistics*, 30(3):647–671, 2015. URL https://doi.org/10.1007/s00180-015-0568-8.

[6] G. Aneiros and P. Vieu. Sparse nonparametric model for regression with functional covariate. *Journal of Nonparametric Statistics*, 28(4):839–859, 2016. URL https://doi.org/10.1080/10485252.2016.1234050.

[7] G. Aneiros, F. Ferraty, and P. Vieu. Variable selection in partial linear regression with functional covariate. *Statistics*, 49(6):1322–1347, 2015. URL https://doi.org/10.1080/02331888.2014.998675.

[8] G. Aneiros, R. Cao, R. Fraiman, C. Genest, and P. Vieu. Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis*, 170:3–9, 2019. URL http://www.sciencedirect.com/science/article/pii/S0047259X1830561X.

[9] G. Aneiros, R. Cao, and P. Vieu. Editorial on the special issue on functional data analysis and related topics. *Computational Statistics*, 34:447–450, 2019. URL https://doi.org/10.1007/s00180-019-00892-0.

[10] G. Aneiros, S. Novo, and P. Vieu. Variable selection in functional regression models: a review. 2021, submitted.

[11] G. Aneiros-Pérez and P. Vieu. Semi-functional partial linear regression. *Statistics & Probability Letters*, 76(11):1102–1110, 2006. URL http://www.sciencedirect.com/science/article/pii/S0167715205004530.

[12] G. Aneiros-Pérez and P. Vieu. Nonparametric time series prediction: A semi-functional partial linear modeling. *Journal of Multivariate Analysis*, 99(5): 834–857, 2008. URL http://www.sciencedirect.com/science/article/pii/S0047259X07000644.

[13] G. Aneiros-Pérez and P. Vieu. Automatic estimation procedure in partial linear model with functional data. *Statistical Papers*, 52:751–7717, 2011. URL https://doi.org/10.1007/s00362-009-0280-2.

[14] S. Bakin. *Adaptive Regression and Model Selection in Data Mining Problems*. PhD thesis, Australian National University, Canberra, 1999. URL https://openresearch-repository.anu.edu.au/handle/1885/9449.

[15] G. Biau, F. Cérou, and A. Guyader. Rates of convergence of the functional k-Nearest Neighbor estimate. *IEEE Transactions on Information Theory*, 56 (4):2034–2040, 2010. URL https://doi.org/10.1109/TIT.2010.2040857.

[16] G. Boente and R. Fraiman. Kernel-based functional principal components. *Statistics & Probability Letters*, 48(4):335–345, 2000. URL http://www.sciencedirect.com/science/article/pii/S0167715200000146.

[17] D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Lecture Notes in Statistics. Springer-Verlag, New York, 2000.

[18] M. Bouraine, A. Aït-Saidi, F. Ferraty, and P. Vieu. Choix optimal de l'indice multi-fonctionnel: méthode de validation croisée. *Revue Roumaine de Mathématiques Pures et Appliquées*, 5(5):355–367, 2010.

[19] P. Breheny and J. Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25:173–187, 2015. URL https://doi.org/10.1007/s11222-013-9424-2.

[20] L. Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996. URL https://doi.org/10.1214/aos/1032181158.

[21] P. Bühlmann and L. Meier. Discussion: One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1534–1541, 2008. URL https://doi.org/10.1214/07-AOS0316A.

[22] F. Burba, F. Ferraty, and P. Vieu. k-Nearest Neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics*, 21(4):453–469, 2009. URL https://doi.org/10.1080/10485250802668909.

[23] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2392–2404, 2007. URL https://projecteuclid.org/euclid.aos/1201012958.

[24] H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics and Probability Letters*, 45(1):11–22, 1999. URL http://www.sciencedirect.com/science/article/pii/S016771529900036X.

[25] H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–591, 2003. URL http://www3.stat.sinica.edu.tw/statistica/oldpdf/A13n31.pdf.

[26] R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438): 477–489, 1997. URL https://doi.org/10.1080/01621459.1997.10474001.

[27] D. Chen, P. Hall, and H. Müller. Single and multiple index functional regression models with nonparametric link. *Annals of Statistics*, 39(3):1720–1747, 2011. URL https://doi.org/10.1214/11-AOS882.

[28] G. Collomb. Estimation de la régression par la méthode des k points les plus proches: propriétés de convergence ponctuelle, (french). *Comptes Rendus de l'Académie des Sciences*, pages 245–247, 1979.

[29] A. Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2014. URL http://www.sciencedirect.com/science/article/pii/S0378375813000748.

[30] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154, 1982. URL http://www.sciencedirect.com/science/article/pii/0047259X82900884.

[31] C. de Boor. *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer-Verlag, New York, 2001.

[32] A. Delaigle and P. Hall. Methodology and theory for partial least squares applied to functional data. *Annals of Statistics*, 40(1):322–352, 2012. URL https://doi.org/10.1214/11-AOS958.

[33] L. Devroye, L. Gyorfi, A. Krzyzak, and G. Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22(3):1371–1385, 1994. URL https://doi.org/10.1214/aos/1176325633.

[34] J. Dony and U. Einmahl. Uniform in bandwidth consistency of kernel regression estimators at a fixed point. In C. Houdré, V. Koltchinskii, D. M. Mason, and M. Peligrad, editors, *High Dimensional Probability V: The Luminy Volume*, volume 5 of *Collections*, pages 308–325. Institute of Mathematical Statistics, Beachwood, Ohio, 2009. URL https://doi.org/10.1214/09-IMSCOLL520.

[35] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004. URL https://projecteuclid.org/euclid.aos/1083178935.

[36] M. A. Efroymson. Multiple regression analysis. In *Mathematical Methods for Digital Computers*, New York, 1960. Wiley.

[37] J. Fan. Comments on "Wavelets in Statistics: A review" by A. Antoniadis. *Journal of the Italian Statistical Society*, 6:131, 1997. URL https://doi.org/10.1007/BF03178906.

[38] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001. URL https://doi.org/10.1198/016214501753382273.

[39] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00674.x.

[40] J. Fan and J. Lv. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011. URL https://ieeexplore.ieee.org/document/5961830.

[41] J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32(3):928–961, 2004. URL https://doi.org/10.1214/009053604000000256.

[42] M. Febrero-Bande and M. Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012. URL http://www.jstatsoft.org/v51/i04/.

[43] M. Febrero-Bande, P. Galeano, and W. González-Manteiga. Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review*, 85(1): 61–83, 2017. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12116.

[44] S. Feng and L. Xue. Variable selection for partially varying coefficient single-index model. *Journal of Applied Statistics*, 40(12):2637–2652, 2013. URL https://doi.org/10.1080/02664763.2013.823919.

[45] S. Feng and L. Xue. Partially functional linear varying coefficient model. *Statistics*, 50(4):717–732, 2016. URL https://doi.org/10.1080/02331888.2016.1138954.

[46] F. Ferraty and P. Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17:545–564, 2002. URL https://doi.org/10.1007/s001800200126.

[47] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis, Theory and Practice*. Springer Series in Statistics. Springer-Verlag, New York, 2006.

[48] F. Ferraty, A. Peuch, and P. Vieu. Modéle á indice fonctionnel simple. *Comptes Rendus Mathématique de l'Académie des Sciences Paris*, 336(12):1025–1028, 2003. URL https://www.sciencedirect.com/science/article/pii/S1631073X03002395.

[49] F. Ferraty, A. Laksaci, A. Tadj, and P. Vieu. Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical Planning and Inference*, 140(2):335–352, 2010. URL http://www.sciencedirect.com/science/article/pii/S0378375809002316.

[50] F. Ferraty, A. Goia, E. Salinelli, and P. Vieu. Functional projection pursuit regression. *Test*, 22:293–320, 2013. URL https://doi.org/10.1007/s11749-012-0306-2.

[51] I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1993.10485033.

[52] G. M. Furnival and R. W. Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511, 1974. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1974.10489231.

[53] G. Geenens. Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys*, 5:30–43, 2011. URL https://doi.org/10.1214/09-SS049.

[54] A. Goia and P. Vieu. Some advances in semiparametric functional data modelling. In *Contributions in Infinite-Dimensional Statistics and Related Topics*, pages 135–141. Bologna: Esculapio, 2014.

[55] A. Goia and P. Vieu. An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis*, 146:1–6, 2016. URL http://www.sciencedirect.com/science/article/pii/S0047259X15003176.

[56] U. Grenander. Stochastic processes and statistical inference. *Arkiv för Matematik*, 1(3):195–277, 1950. URL https://doi.org/10.1007/BF02590638.

[57] S. Greven and F. Scheipl. A general framework for functional regression modelling. *Statistical Modelling*, 17(1–2):1–35, 2017. URL https://doi.org/10.1177/1471082X16681317.

[58] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. New York: Springer, 2002. URL http://dx.doi.org/10.1007/b97848.

[59] W. Härdle and S. Marron. Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, 13(4):1465–1481, 1985. URL https://doi.org/10.1214/aos/1176349748.

[60] W. Härdle, P. Hall, and H. Ichimura. Optimal smoothing in single-index models. *Annals of Statistics*, 21(1):157–178, 1993. URL https://doi.org/10.1214/aos/1176349020.

[61] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer-Verlag, Berlin, Heidelberg, 2004. URL https://doi.org/10.1007/978-3-642-17146-8.

[62] T. Hastie, R. Tibshirani, and J. Friedman. Linear methods for regression. In *The Elements of Statistical Learning*, New York, 2009. Springer Series in Statistics. URL https://doi.org/10.1007/978-0-387-84858-7_3.

[63] T. Hsing and R. Eubank. *Theoretical Foundations to Functional Data Analysis with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 2015. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781118762547.

[64] J. Huang, J. L. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, 36(2):587–613, 2008. URL https://projecteuclid.org/euclid.aos/1205420512.

[65] J. Huang, S. Ma, and C-H. Zhang. Adaptive LASSO for sparse high-dimensional regression. *Statistica Sinica*, 18:1606–1618, 2008.

[66] L. Kara-Zaitri, A. Laksaci, M. Rachdi, and P. Vieu. Data-driven kNN estimation in nonparametric functional data analysis. *Journal of Multivariate Analysis*, 153:176–188, 2017. URL http://www.sciencedirect.com/science/article/pii/S0047259X16301105.

[67] L. Kara-Zaitri, A. Laksaci, M. Rachdi, and P. Vieu. Uniform in bandwidth consistency for various kernel estimators involving functional data. *Journal of Nonparametric Statistics*, 29(1):85–107, 2017. URL https://doi.org/10.1080/10485252.2016.1254780.

[68] K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae*, 7, 1946.

[69] A. N. Kolmogorov and V. M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspekhi Mat. Nauk*, 14:3–86, 1959.

[70] N. Krämer, A-L. Boulesteix, and G. Tutz. Penalized partial least squares with applications to B-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):60–69, 2008. URL http://www.sciencedirect.com/science/article/pii/S0169743908001214.

[71] N. L. Kudraszow and P. Vieu. Uniform consistency of kNN regressors for functional variables. *Statistics and Probability Letters*, 83(8):1863–1870, 2013. URL http://www.sciencedirect.com/science/article/pii/S0167715213001387.

[72] P. Lai and Q. Wang. Partially linear single-index model with missing responses at random. *Journal of Statistical Planning and Inference*, 141(2):1047–1058, 2011. URL http://www.sciencedirect.com/science/article/pii/S0378375810004234.

[73] Y. Li and T. Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, 38(6):3321–3351, 2010. URL https://doi.org/10.1214/10-AOS813.

[74] H. Lian. Functional partial linear model. *Journal of Nonparametric Statistics*, 23(1):115–128, 2011. URL https://doi.org/10.1080/10485252.2010.500385.

[75] H. Liang, X. Liu, R. Li, and C-L. Tsai. Estimation and testing for partially linear single-index models. *The Annals of Statistics*, 38(6):3811–3836, 2010. URL https://doi.org/10.1214/10-AOS835.

[76] N. Ling and P. Vieu. Nonparametric modelling for functional data: selected survey and tracks for future. *Statistics*, 52(4):934–949, 2018. URL https://doi.org/10.1080/02331888.2018.1487120.

[77] N. Ling, G. Aneiros, and P. Vieu. kNN estimation in functional partial linear modeling. *Statistical Papers*, 61:423–444, 2017. URL https://doi.org/10.1007/s00362-017-0946-0.

[78] N. Ling, R. Kan, P. Vieu, and S. Meng. Semi-functional partially linear regression model with responses missing at random. *Metrika*, 82:39–70, 2019. URL https://doi.org/10.1007/s00184-018-0688-6.

[79] N. Ling, S. Meng, and P. Vieu. Uniform consistency rate of kNN regression estimation for functional time series data. *Journal of Nonparametric Statistics*, 31(2):451–468, 2019. URL https://doi.org/10.1080/10485252.2019.1583338.

[80] X. Luo, L. Zhu, and H. Zhu. Single-index varying coefficient model for functional responses. *Biometrics*, 72(4):1275–1284, 2016. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12526.

[81] S. Ma. Estimation and inference in functional single-index models. *Annals of the Institute of Statistical Mathematics*, 68(1):181–208, 2016. URL https://doi.org/10.1007/s10463-014-0488-3.

[82] J. S. Marron. An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Annals of Statistics*, 13(3):1011–1023, 1985. URL https://doi.org/10.1214/aos/1176349653.

[83] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the LASSO. *The Annals of Statistics*, 34(3):1436–1462, 2006. URL https://doi.org/10.1214/009053606000000281.

[84] H-G. Müller. Peter Hall, functional data analysis and random objects. *Annals of Statistics*, 44(5):1867–1887, 2016. URL https://doi.org/10.1214/16-AOS1492.

[85] S. Müller and J. Dippon. k-NN kernel estimate for nonparametric functional regression in time series analysis. Technical report, University of Stuttgart, Fachbereich Mathematik, 2014.

[86] E. A. Nadaraya. On estimating regression. *Theory of Probability and Application*, 9:141–142, 1964.

[87] S. Novo, G. Aneiros, and P. Vieu. Automatic and location-adaptive estimation in functional single-index regression. *Journal of Nonparametric Statistics*, 31(2):364–392, 2019. URL https://doi.org/10.1080/10485252.2019.1567726.

[88] S. Novo, G. Aneiros, and P. Vieu. Sparse semiparametric regression when predictors are mixture of functional and high-dimensional variables. *TEST*, 30:481–504, 2021. URL https://doi.org/10.1007/s11749-020-00728-w.

[89] S. Novo, G. Aneiros, and P. Vieu. A kNN procedure in semiparametric functional data analysis. *Statistics & Probability Letters*, 171:109028, 5 pages, 2021. URL http://www.sciencedirect.com/science/article/pii/S016771522030331X.

[90] S. Novo, G. Aneiros, and P. Vieu. Fast and efficient algorithms for sparse semiparametric bi-functional regression. 2021, submitted.

[91] C. Preda and G. Saporta. PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 48(1):149–158, 2005. URL http://www.sciencedirect.com/science/article/pii/S0167947303002366.

[92] C. Preda, G. Saporta, and C. Lévéder. PLS classification of functional data. *Computational Statistics*, 22:223–235, 2007. URL https://doi.org/10.1007/s00180-007-0041-4.

[93] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

[94] J. O. Ramsay. When the data are functions. *Psychometrika*, 47:379–96, 1982. URL https://doi.org/10.1007/BF02293704.

[95] J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):

539–561, 1991. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1991.tb01844.x.

[96] J. O. Ramsay and B. Silverman. *Applied Functional Data Analysis Methods and Case Studies*. Springer Series in Statistics. Springer-Verlag, New York, 1st edition, 2002.

[97] J. O. Ramsay and B. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2nd edition, 2005.

[98] J. B. Ramsey. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(2):350–371, 1969. URL http://www.jstor.org/stable/2984219.

[99] P. T. Reiss and R. T. Ogden. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996, 2007. URL https://doi.org/10.1198/016214507000000527.

[100] E. M. Scott. The role of Statistics in the era of big data: Crucial, critical and under-valued. *Statistics & Probability Letters*, 136:20–24, 2018. URL http://www.sciencedirect.com/science/article/pii/S0167715218300956.

[101] H. L. Shang. Bayesian bandwidth estimation for a semi-functional partial linear regression model with unknown error density. *Computational Statistics*, 29:829–848, 2014. URL https://doi.org/10.1007/s00180-013-0463-0.

[102] Y. Shi, J. Huang, Y. Jiao, and Q. Yang. A semismooth Newton algorithm for high-dimensional nonconvex sparse learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2993–3006, 2020. URL https://ieeexplore.ieee.org/document/8835076.

[103] B. W. Silverman. Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, 24(1):1–24, 1996. URL https://doi.org/10.1214/aos/1033066196.

[104] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996. URL https://doi.org/10.1007/s00180-013-0463-0.

[105] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00490.x.

[106] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996. URL https://doi.org/10.1007/978-1-4757-2545-2.

[107] P. Vieu. Nonparametric regression: Optimal local bandwidth choice. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):453–464, 1991. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1991.tb01837.x.

[108] P. Vieu. On dimension reduction models for functional data. *Statistics & Probability Letters*, 136:134–138, 2018. URL http://www.sciencedirect.com/science/article/pii/S0167715218300774.

[109] G. Wang and Z. Zhu. Variable selection for the partial linear single-index model. *Acta Mathematicae Applicatae Sinica, English Series*, 33:373–388, 2017. URL https://doi.org/10.1007/s10255-017-0666-1.

[110] G. Wang, X-N. Feng, and M. Chen. Functional partial linear single-index model. *Scandinavian Journal of Statistics*, 43(1):261–274, 2016. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12178.

[111] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the LAD-LASSO. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

[112] J. L. Wang, J. M. Chiou, and H. G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, 2016. URL https://doi.org/10.1146/annurev-statistics-041715-033624.

[113] G. Watson. Smooth regression analysis. *Sankhyã: The Indian Journal of Statistics Series A*, 26:359–372, 1964.

[114] S. Weisberg. *Applied Linear Regression*. Wiley, New York, 1980.

[115] Y. Xia and W. K. Li. On single-index coefficient regression models. *Journal of the American Statistical Association*, 94(448):1275–1285, 1999. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473880.

[116] H. Xie and J. Huang. Scad-penalized regression in high-dimensional partially linear models. *Annals of Statistics*, 37(2):673–696, 2009. URL https://doi.org/10.1214/07-AOS580.

[117] X. Zhao and Z. Huang. Varying-coefficient single-index measurement error model. *Journal of Applied Statistics*, 45(12):2128–2144, 2018. URL https://doi.org/10.1080/02664763.2017.1410528.

[118] H. Zhu, R. Zhang, and G. Zhu. Estimation and inference in semi-functional partially linear measurement error models. *Journal of Systems Science and Complexity*, 33:1179–1199, 2020. URL https://doi.org/10.1007/s11424-019-8045-z.

[119] H. Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. URL https://doi.org/10.1198/016214506000000735.

[120] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2):301–320, 2005. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x.