Pointwise forecast, confidence and prediction intervals in electricity demand and price

Paula Raña Míguez

Doctoral Thesis

Universidade da Coruña

2016

Pointwise forecast, confidence and prediction intervals in electricity demand and price

Paula Raña Míguez

Doctoral Thesis 2016

Supervised by: Juan M. Vilar Fernández, Germán Aneiros Pérez

Programa Oficial de Doutoramento en Estatística
e Investigación Operativa

Departamento de Matemáticas
Universidade da Coruña

UNIVERSIDADE DA CORUÑA

Los abajo firmantes hacen constar que son los directores de la Tesis Doctoral titulada "Pointwise forecast, confidence and prediction intervals in electricity demand and price", realizada por Paula Raña Míguez 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 su autora proceda a su presentación y posterior defensa.

Os abaixo asinantes fan constar que son os directores da Tese Doutoral titulada "Pointwise forecast, confidence and prediction intervals in electricity demand and price", desenvolta por Paula Raña Míguez na Universidade da Coruña no marco do programa interuniversitario (UDC, USC e UVigo) de doutoramento en Estatística e Investigación de Operacións, dando o seu consentimento para que a súa autora proceda a súa presentación e posterior defensa.

A Coruña, 26 de setembro de 2016.

Directores:

Dr. Juan M. Vilar Fernández

Dr. Germán Aneiros Pérez

Doctoranda:

Paula Raña Míguez

Si caminas solo, irás más rápido;
si caminas acompañado, llegarás más lejos.
Proverbio chino

Gracias

A Diego, por su paciencia y su apoyo incondicional.
A mis padres, porque a vosotros os lo debo todo.
A Ramón, Almu y al recién llegado Mateo.
A mis amigos, por esos momentos de desconexión tan necesarios.

A mis directores de tesis, Juan y Germán,
por todo el tiempo y el trabajo invertidos a lo largo de estos años.
A Philippe y a su familia, por su buena acogida en Toulouse.

A mis compañeros del departamento, que han sido en algún momento
también compañeros de cafés, de comidas de tupper o de viajes de
congresos (tanto los de Coruña como los de Santiago) y, en especial, a Ana,
con la que he compartido nuestros primeros pasos en la docencia.

A todos los que, de una forma u otra, han
contribuido a la culminación de este proyecto.

La doctoranda ha sido beneficiaria de una beca FPI (BES-2012-054691), asociada al proyecto MTM2011-22392 del Ministerio de Economía y Competitividad, y de una ayuda a la movilidad predoctoral para la realización de estancias breves (EEBB-I-15-10246) del mismo organismo. Ha sido también parcialmente financiada por los proyectos MTM2014-52876-R y CN2012/130 del Ministerio de Economía y Competitividad y Xunta de Galicia, respectivamente. Los datos de temperatura empleados en esta tesis han sido proporcionados por AEMET.

Abstract

Analysis of the electricity demand and price is presented, within the Spanish Electricity Market, applying statistical tools from the field of functional data. It begins with a descriptive analysis of the electrical data, studying its particular features. This kind of data conform a functional time series. Functional outlier detection methods are proposed to deal specifically with functional time series, taking dependence in this data structure into account. Then, a comparative study among different prediction techniques for next-day electricity demand and price is performed. It includes naïve procedures, time series ARIMA models and robust functional principal components analysis. The use of functional regression methods is proposed in this field. Specifically, the functional nonparametric regression model is used together with the semi-functional partial linear regression model, which allows incorporating external covariates as temperature and wind power production. Bootstrap procedures are proposed to build confidence intervals for the considered functional regression models. Validity of these bootstrap procedures is proved theoretically and they are applied to both a simulation study and the electricity demand and price data. Finally, bootstrap procedures are proposed to build prediction intervals and prediction density, which are also applied to the electrical data.

Resumen

Se presenta un análisis de la demanda y el precio de la electricidad, dentro del Mercado Eléctrico Español, aplicando técnicas estadísticas del ámbito de los datos funcionales. En primer lugar, se realiza un análisis descriptivo de los datos eléctricos, en el que se estudian sus principales características. Este tipo de datos conforman una serie de tiempo funcional. Se proponen métodos de detección de atípicos diseñados específicamente para series de tiempo funcionales, teniendo en cuenta la dependencia presente en esta estructura de datos. A continuación, se realiza un estudio comparativo de diferentes técnicas para la predicción de la demanda y precio de la electricidad al día siguiente. Este estudio incluye métodos naïve, modelos ARIMA de series de tiempo y métodos basados en componentes principales funcionales robustas. Se propone el uso de métodos de regresión funcional en este ámbito. En concreto, se utiliza el modelo de regresión funcional no paramétrico y el modelo semi-funcional parcialmente lineal, en el que se incorporan covariables externas como la temperatura y la producción de energía eólica. Considerando los métodos de regresión funcional indicados, se proponen procedimientos bootstrap para el cálculo de intervalos de confianza, cuya validez se prueba teóricamente y se aplican en un estudio de simulación y en los datos eléctricos de demanda y precio. Finalmente, se proponen procedimientos bootstrap para construir intervalos y densidades de predicción, los cuales se aplican al mismo conjunto de datos eléctricos.

Resumo

Preséntase unha análise da demanda e prezo da electricidade, dentro do Mercado Eléctrico Español, aplicando técnicas do ámbito dos datos funcionais. En primeiro lugar, realízase unha análise descritiva dos datos eléctricos, estudando as súas principais características. Este tipo de datos conforman unha serie de tempo funcional. Propóñense métodos de detección de atípicos deseñados especificamente para series de tempo funcionais, tendo en conta a dependencia presente nesa estrutura de datos. A continuación, lévase a cabo un estudo comparativo de diferentes técnicas para predición da demanda e prezo da electricidade no día seguinte. Este estudo inclúe métodos naïve, modelos ARIMA de series de tempo e métodos baseados en compoñentes principais funcionais robustas. Proponse o uso de métodos de regresión funcional neste ámbito. En concreto, utilízase o modelo de regresión funcional non paramétrico e o modelo semi-funcional parcialmente lineal, no que se incorporan covariables externas como a temperatura e a produción de enerxía eólica. Considerando os métodos de regresión funcional indicados, propóñense procedementos bootstrap para o cálculo de intervalos de confianza, nos que a súa validez se proba na teoría e que son aplicados tanto nun estudo de simulación como nos datos eléctricos de demanda e prezo. Finalmente, propóñense procedementos bootstrap para construír intervalos e densidades de predición, que se aplican ao mesmo conxunto de datos eléctricos.

Contents

Preface

The present memory pretends to summarize all the study developed along the Phd trajectory. Mainly, it is focused on the study of prediction for electricity demand and price data from the Spanish Electricity Market, from a statistical point of view. Specifically, the use of techniques from Functional Data Analysis is introduced in this problem.

Since electricity power is a non-storable product, it is of main importance to anticipate decisions and to prevent misleading measures in order to optimize benefits and to reduce expenses for the agents involved in it. Electricity data, both demand and price, have some particular features in their behaviour that make them hard to analyse. This kind of data has been studied from many points of view, most of them in engineering.

Some statistical tools and procedures are analysed and proposed along this memory, to study this kind of data: electricity demand and price, which is the main objective of the thesis. Briefly, after an introduction to the Spanish Electricity Market, to functional data and a descriptive analysis of our dataset, the study will start from outlier detection methods in the context of functional time series and to forecasting within this field. It keeps going proposing bootstrap procedures to build confidence and prediction intervals for two functional regression models: Functional Nonparametric regression (FNP) and Semi-Functional Partial Linear regression model (SFPL). Asymptotic theorems to prove the validity of the bootstrap confidence intervals in both models are proved in the theoretical part of the memory. Confidence intervals are analysed with simulated data and applied to the electrical dataset, while the case of prediction ones is focussed on the application to real data.

The contents of the memory are organized as follows: Chapter 1 is devoted to introduce the reader to the context in which the study is carried out: electricity data. It begins with an explanation of the Spanish Electricity Market operation in Section 1.2. The transactions that take place in this

market share some properties with a stock market, but the timing and agents involved in it are peculiar. Since this study is based on functional techniques, an introduction to this topic, including some definitions and a small review of the basic concepts is given in Section 1.3. Coming up next, Section 1.4 includes a detailed descriptive analysis of the electricity data used in this study: electricity demand and price in Spain along an entire year. This section dissects each feature of this data, analysing the different performance of demand and price, depending on the day of the week, the weekend, the holidays, the month and season of the year, etc. It is know that external factors affect the demand and the price and so, some additional data will be included in our study. This is the case of the temperature or the wind power production, which affect demand and price, respectively. This additional data will be described in Section 1.5.

Chapter 2 includes an extensive study of functional outliers in the context of this project: the functional time series. Outlier detection can be seen as a first step when dealing with a real dataset, prior to any kind of statistical analysis, as prediction. It begins with a detailed review, in Section 2.2, of some tools in the statistical literature devoted to outlier detection in functional data. Up to our knowledge, there is no specific procedure to detect outliers in the context of functional time series. Three new tools are proposed, based on the procedures included in the review, that are adapted to this context by taking dependence into account, among other variations. They are structured in two sections: Section 2.3 presents the first proposal, which works with functional depths and hypothesis tests. This procedure is based on the method developed in Febrero et al. (2008) and, after its statement, it is analysed and compared with other available methods in a simulation study.

After that, two other new proposals to detect outliers in functional time series are given in Section 2.4. Two different procedures are presented in this section, being both of them based on Functional Principal Component Analysis (FPCA). Specifically, the robust FPCA proposed in Hyndman and Ullah (2007) is applied. Again, a simulation study is performed in order to show the behaviour of our proposals and to compare it with other methods in the literature, showing the improvement of taking dependence of the functional time series into account. Finally, an application of these tools together to detect outliers in electricity demand and price is given in Section 2.5.

Probably the biggest issue when dealing with electrical data is to obtain accurate predictions. So, once outlier detection methods are applied, the

problem of electricity demand and price forecasting is introduced. Chapter 3 includes a comparative study of different techniques focussed in this problem. Not only functional data methods are considered, but also other statistical tools that are very popular in this field, as naïve prediction or time series. From the functional data point of view, the proposal is to consider functional regression to deal with electricity data forecasting, comparing it also with robust FPCA. Specifically, two models are considered: FNP regression and also SFPL regression model, in which it is possible to introduce additional covariates as temperature and wind power production. A comparative study, applied to electricity demand and price prediction concludes this chapter. This analysis extends and complements the methods and results developed in Vilar et al. (2012).

Chapters 4 and 5 contain the theoretical part of this memory and, probably, the main contribution of the thesis to the functional data literature. Both are devoted to propose bootstrap procedures for FNP and SFPL regression models, respectively. These two chapters follow the same structure: they begin with an introduction to the models and their estimators (that have been already used for prediction in Chapter 3) and establish two bootstrap procedures, Naïve bootstrap for homoscedastic models and Wild bootstrap for heteroscedasticity. The central part of the chapters includes the theorems that establish the validity of the bootstrap, together with the assumptions needed for them and also their detailed proofs. Finally, the bootstrap procedures are applied to build confidence intervals in these functional regression models, through both a simulation study and an application to electricity data.

Validity of the bootstrap procedures applied to the functional nonparametric regression model has been already studied in a context of independent data in Ferraty et al. (2010). In this study, this result is extended to the case of dependent data, taking use of the asymptotic distribution of the estimator given by Delsol (2009). However, when dealing with the SFPLR model, there is no preceding, nor in the context of functional data neither in classical regression under random design. Thus, this contribution is, up to our knowledge, the first approach to the proposed bootstrap procedures in this kind of partial linear regression considering both linear and nonparametric components of the model. This study is focussed on the context of dependent functional data, but it can be applied to independent functional data as a particular case.

Last chapter, Chapter 6, includes an application of the same bootstrap procedures, together for FNP and SFPL regression models, to build prediction intervals and also prediction density. Again, the accuracy of the prediction intervals is shown with an application to electricity data, following the same structure as in Chapters 4 and 5. In this case, this chapter is of practical usefulness.

The remaining part of the memory includes some conclusions and open problems for future work, some appendix with auxiliary results, an extended summary of the memory in Spanish and the bibliographic references employed along the study.

As it can be extracted from the structure of the memory, it follows a logical time-line through this entire statistical project. It begins with an introduction to the context in which it is developed, analysing in detail the dataset used along the whole memory. One of the main objectives of the study is to provide accurate predictions for the electricity demand and price, and to go deeper into these predictions by building bootstrap confidence and prediction intervals. For that purpose, the first step is to obtain a manageable dataset and this is done by detecting the functional outliers, taking into account that our dataset corresponds to a functional time series. Then, predictions can be computed using the proposed functional regression models and comparing them with other different tools available in the literature. After that, bootstrap procedures are proposed for both functional regression models, their validity is proved theoretically and they are applied to build both confidence and prediction intervals, concluding the aims and scopes of the project.

Chapter 1

Spanish Electricity Market Data

1.1 Introduction

First chapter is devoted to introduce the area of application of the thesis. All the procedures and methodologies presented in this study were applied to, or developed to work with, electricity data: demand and price of electrical power in the Spanish Electricity Market. Along this study, this data is used to detect outliers, to obtain predictions and to build confidence and prediction intervals. It is then necessary to analyse in detail this data, as a first step in this study, before going through the statistical methodology present in the following chapters.

Chiefly, the main features of the Spanish Electricity Market are given in Section 1.2 in order to provide a general overview of this market behaviour. Section 1.3 serves to locate our work, within the statistics, in the field of functional data. It contains a brief definition of the kind of data to be used. Finally, the main part of this chapter is the descriptive analysis of the electrical data used along the study, which is given in Section 1.4. Also some additional data is described in Section 1.5.

1.2 Overview of the Spanish Electricity Market

The Spanish Electricity Market is a compound of all the markets involved in the trade of electrical power in Spain since 1997, year of the deregulation of the market. Until 1997, the Spanish Government managed the electrical system paying the costs to a set of private electrical companies. The exception was the public company Endesa. Liberalization of this market was made in order to introduce competence and increase the efficiency of the electrical

sector. It was a European policy, pretending to avoid the abuse of the energy monopoly, and Spain was the first continental country in implementing this measure.

Since 1997, the electricity market was divided in 4 components: generation, transport, distribution and merchandising. At the same time, these four components are grouped in two classes of activities: partially liberalized activities (generation and merchandising) and regulated activities (transport and distribution).

Generation and merchandising of the electrical power, which are the partially liberalized activities, can be made by any agent. Meanwhile, transport and distribution is subject to a specific supervision. Also part of the generation is regulated. This is the case of renewable energies, among others, which are subsidized.

Spanish Electricity Market is then referred to as the liberalized activities, including wholesale and retail sector. This study will focus on the wholesale sector, which is called "Mercado Ibérico de la Elecricidad" (MIBEL) meaning Iberian Electricity Market. MIBEL integrates two main operators involved in the transactions of the market. On the one hand, the market operator is devoted to the economic management. It is called "Operador del Mercado Ibérico Español" (OMIE) meaning Spanish-Iberian Market Operator. On the other hand, the system operator is responsible of the technical management and is made by "Red Eléctrica de España" (REE).

There are different kinds of organized markets involved in MIBEL, which are usually called "pool" and which works somehow like a stock market, adjusting electricity supply and demand. The most important one is the daily market, in which the energy for the next-day 24 hours is negotiated and which marks the "market price". This daily market is controlled by OMIE, whose function is the market management and to ensure transparency, objectivity and independence in the recruitments.

Daily market behaves as follows: buyers and sellers present their electricity sale offers one day before, for each one of the next 24 hours. That is, there are 24 different components in the market. Each sale offer is composed by a pair of a quantity of power (measured in Mega Watts per hour, MWh) and its price (measured in €/MWh). Market operator matches the sale offers in a matching, crossing the aggregated curves for sell and purchase and looking at the coincidence point. Then, the prevision for the electricity demand, to-

gether with its price, for the 24 hours of the next day is obtained. The price is obtained by simple adjusting of the offers, which is called uniform-price auction, and that means that all the buyer agents whose offers are greater or equal than the fixed price will pay this fixed price. Meanwhile, seller agents whose offers are lower or equal than the fixed price will also be remunerated based on that fixed price, independently of their initial offer.

After the daily market, the intra-day market is introduced. Its main role is to allow the agents to adjust the offers fixed in the daily market in some intra-day markets, by selling and purchasing electrical power. Once the total demand and price offers are fixed, they are communicated to the System Operator, which is responsible of its execution. As the daily market does not take into account technical restrictions, it may be modified by the System Operator in order to ensure a good operation, taking care of the capability limits on the network.

This memory deals with the electricity demand and price given by the daily market. Specifically, it considers the "Total energy purchase for the Spanish system" and "Marginal price for the Spanish system", also called "spot price" that is obtained as described above. Data source was the OMIE web page (http://www.omie.es), in which they recorded the transactions of the daily market along the time. Although Section 1.4 includes an extensive descriptive analysis of the dataset involved in this study, some of the main features of the electricity data in Spain will be also included here.

On the one hand, demand is affected mainly by working patterns and meteorological conditions, while spot price has an important strategic component, as it depends on the way the agents build their offers. It is then important to take into account this kind of influences with the aim of better understanding its behaviour.

Main features of electricity demand, which are going to be extensively analysed in Subsection 1.4.1, can be summarized in the daily and weekly seasonality, the calendar effect on the weekend, the presence of outliers and the weather influence. Human patterns clearly affect the energy consumption. During the weekdays, from Monday to Friday, most of the people is working, the business and industry are at full capacity and they follow more o less the same routine, day after day. This makes energy consumption to repeat the same, or similar, pattern for each weekday. Electricity demand follows also this shape: during the night the activity decreases and so does the demand, in the mornings electricity demand rises as people start to ac-

tivate and to consume energy. Along the day there are some hours at which demand fluctuates, but keeping always medium-high values. This continues until the night, when the demand decreases again. The repetition of this pattern originates the daily seasonality.

Weekend follows a particular behaviour, different than weekdays. On Saturdays most of the business and industrial activities work half-heartedly and so, demand behaves at a lower level than on weekdays. This is even more remarkable on Sundays, as this is the traditional holiday for workers, both business and commerce or industries. On Sundays demand reaches its lowest values, with also a smoother pattern. Again, as this behaviour is repeated in a similar way, week after week, it originates the weekly seasonality, due to the calendar effect on the weekend.

Finally, it is important to note the presence of outliers in the demand, most of them due to holidays, strikes or special days for other reasons. They induce non-usual patterns in the weekly routine, affecting also the electricity demand. Furthermore, other external factors affect the demand, being probably the most remarkable ones the weather variables and, among them, the temperature. Traditionally, when the temperature is low or high, people take use of climatization systems that need energy, increasing the electricity demand. On the contrary, medium temperature does not affect demand directly. This relation will be studied in detail in Subsection 1.5.1.

Electricity price is traditionally more complex than demand, as it is affected by economic rules besides the usual "supply and demand" market rules. Energy source is also a parameter in the price fixing. In Spain, renewable energies as subsidised by the Government, allowing them to enter on the daily market in a preferential position with zero price. In this situation, the electricity demand is first covered by those subsidised energies (wind power, hydroelectric or nuclear) and then, the rest of the remaining demand is completed with the other energy sources which really mark the spot price. The main importance of this fact is that price usually decreases when the input of renewable energies increases, resulting in unexpected prices. This is one of the key points when dealing with electricity price predictions that will also be analysed in detail in Subsection 1.5.2.

Usually along the year, one can find some time intervals in which wind power production covers an out of the ordinary amount of demand. As a consequence, the spot price reduces reaching even zero prices during this time, that can last since some hours to some days. Wind power production is an

increasing energy source in Spain. Next extract from a specialized article in the journal The Guardian (from January 2014) gives an idea of the importance of wind power in Spain:

Wind power was Spain's top source of electricity in 2013.
(...) Red Eléctrica de España (REE) released a preliminary report on the country's power system late last month, revealing that for "the first time ever, [wind power] contributed most to the annual electricity demand coverage". (...) wind turbines met 21.1% of electricity demand on the Spanish peninsular, narrowly beating the region's fleet of nuclear reactors, which provided 21% of power.
In total, wind farms are estimated to have generated 53,926 gigawatt hours of electricity, up 12% on 2012.(...) "Throughout 2013, the all-time highs of wind power production were exceeded" (...) in January, February, March and November wind power generation was the technology that made the largest contribution towards the total energy production of the system.(...)

1.3 Functional data

Functional data is a relative recent field in Statistics, which has been increasing its presence over the last years. It is based on extending the dimension of the data which statistics works with, allowing to deal directly with curves, images, etc. The first monograph about this topic was published within the last two decades. Specifically in 1997 by Ramsay and Silverman (book entitled Functional Data Analysis), with a second edition in 2005. This gives an idea of the novelty of functional data. Other main monographs are the book by Ferraty and Vieu (2006), in which they analyse functional data from the nonparametric statistical point of view, Horvath and Kokoszka (2012), focussed on inference, and, more recently, Hsing and Eubank (2015), which is a compendium of the key mathematical concepts and results that are relevant for the theoretical development of functional data analysis.

In spite of its novelty, several authors had contributed over these years to the growth of the research in functional data and to spread it to a large number of areas. Most statistical techniques have been generalized to the functional context. This includes linear regression models (Cardot, Ferraty and Sarda, 1999; Li and Hsing, 2007; García-Portugués, González-Manteiga and Febrero-Bande, 2014), nonparametric smoothing methods (Ferraty and Vieu, 2002; Delsol, Ferraty and Vieu, 2011; Shang, 2014), classification (Cuevas, Febrero and Fraiman, 2007; Baíllo, Cuesta-Albertos and Cuevas,

2011; Sguera, Galeano and Lillo, 2014), dimension reduction (Boente and Fraiman, 2000; Hall, Müller and Wang, 2006) and bootstrap methods (González-Manteiga and Martínez-Calvo, 2011; Ferraty, van Keilegom and Vieu, 2012). In addition, FDA has been successfully applied in a wide range of fields such as climatology (Besse, Cardot and Stephenson, 2000), chemometrics (Ferraty and Vieu, 2002), environmetrics (Aneiros-Pérez et al., 2004), demography (Hyndman and Ullah, 2007), social sciences (Ocaña, Aguilera and Escabias, 2007) and the electricity market (Aneiros et al., 2013 and 2016). See Cuevas (2014) for an overview.

In the following paragraphs, some basic concepts in FDA are introduced in order to provide a general overview of this field.

1.3.1 Definition of functional data

In general, a functional datum can be defined as an observation from a random variable taking values in an infinite dimensional space (see Ferraty and Vieu, 2006). This definition can include, for example, curves, surfaces or images.

This memory is going to focus on curves coming from Spanish Electricity Market: daily demand and price curves.

In this setting, a functional datum (a curve) will be denoted by

$$\chi = \{\chi(t) : t \in T\}$$

where T is an interval, $T \subset \mathbb{R}$.

1.3.2 Semi-metrics

Measures for proximity between mathematical objects are, as a rule, a main issue into many statistical methodologies. Classical norms can be used to measure how close two objects are.

Within a finite dimensional space, one may assume equivalence between all norms, which means that the election of the norm is a minor question. Meanwhile, when dealing with infinite–dimensional spaces, this is not the case. As equivalence between norms cannot be assumed, their election becomes crucial and some features of the data (as the shape, if they are smooth of rough) may help this choice.

In the following, semi-metrics will be considered as a closeness measure between functional data (as metrics may be too restrictive in some situations, revealing less structure in the data, and semi-metric spaces are better adapted than metric spaces).

Some of the semi-metrics used along this memory are based on derivatives or Principal Component Analysis. For example, when dealing with smooth curves, one can consider the semi-metric based on the ν–th derivative of the curve, $d_\nu^{deriv}(\cdot,\cdot)$, where

$$d_\nu^{deriv}(\boldsymbol{\chi}_i,\boldsymbol{\chi}_j)=\sqrt{\int(\boldsymbol{\chi}_i^{(\nu)}(t)-\boldsymbol{\chi}_j^{(\nu)}(t))^2dt}.$$

However, if one deals with rough but balanced functional data, one can use the semi-metric based on the projection on the first s eigenvectors, $v_1(\cdot),\ldots,v_s(\cdot)$, associated with the s largest eigenvalues of the empirical covariance operator of the functional predictor $\boldsymbol{\chi}$, which is $d_s^{proj}(\cdot,\cdot)$:

$$d_s^{proj}(\boldsymbol{\chi}_i,\boldsymbol{\chi}_j)=\sqrt{\sum_{k=1}^{s}\left(\int(\boldsymbol{\chi}_i(t)-\boldsymbol{\chi}_j(t))v_k(t)dt\right)^2}.$$

Another kind of semi-metrics could be considered, as the semi-metric based on Partial Least Squares.

1.3.3 Centrality measures: Functional depths

A first step into a descriptive analysis of a dataset could be to obtain the classical measures of centrality: mean, median and mode.

The functional mean for a set of curves $S=\{\boldsymbol{\chi}_1\ldots,\boldsymbol{\chi}_n\}$ can be easily computed as follows:

$$\boldsymbol{\chi}_{mean,S}(t)=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\chi}_i(t),\ \forall t\in T.$$

However, if one wants to estimate the functional median or mode, in which notions of distances and densities appear, the problem becomes more difficult and it will make use of functional depths.

The concept of depth was first introduced in the context of multivariate data (as Tukey or halfspace depth, simplicial depth or location depth) and

then, it was extended to the infinite dimensional case of functional data.

Depths of functional data were introduced to measure how deep (central) or outlying an observation is, with respect to a functional dataset. This allows to order a set of functional data from the centre-outward, so that the most central or interior data in the sample will have higher depth. Thus, functional depths may indicate which observations (if any) can be considered as outliers: those with unusually low depths.

Along this memory, three different functional depths are used, which are briefly described in the next list:

- h-modal depth (MD): proposed by Cuevas, Febrero and Fraiman (2006). Based on the concept of mode, these authors defined a functional mode as the curve most densely surrounded by the rest of the curves, which corresponds to the maximum value of this depth. The h-modal depth of a curve χ, regarding a set of curves $\chi_1, \ldots, \chi_n$ is given by:

$$MD_n(\chi, h) = \sum_{i=1}^{n} K\left(\frac{\|\chi - \chi_i\|}{h}\right),$$

  where $\| \ \|$ is a norm in the functional space, K is a kernel function and h is a bandwidth.

- Band depth (BD): This depth is a graph based approach proposed by López-Pintado and Romo (2009). In summary, it counts the number of times that one curve is contained in a band, built from the rest of the curves in the sample. This band can be constructed using two or more curves. In the case when two curves are employed to build the band, the following expression indicates the proportion of bands, $B(\chi_{i_1}, \chi_{i_2})$, determined by two different curves, χ_{i_1} and χ_{i_2}, containing the graph of χ, $G(\chi)$, regarding a set of curves $\chi_1, \ldots, \chi_n$:

$$BD_n^{(2)}(\chi) = \binom{n}{2}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq n} 1_{\{G(\chi) \subseteq B(\chi_{i_1}, \chi_{i_2})\}}. \tag{1.1}$$

  Finally, the band depth for the curve χ is given by the aggregation of the latter expression when the number of curves employed to build the band varies:

$$BD_n(\chi) = \sum_{j=1}^{J} BD_n^{(j)}(\chi), \ J \geq 2.$$

- Modified band depth (MBD): This is a more flexible version of the Band depth, also proposed in López-Pintado and Romo (2009), which enables to change the indicator function in expression (1.1) to the "proportion of time" that the curve is in the band. This allows the result to take intermediate values between 0 and 1, giving a more flexible approximation. That is, the set of points, in the interval T where the curves take values, where the curve χ is contained in the band built from two curves is:

$$A_2(\chi) = \left\{ t \in T : \min_{r \in \{i_1, i_2\}} \chi_r(t) \leq \chi(t) \leq \max_{r \in \{i_1, i_2\}} \chi_r(t) \right\},$$

and considering the Lebesgue measure in T, $\lambda_r(A_2(\chi)) = \dfrac{\lambda(A_2(\chi))}{\lambda(T)}$, one has the analogous expression as (1.1):

$$MBD_n^{(2)}(\chi) = \binom{n}{2}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq n} \lambda_r(A_2(\chi)).$$

Finally, the MBD for the curve χ is given by:

$$MBD_n(\chi) = \sum_{j=1}^{J} MBD_n^{(j)}(\chi), \ J \geq 2.$$

Once the concepts of functional depth have been introduced, one may define the functional median as the deepest curve in a dataset, with respect to a specific depth. The functional mode will be considered as the deepest curve when one considers the modal depth defined above.

1.3.4 Functional Principal Component Analysis

Classical Principal Component Analysis (PCA) is one of the most used techniques to reduce dimension of multivariate data. This tool can be extended to the context of functional data, resulting the Functional Principal Components Analysis (FPCA). Along this memory, FPCA is applied to the detection of functional outliers in Chapter 2 and also for prediction in Chapter 3. Next paragraphs present a brief review on this topic, from the finite-dimensional case to the functional space L_2. Without loss of generality, it is assumed that the considered (multivariate or functional) random variable has zero mean.

PCA is a standard approach to explore variability in multivariate data, $\mathbf{X} \in \mathbb{R}^d$. This approach specifies the d directions, $\{\mathbf{v}_k\}_{k=1}^d \in \mathbb{R}^d$, that maximize the variance along each component, subject to the orthonormality condition. That is, the aim is to find the vectors $\mathbf{v}_k$ such that the variance of

$$\alpha_k = \mathbf{v}_k^T \mathbf{X}$$

is maximized subject to

$$\mathbf{v}_k^T \mathbf{v}_k = 1 \ (k = 1, \ldots, d) \text{ and } \mathbf{v}_k^T \mathbf{v}_j = 0 \ (k \neq j).$$

Vectors $\mathbf{v}_k$ can be obtained by solving the eigenfunction

$$\mathbf{B}\mathbf{v} = \lambda \mathbf{v}, \tag{1.2}$$

where $\mathbf{B}$ denotes the covariance matrix of $\mathbf{X}$, and $\mathbf{v} \in \mathbb{R}^d$. The eigenvector $\mathbf{v}_k$ is known as the k-th principal component, assuming that the corresponding eigenvalue λ_k satisfies $\lambda_k \geq \lambda_{k+1}$ $(k = 1, \ldots, d-1)$. Thus, the direction $\mathbf{v}_k$ corresponds to the k-th most important mode of variation, while

$$\frac{\lambda_k}{\sum_{k=1}^d \lambda_k}$$

is the proportion of total variance explained by $\mathbf{v}_k$. PCA is often the first step in reducing the dimension of the data, while maintaining most of the information. This is done by means of the approximation

$$\mathbf{X} \approx \sum_{k=1}^{d'} \alpha_k \mathbf{v}_k, \tag{1.3}$$

where $d' < d$ and $\sum_{k=1}^{d'} \lambda_k$ is close to $\sum_{k=1}^{d} \lambda_k$ (note that $\sum_{k=1}^{d'} \lambda_k / \sum_{k=1}^{d} \lambda_k$ is the proportion of variability of $\mathbf{X}$ explained by the representation (1.3)). Of course, in practice $\mathbf{B}$ is unknown and must be estimated. For details, see e.g. Johnson and Wichern (2002).

Reducing the dimension is especially important when data are infinite dimensional, this being the case of functional data. If we focus on curves observed in $T = [a, b]$ $(-\infty < a < b < \infty)$ and square integrable, then, if χ denotes a functional random variable, PCA can be easily generalized to FPCA. The aim is to find the functions $\phi_k : [a, b] \to \mathbb{R}$ such that the variance of

$$\beta_k = \int_a^b \phi_k(t)\chi(t)dt \tag{1.4}$$

is maximized subject to the constraints

$$\int_a^b \phi_k^2(t)dt = 1 \text{ and } \int_a^b \phi_k(t)\phi_j(t)dt = 0 \ (k \neq j). \tag{1.5}$$

As in the finite-dimensional case, the functional principal components, $\phi_k(\cdot)$, can also be defined as the orthonormal functions verifying

$$\int_a^b \mathbf{C}(t,s)\phi_k(s)ds = \lambda_k\phi_k(t) \ (t \in [a,b], \ k = 1,2,\dots), \tag{1.6}$$

where $\mathbf{C}(t,s)$ denotes the covariance between $\chi(t)$ and $\chi(s)$. Finally, dimension reduction is performed by considering the approximation

$$\chi(\cdot) \approx \sum_{k=1}^{K} \beta_k\phi_k(\cdot), \tag{1.7}$$

where $K < \infty$ and $\sum_{k=1}^{K}\lambda_k$ is close to $\sum_{k=1}^{\infty}\lambda_k$ (we have assumed that $\lambda_k \geq \lambda_{k+1}$, $k = 1,2,\dots$). For details, see e.g. Ramsay and Silverman (2005).

Functional principal components, $\phi_k(\cdot)$, depend on the unknown covariance operator $\mathbf{C}(\cdot,\cdot)$. Assuming that one has observations $\{\chi_i\}_{i=1}^n$ identically distributed from the functional random variable χ, estimates for $\phi_k(\cdot)$ can be obtained by using

$$\widehat{\mathbf{C}}(t,s) = \frac{1}{n}\sum_{i=1}^{n}(\chi_i(t) - \overline{\chi}(t))(\chi_i(s) - \overline{\chi}(s)), \text{ where } \overline{\chi}(t) = \frac{1}{n}\sum_{i=1}^{n}\chi_i(t),$$

instead of $\mathbf{C}(t,s)$ in (1.6). See Horváth and Kokoszka (2012) for the consistency of $\widehat{\mathbf{C}}$ and the corresponding eigenfunctions and eigenvalues, under either independent curves or weakly dependent functional time series.

It is worth noting that, apart being used for dimension reduction, FPCA can also be used as a tool for outlier detection. Nevertheless, as noted in the previous paragraph, the estimation of functional principal components is based on the estimated covariance operator $\widehat{\mathbf{C}}(\cdot,\cdot)$, which is known to be sensitive to outliers. Thus, if the goal is to construct an approach based on principal components to identify functional outliers, robust FPCA should be considered (see e.g. Hyndman and Ullah 2007; Hyndman and Shang 2010; Sawant, Billor and Shin 2012).

Hyndman and Ullah (2007) propose estimating the functional principal components by means of the functions $\widehat{\phi}_k(\cdot)$ that maximize the variance of the scores

$$z_{i,k} = w_i \int_a^b \phi_k(t)\chi_i(t)dt \tag{1.8}$$

subject to the constraints (1.5). The weights w_i are computed as

$$w_i = \begin{cases} 1 & \text{if} \quad v_i < s + \lambda\sqrt{s} \\ 0 & \text{otherwise} \end{cases}$$

where

$$v_i = \int_a^b (\chi_i(t) - \sum_{k=1}^{K} \widetilde{\beta}_{i,k}\widetilde{\phi}_k(t))^2 dt \tag{1.9}$$

with $\widetilde{\phi}_k(\cdot)$ being initial (highly robust) projection-pursuit estimates of $\phi_k(\cdot)$ obtained from the RAPCA algorithm (see Hubert, Rousseeuw and Verboven 2002) considering equal weights w_i in (1.8), while $\widetilde{\beta}_{i,k} = \int_a^b \widetilde{\phi}_k(t)\chi_i(t)dt$. In addition, s is the median of $\{v_1, \ldots, v_n\}$ and $\lambda > 0$ is a tuning parameter to control the degree of robustness. Once the robust estimates $\widehat{\phi}_k(\cdot)$ are obtained, the coefficients corresponding to the curve χ_i are constructed as

$$\widehat{\beta}_{i,k} = \int_a^b \widehat{\phi}_k(t)\chi_i(t)dt. \tag{1.10}$$

Note that, given the definition of $\{w_i\}$, outlying curves receive low weight (for details, see Hyndman and Ullah 2007). For this reason, the estimates $\widehat{\phi}_k(\cdot)$ are robust.

1.3.5 Functional Regression

Regression is a powerful tool to analyse the relation between variables and to obtain predictions. As many other classical statistical techniques, also regression analysis can be extended to deal with functional data, resulting Functional Regression.

Probably, the most commonly used regression model is the classical linear model where the variables are scalar. If one introduces curves instead of scalar values, one can obtain three different regression models:

- Fully functional model: both response and regressors are curves.
- Scalar response model: scalar response and functional regressors.

- Functional response model: functional response and scalar regressors.

This classification given for the linear regression model can be applied to other kind of models, as partial-linear or nonparametric regression models.

Within this memory, functional regression is one of the main topics. First, in Chapter 3 some of them are applied to predict electricity demand and price in the context of functional autoregression. Later, on Chapters 4 and 5, the FNP and SFPL regression models, respectively, are analysed in detail proposing bootstrap procedures to build confidence intervals, which are also extended in Chapter 6 to build prediction intervals. Thus, a brief introduction to these models is given.

FNP regression model, with scalar response:

$$Y_i = m(\boldsymbol{\chi}_i) + \varepsilon_i,$$

where the response, Y, is scalar while the covariate, $\boldsymbol{\chi}$, is valued in some infinite-dimensional space, $\mathcal{H}$, which is endowed with a semi-metric $d(\cdot,\cdot)$. Finally, $m(\cdot)$ is an unknown smooth real-valued operator and the corresponding random errors $\{\varepsilon_i\}$ are i.i.d. as ε.

The regression function $m(\cdot) = \mathbb{E}(Y \mid \boldsymbol{\chi} = \cdot)$ can be estimated by $\widehat{m}_h(\cdot)$; that is,

$$\widehat{m}_h(\chi) = \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi) Y_i.$$

Nadaraya-Watson type weights can be used:

$$w_h(\boldsymbol{\chi}_i, \chi) = \frac{K(d(\boldsymbol{\chi}_i, \chi)/h)}{\sum_{i=1}^{n} K(d(\boldsymbol{\chi}_i, \chi)/h)},$$

where $K(\cdot)$ is a real function (the kernel) and $h > 0$ is a smoothing parameter.

SFPL regression model, also with scalar response:

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

where the response Y is scalar, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown real parameters, m is an unknown smooth real-valued operator and ε_i are i.i.d. mean zero random errors. The explanatory random variables $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ and $\boldsymbol{\chi}_i$ are valued in $\mathbb{R}^p$ and in some infinite-dimensional space, $\mathcal{H}$, respectively. Let us denote

$$\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T,\ \mathbf{Y} = (Y_1, \ldots, Y_n)^T,\ \mathbf{W}_h = (w_h(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)), i, j = 1, \ldots, n$$

and, for any $(n \times q)$ matrix $\mathbf{A}$ $(q \geq 1)$,

$$\widetilde{\mathbf{A}}_h = (\mathbf{I} - \mathbf{W}_h)\mathbf{A}.$$

The following estimators $\widehat{\boldsymbol{\beta}}_h$ and $\widehat{m}_h(\cdot)$ of the vector parameter $\boldsymbol{\beta}$ and the function $m(\cdot)$ will be considered:

$$\widehat{\boldsymbol{\beta}}_h = (\widetilde{\mathbf{X}}_h^T \widetilde{\mathbf{X}}_h)^{-1} \widetilde{\mathbf{X}}_h^T \widetilde{\mathbf{Y}}_h$$

and

$$\widehat{m}_h(\chi) = \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)(Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}_h),$$

respectively. Nadaraya-Watson type weights are also used.

Kernel local weighting

As it could be seen in the functional regression models considered above, when dealing with nonparametrics one may consider kernel local weighting, based on both a kernel function and a smoothing parameter. The idea of this concept in the one dimensional case is to give a weight to each value around a point x, taking into account the difference between each value and that x:

$$\frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

The weights satisfy that bigger difference/distance corresponds to smaller weights.

Focussing on the kernel functions, classical symmetrical kernel functions can be considered in the real case: box kernel, triangle kernel, quadratic kernel or Gaussian kernel. If one extends this kernel local weighting to functional case (also with multivariate data), one may consider asymmetric kernels as the distance between observations will be always a positive value and so, the support of the kernel must be also positive. Note that in this case, the kernel function is applied to the distance (in functional data, the semi-metric) between a curve $\boldsymbol{\chi}_i$ and χ:

$$K\left(\frac{d(\boldsymbol{\chi}_i, \chi)}{h}\right).$$

Thus, one can extend Nadaraya Watson weights to functional data, resulting:

$$\frac{K(d(\boldsymbol{\chi}_i, \chi)/h)}{\sum_{j=1}^{n} K(d(\boldsymbol{\chi}_j, \chi)/h)}.$$

This fact implies the use of the asymmetric box kernel, asymmetric triangle kernel, asymmetric quadratic kernel or asymmetric Gaussian kernel, for instance.

1.3.6 Functional Time Series

When dealing with functional data which is recorded along the time, one must take into account the temporal ordering and temporal dependence present in it. This is the case of the daily curves of electricity demand and price. As each curve corresponds to one day, there is an implicit temporal dependence within each curve. While, as each day is also connected to the previous and following ones, there is another kind of temporal dependence between the curves. This kind of situations are matched up in Functional Time Series (FTS).

A Functional Time Series can be seen as the natural extension from the univariate time series to functional data in which each observation of the time series is, for example, a curve. Specifically, to define the FTS, $\{\boldsymbol{\chi}_i\}_{i=1}^n$, which is going to be used along this study, a real-valued continuous time stochastic process $\{\boldsymbol{\chi}(t)\}_{t\in\mathbb{R}}$ is considered. Then, it is assumed that such process is seasonal with seasonal length τ and it is observed on the interval $[a, b)$ with $b = a + n\tau$. The definition of the functional time series $\{\boldsymbol{\chi}_i\}_{i=1}^n$ in terms of $\{\boldsymbol{\chi}(t)\}_{t\in\mathbb{R}}$ is:

$$\boldsymbol{\chi}_i(t) = \boldsymbol{\chi}(a + (i-1)\tau + t) \text{ with } t \in [0, \tau). \tag{1.11}$$

1.4 Exploratory analysis of the electrical data

Along all this analysis, the methods and procedures developed will be applied and compared using the same functional database. This database consists mainly in electricity demand and price from the Spanish Electricity Market. Also other covariates, regarding weather information and source of power generation will be included in the study.

Data corresponding to electricity demand and price were collected in a database used along all this thesis. Data source was OMIE. Main features of the electrical data will be analysed in detail for the demand and the price. They share some characteristics that can be summarized in the daily and weekly seasonality, the calendar effect on weekend and also the presence of outliers. Both kinds of data present other particularities that may be studied in detail for each case.

The database consists in hourly demand and price in Spain for the years 2011 and 2012. Specifically, it recorded the total energy traded in the daily market (measured in MWh: Mega Watts per hour) and the marginal price in the Spanish system (measured in Cent/kWh). All this data was registered for each hour of the day within the years 2011 and 2012.

The rest of the section is organized as follows. Subsection 1.4.1 contains a descriptive study of the electricity data for the demand and Subsection 1.4.2 for the price. Demand and price will be analysed separately in order to distinguish the main features and to study their behaviour along the selected period. Main patters of the electricity data remain stable year by year and so, only data for 2012 will be employed. In that way, disturbances due to the years will not change the descriptive study as the years 2011 and 2012 were similar in terms of electricity. After this detailed analysis of the electricity data, the additional data employed will be described in Section 1.5, again divided in Subsection 1.5.1, for the temperature data, and Subsection 1.5.2, for the source of the energy.

1.4.1 Electricity Demand

In this subsection, a descriptive analysis for the electricity demand during 2012 in the Spanish Electricity Market will be carried out.

First of all, the historical demand (MWh) for the year 2012 in Spain is represented as a time series in Figure 1.1. Vertical red lines are indicators of the different months in which the year is divided, in order to distinguish the behaviour of the demand along the year. The presence of trend in this data will be taken into account in different parts of this memory, for example when it will be used to detect outliers. However, in this section, the raw data is analysed as it was really recorded in the daily market.

At the first sight, one can see how the demand fluctuates along the year. First and last months of the year corresponds to high values of demand, as in some other central parts of the year. This fact, suggest different patterns depending on the period of the year possibly due to the different climate, among other reasons.

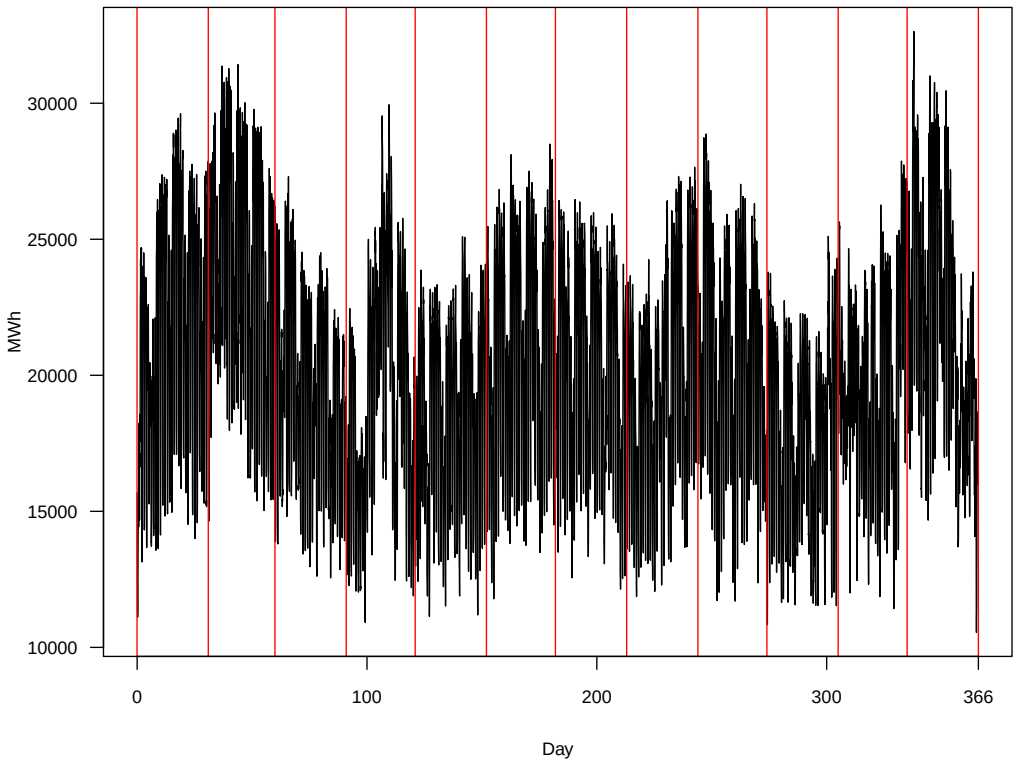


Figure 1.1: Functional time series of the electricity demand in Spain in 2012.

In Figure 1.2, only data form January is represented. One can clearly distinguish how the demand varies in each day and also each week, which is very difficult to see when the whole year is plotted together. In this new graph, as it was said before, only data corresponding to January is plotted, that is, the first 744 hours of 2012.

Daily and weekly seasonality is now explicit. The first day of the year 2012 was Sunday but, from the second day, one can see how the pattern of the week is: the first 5 days of the week are very similar, followed by the weekend that has lower values. Also this pattern can be disturbed in the first complete week of the year due to the presence of a holiday in January the sixth, which is more similar to the weekend that the weekdays. This fact allows to intuit other of the main features of this kind of data: the presence of outliers that change the usual behaviour. In the next three complete weeks plotted in the graph, one can see a more stable period, distinguishing clearly the five weekdays and the weekend.

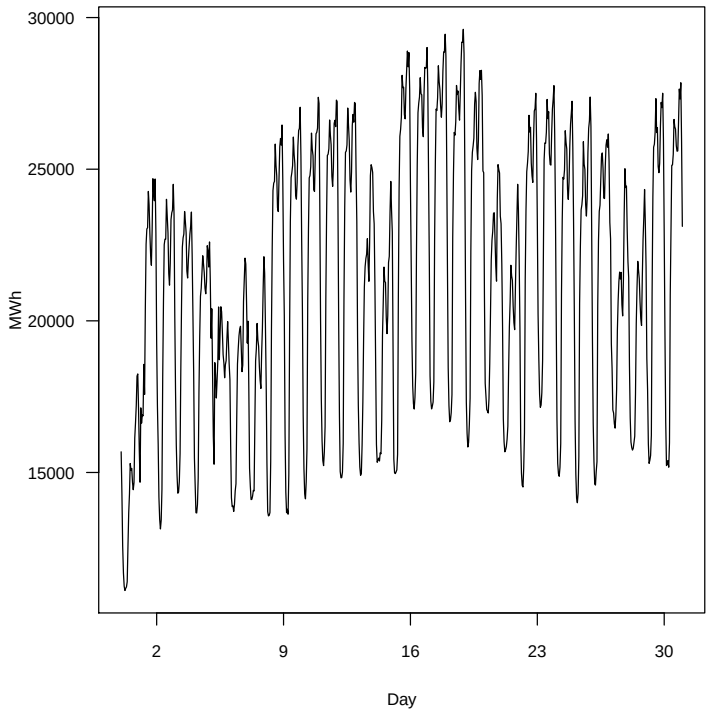


Figure 1.2: Time series of the electricity demand in Spain in January 2012.

This kind of data, recorded as an hourly time series, can be treated as functional data. Each 24 hourly observations for a day will be used to build a curve, resulting 366 daily curves for the year 2012 (remember that 2012 was a leap-year, so it has 366 days instead the common 365 days). In Figure 1.3, one can see all the daily demand curves. Demand swing between 10000 and 30000 MWh depending on the day and it repeats mainly the same shape. Minimum values are recorded in the early morning, around 5 a.m., and maximum ones between 12 a.m. and 10 p.m. It is expected that electricity consumption reduces during the night, corresponding to the period of the day with less activity. Maximum values are recorded during the central part of the day, having also a little slump in the first hours of the afternoon. This behaviour is consistent with the performance of our society.

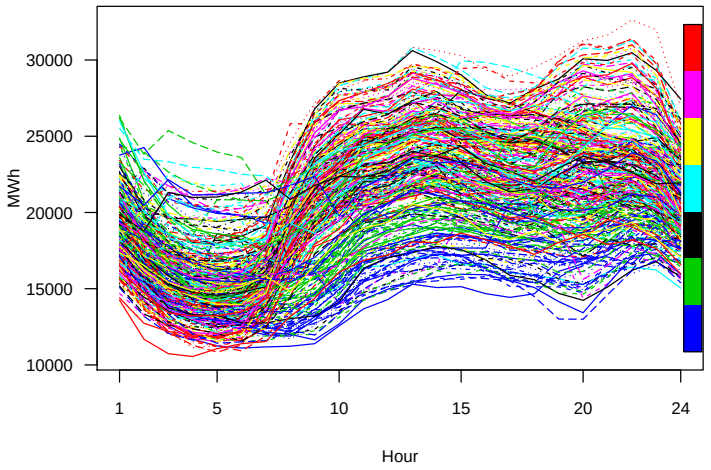


Figure 1.3: Electricity demand daily curves. Colour scale distinguishes the day of the week from Monday (red) to Sunday (blue).

Colour scale in Figure 1.3 is plotted according to the day of the week. Red is for Monday, pink for Tuesday and so on, until green for Saturday and blue for Sunday. In general, lines are intermingled but it can be seen that specially Sundays, in blue line, are plotted together in the bottom of the graph. This gives the idea of studying each day of the week separately, in order to see their behaviour.

Figure 1.4 shows the daily demand curves separately for the weekdays, from Monday to Friday, in different colours, overlaying the curves for all the year in grey. Also the functional mean for each group of days is plotted in black dashed line. Generally, the pattern is repeated along all the weekdays,

occupying the medium and high values of the demand. From this graph one can establish that all the weekdays work in the same way so, they can be treated as an only group.

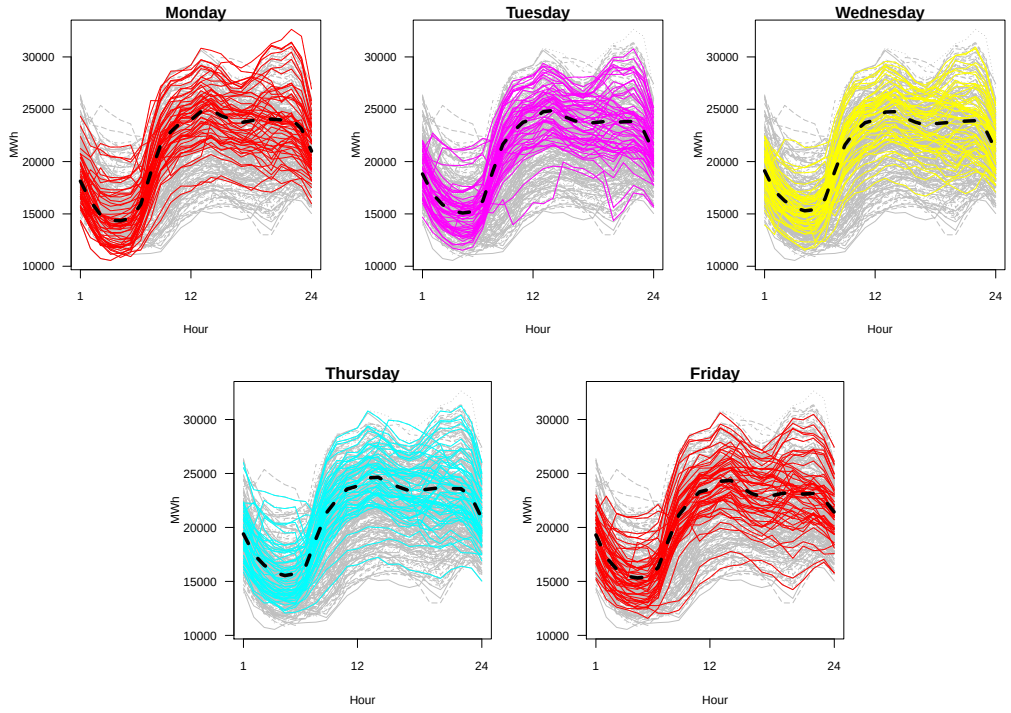


Figure 1.4: Weekdays daily demand curves.

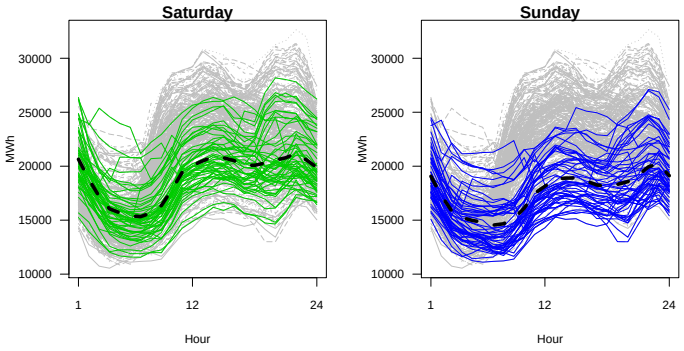


Figure 1.5: Daily electricity demand in the weekend, separately for Saturday and Sunday.

Figure 1.5 is analogous to the last one, but now the weekend is represented separately for Saturdays and Sundays. Comparing it to the weekdays, one can see a clear difference, as the weekend takes almost always low values of the demand.

There is also a difference between Saturday and Sunday, because in this last day the demand curves are generally in the minimum values along the year, while the demand on Saturdays is not so low. This behaviour is reasonable because in the weekend the demand decreases due to the reduction in the industrial production and the change of the workers routine. In weekdays all the industries and economical activities are working at their usual rhythm, in Sundays most of them are close in the day off. Nevertheless, Saturdays are in the middle of this situation, as a big amount of people don't work and schools are closed but, for some economical activities related to shopping, leisure, restaurants, etc. is a very busy day. As a consequence, Saturdays and Sundays will be studied separately.

Due to the analysis developed within the different days of the week, three groups of days will be considered, in which the behaviour of the electricity demand is very similar: weekdays, Saturday and Sunday. These three groups of days are plotted in Figure 1.6.

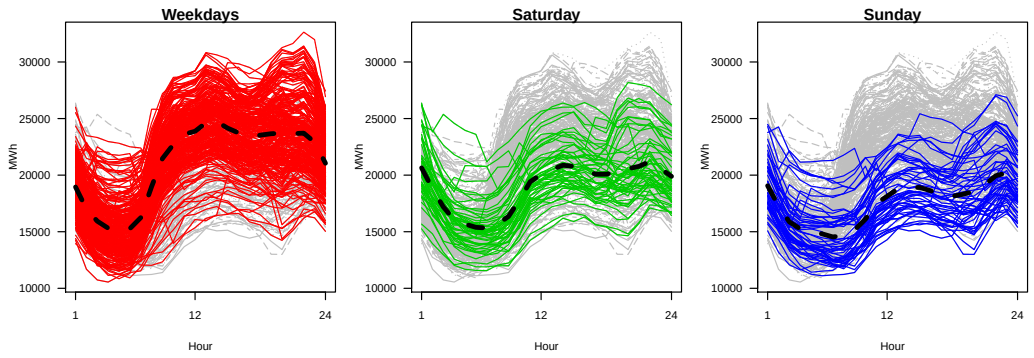


Figure 1.6: From left to right: Weekday, Saturday and Sunday daily curves of electricity demand. Dotted black lines represent the functional mean of each group of days.

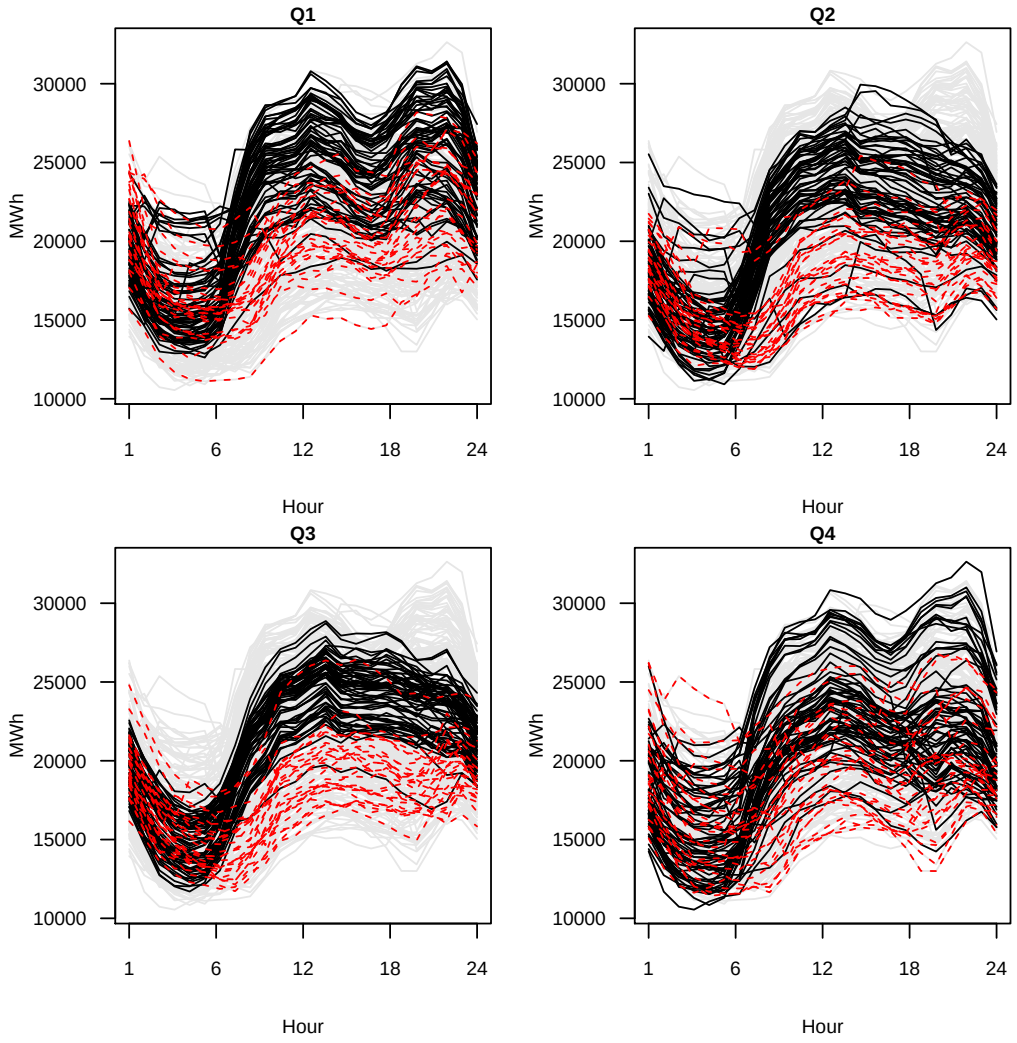


Figure 1.7: Daily curves of electricity demand for each quarter of the year 2012. Dotted red lines correspond to weekend.

Due to the features of the electricity consumption and its relation with the weather and the economic and industrial activities, it is expected that demand varies also looking at the different periods of the year and not only at the day of the week. This was already pointed out in Figure 1.1. In Figure 1.7 each quarter of the year is represented separately and, within each quarter, weekdays are plotted in black and weekend in red. Again, it is easy to see the different behaviour in the demand of a working day and in the weekend. Moreover, during the first and the fourth quarter of the year, when the weather is cold, the demand is generally higher than in the middle of the year.

Last quarter also seems to be more unstable and variable than the rest of the year.

Following this reasoning, Figure 1.8 represents the daily demand curves separately for each month. As a reference, the daily curves for all the year keep on the background in grey colour and, within each month, the weekdays are represented in black and the weekend in red (as was seen before how different is the behaviour between these groups of days).

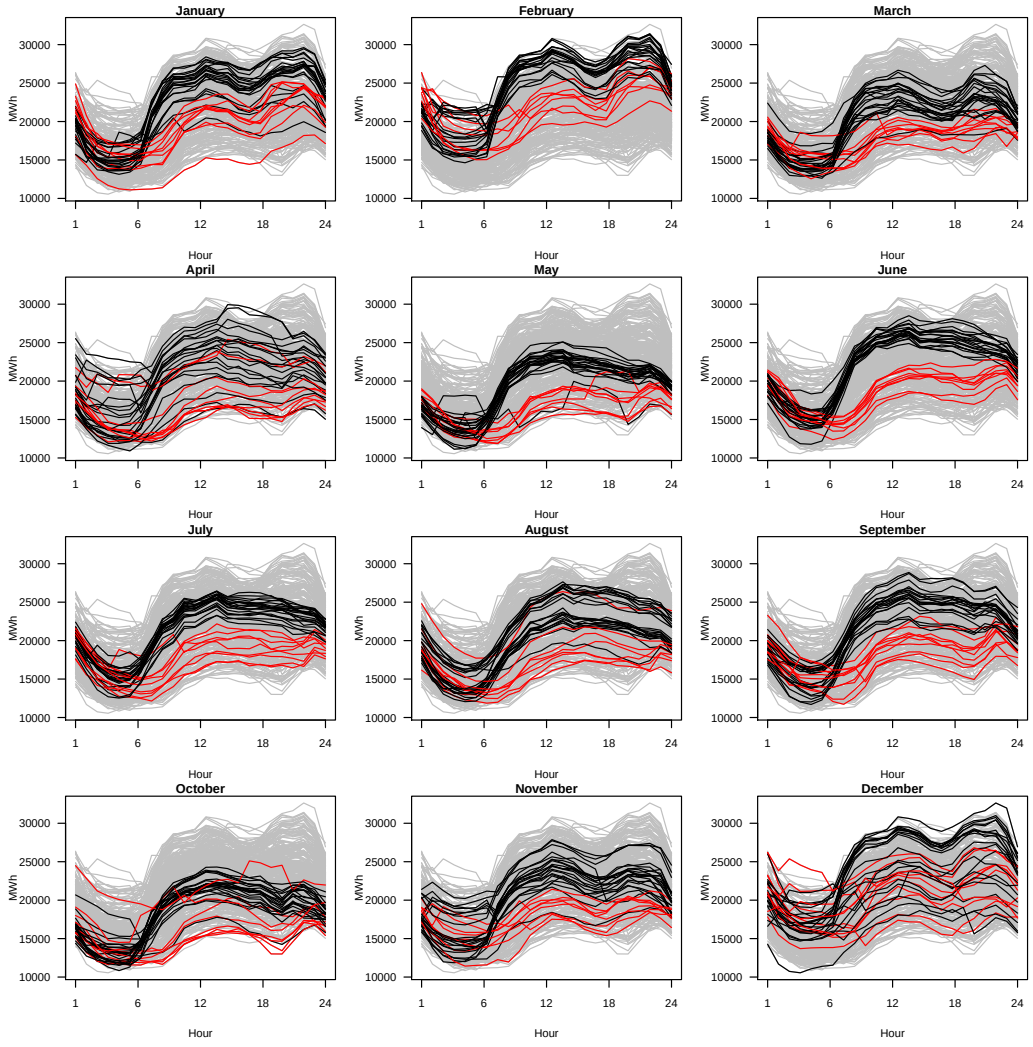


Figure 1.8: Daily curves of electricity demand for each month of the year 2012. Dotted red lines correspond to weekend.

One can distinguish how the daily curves of the demand vary along the year. First months, specially February, reach high values of demand. Second quarter starts in April with a very unstable period, maybe due to the mix of climate conditions, as in April the weather can be typically wintry or spring.

In the central part of the year, the demand is generally lower and also more concentrated, it is more difficult to see large fluctuations. This also makes the differences between weekdays and weekend more remarkable, since the curves corresponding to these groups are almost always separated. Finally, in the last part of the year the demand rises again coming up to December, in which demand is not only high but also very changeable. Taking into account that in December there are a lot of holidays in Spain, it comes out that these variations can be due to these special days. It is expected than in a holiday the demand behaves as in a Sunday, in which most of the business are close, and this is probably the reason why some of the weekdays curves in December are mixed with the weekend ones.

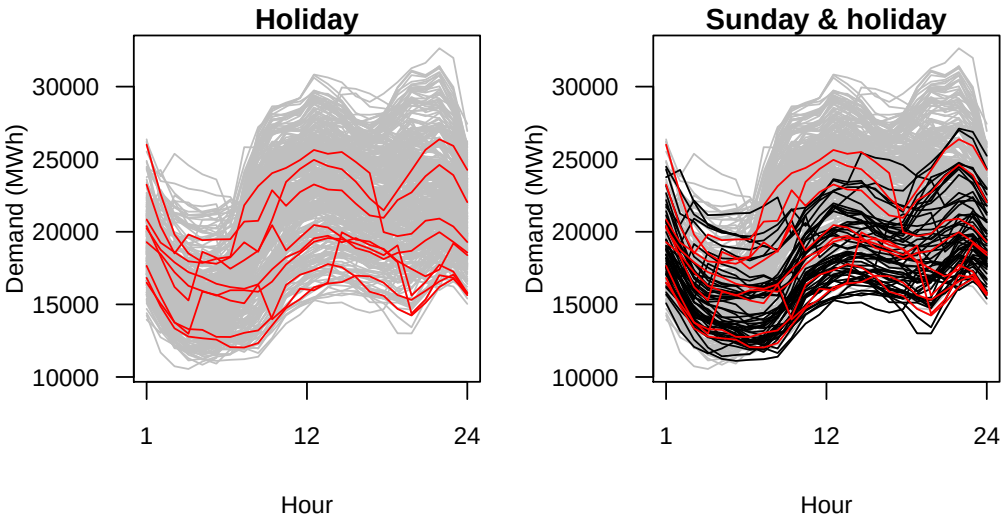


Figure 1.9: Left panel: Demand curves for holidays in red. Right panel: Demand curves for holidays (in red) and for Sundays (in black).

The next step is to analyse the holidays along the year. In Figure 1.9 the holidays are plotted over the daily demand curves for all the year. In the left, one can see only these holiday curves while, in the right panel, holidays are plotted together with Sunday demand curves. First of all, looking at the holiday curves, one can see that all of them take medium or low values. They will never reach the highest demand of the year and some of them present

also a different pattern along the day, maybe with a more flat shape. These holidays are very similar to Sundays, both in terms of demand curves and also in terms of social behaviour, as can be seen in the right panel of Figure 1.9.

One can analyse also the typical measures of centrality within a group of trajectories when working with functional data. That is the mean, median and mode. Functional median corresponds to the deepest curve with the order given by a depth measure. Figure 1.10 plots together the functional mean, the functional median (according to the L^2 norm) and the functional mode (according to the mode depth). See Section 1.3 for details about these measures. The three curves are very similar, with slight disturbances among them.

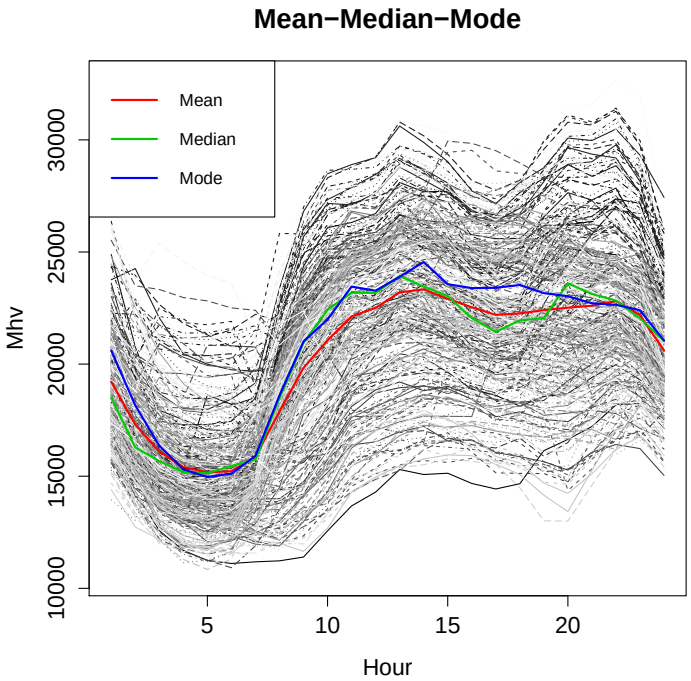


Figure 1.10: Mean, median and mode for the electricity demand in Spain, year 2012.

1.4.2 Electricity Price

Electricity price shares some of the characteristics with the demand. However, it has some particular properties. The most notable one are the days with price zero. Looking at the time series of the electricity price (Cents/kWh) along 2012 in Figure 1.11, one can see some days in which price decreases, reaching the value zero. This is the real price obtained in the bidding of the daily energy market. The price decreases in some days due to the overproduction of wind power. In Spain, renewable energies are subsidized and have a preferential position in the auction, entering free of charges. For that reason, when the production of wind power increases, the price generally decreases, even until zero.

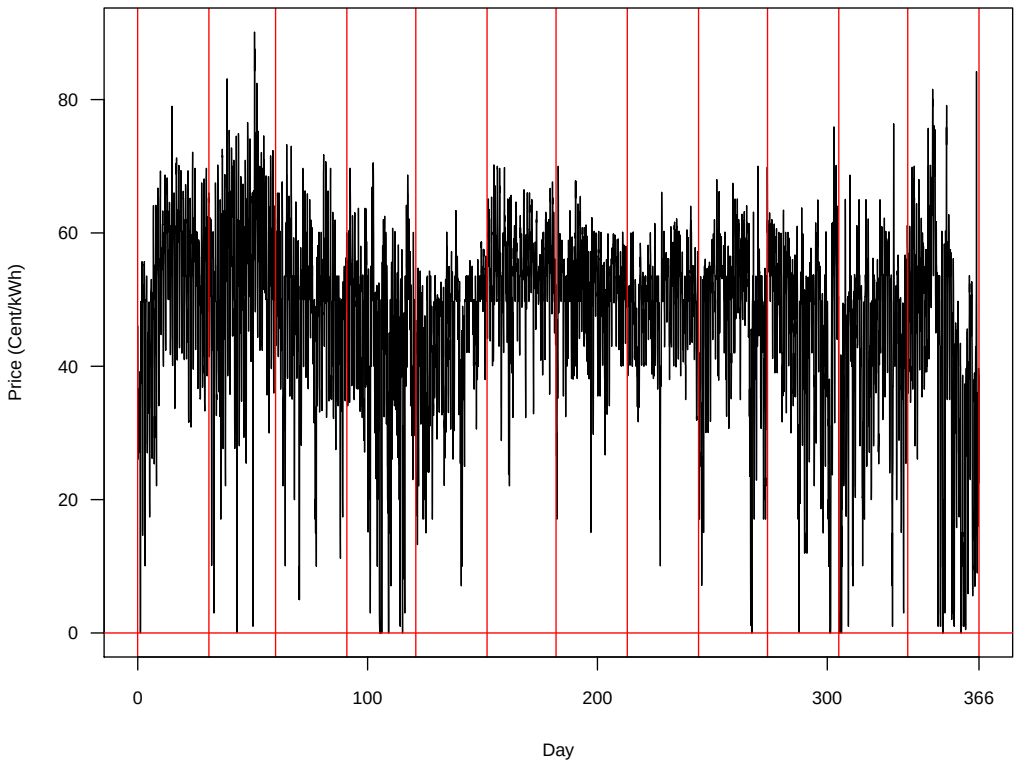


Figure 1.11: Functional time series of the electricity price in Spain in 2012.

The different scale between demand and price is also remarkable. Demand values covers from 10000 until 30000 MWh while price goes from zero to 80 Cents/kWh. However, even if they are in different levels, one can observe similar patterns like the descent at the first hours of the day. There are also some variations in the behaviour of the weekdays and the weekend, but it is not as significant as in the demand.

Zero-price days are concentrated in some parts of the year which are maybe related to the climate conditions that are favourable to the production of wind power. Specifically, zero-price was reached during some hours of the next days in 2012: January: 2; April: 15, 16, 19 and 25; September: 24; October: 28; November: 1 and 2 and December: 16 and 24.

This fact will be analysed in detail in Subsection 1.5.2, in which the relation between wind power production and electricity price will be studied.

The last graph is enlarged to show only the period correspondent to one month of the year. One can appreciate the details of the price behaviour. Figure 1.12 shows the time series for the price only during January 2012. It is clearly appreciated a daily periodicity, a seasonal component of period 24. Nevertheless, the weekly periodicity of the demand is not so clear in the case of the price. One can see a slight decrease in some days corresponding to weekends but their values are very close to weekdays.

In Figure 1.13, all the daily curves for the price along 2012 are represented. Colour scale corresponds to the temporal ordering, from January (red) to December (violet). Days are overlapping as long as the time passes. First days of the year are barely distinguished in the surroundings of the curve cloud.

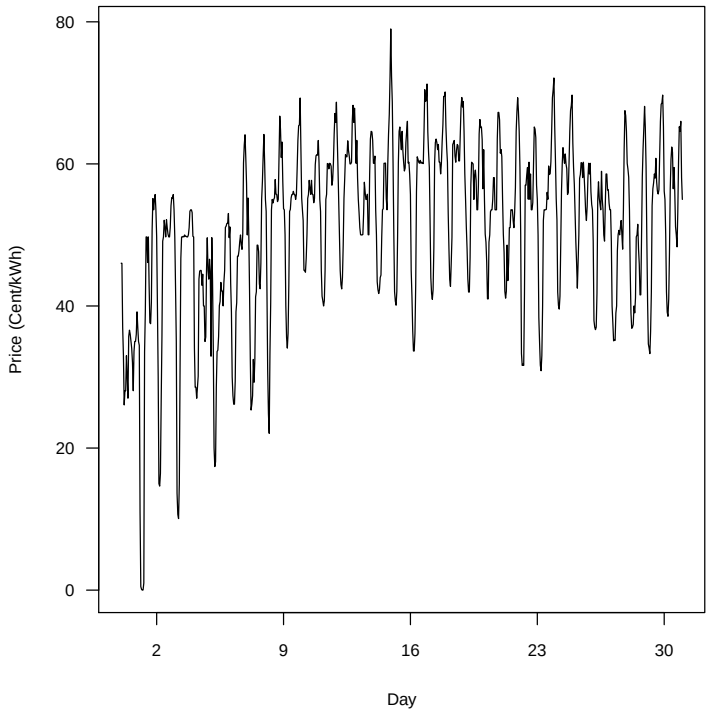


Figure 1.12: Time series of the electricity price in Spain in January 2012.

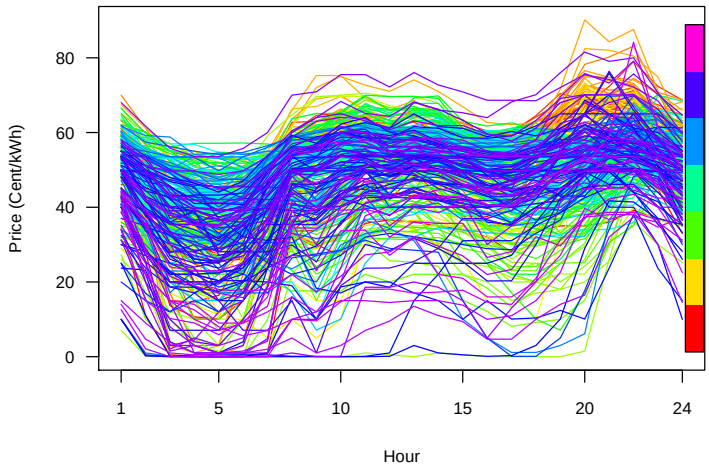


Figure 1.13: "Rainbow" plot for the daily price curves. Colour scale corresponds to the temporal ordering, from January to December.

Electricity price can be analysed in the different days of the week, as it was already done for the case of electricity demand. Figure 1.14 plots separately the price curves for weekdays, from Monday to Friday. The functional mean for each group is represented with a dotted line. The five weekdays are very similar among them. One cannot appreciate changes in their patterns, maybe a barely decline in the first hours of Mondays, but not very significant.

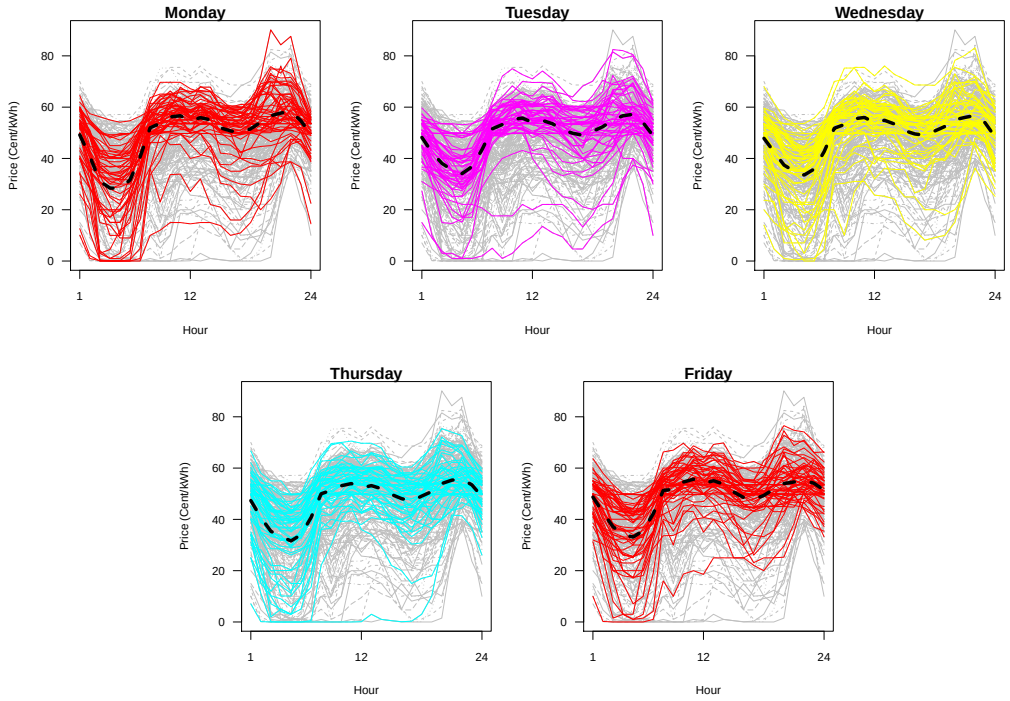


Figure 1.14: Weekdays daily price curves.

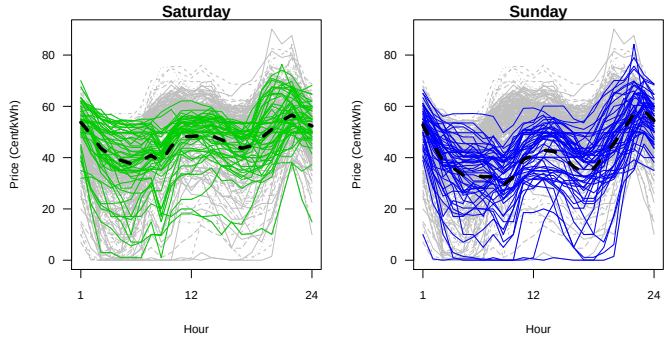


Figure 1.15: Daily price curves for Saturday and Sunday.

Figure 1.15 represents the price curves for the Saturdays and the Sundays. In this case, one can see more differences with respect to the weekdays curves. Saturdays have a smooth pattern, taking values very stable along the hours of the day. Sundays are a little bit difference from the rest of the days, because their values are generally lower specially in the central part of the day. This fact enforces to keep the same distribution in three groups of days: weekdays, Saturdays and Sundays.

In Figure 1.16 the three considered groups of days are represented. This classification is coincident with the one of electricity demand, even if the differences in the case of the price are not so pronounced.

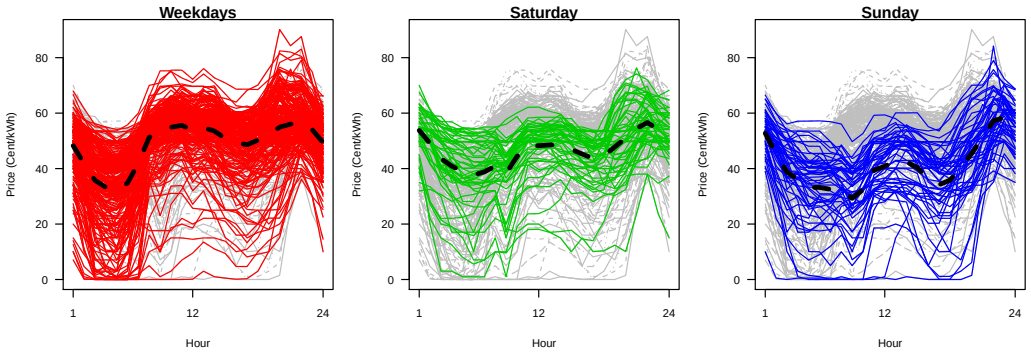


Figure 1.16: From left to right: Weekday, Saturday and Sunday daily curves. Dotted black lines represent the functional mean of each group of days.

Following the outline of the demand analysis, the daily curves for the electricity price will be compared now along the year. Figure 1.17 plots separately the four quarters of the year. Again, weekdays are plotted in black and the weekend in red. The first thing that comes to the mind, looking at this graph, is that price remains in the same values along the whole year. As for the demand, the last quarter of the year corresponds to a turbulent period in the market in which the variation is higher, reaching both lowest and highest values of the year at different days. Also some differences between weekdays and the weekend can be seen, as the weekend usually occupies the central band of the graph. Generally, lowest values are reached in the first hours for some weekdays. It is clear that, unlike the demand case, in which difference between weekdays and weekend were related to the level, for the price this difference is more accused looking at the shape of the curves.

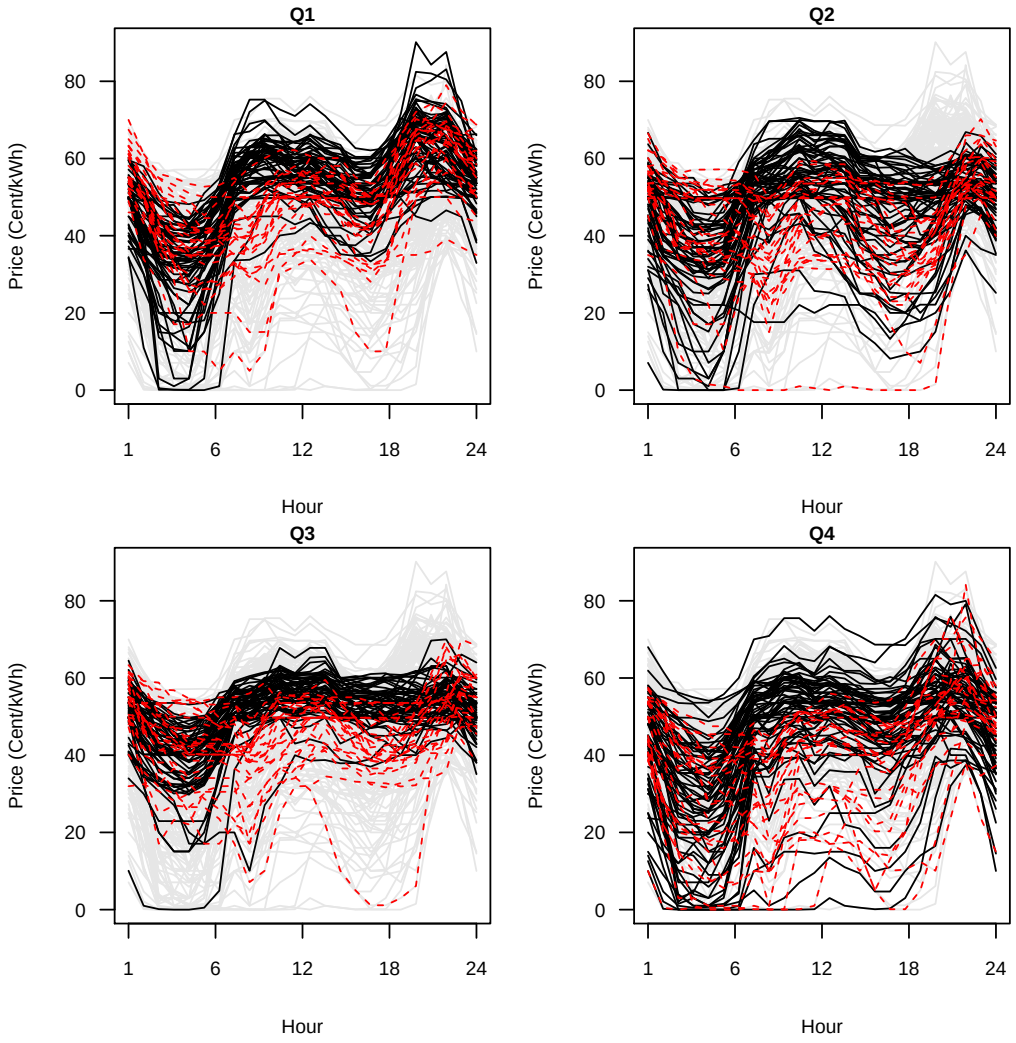


Figure 1.17: Daily curves of electricity price for each quarter of the year 2012. Red lines correspond to weekend.

Figure 1.18 goes a step forward analysing the price curves at each month of the year. There are some months, like April or December in which the fluctuation is higher and do not present any stable pattern. However in the centre of the year, specially June, July and August, all the curves are concentrated in a band that is also very flat. This is an indicator of stability in the Electricity Market. The first and last months of the year take the highest values for the price but, again, not so remarkable as in the case of the demand.

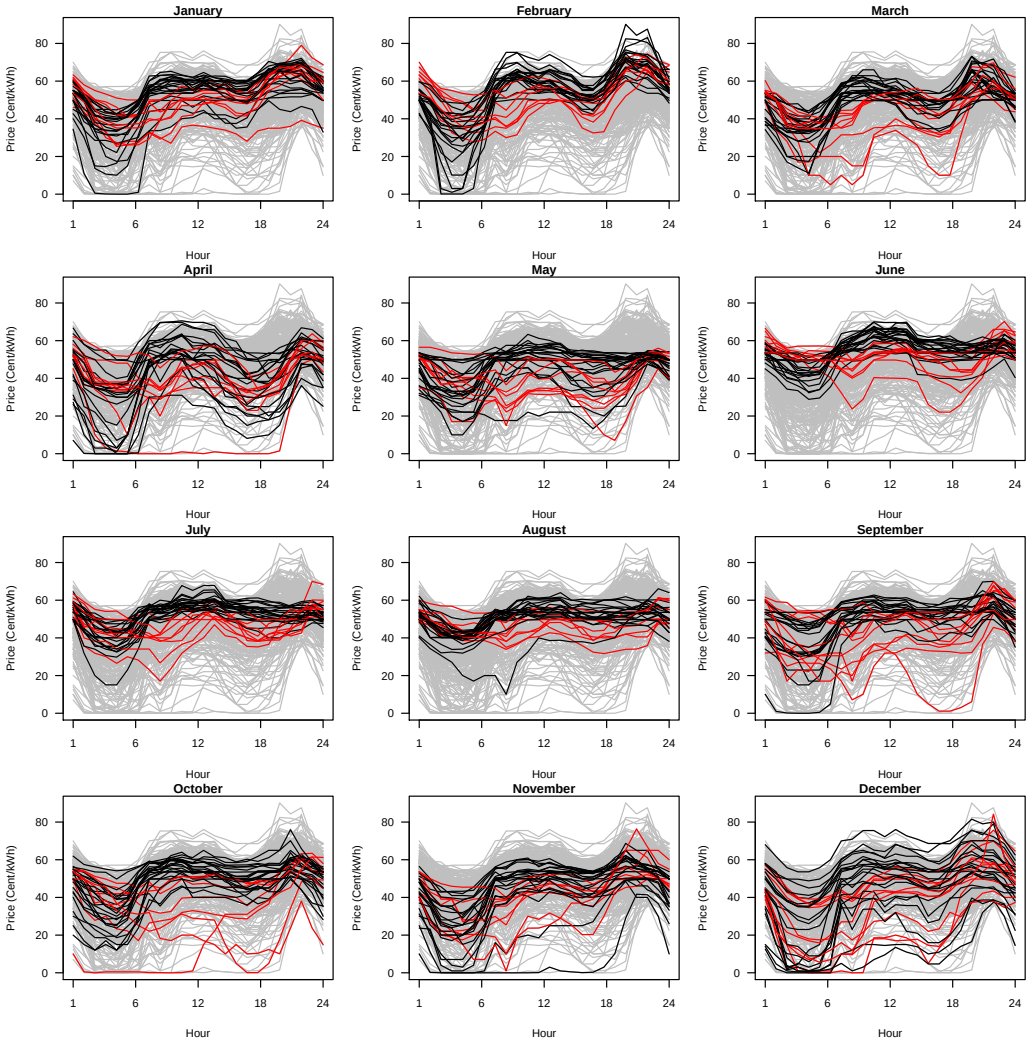


Figure 1.18: Daily price curves for each month of the year 2012.

Figure 1.19 plots the daily price curves correspondent to the holidays. These curves are represented alone and together with the Sundays. It is clear that they take always medium or low prices and that they are very similar to the Sundays.

Classical measures of centrality, as the functional mean, median and mode are represented in Figure 1.20. Median and mode are coincident in this case, being both slightly higher than the mean. Functional mean is lower, specially in the first hours of the day due to the influence of the zero price.

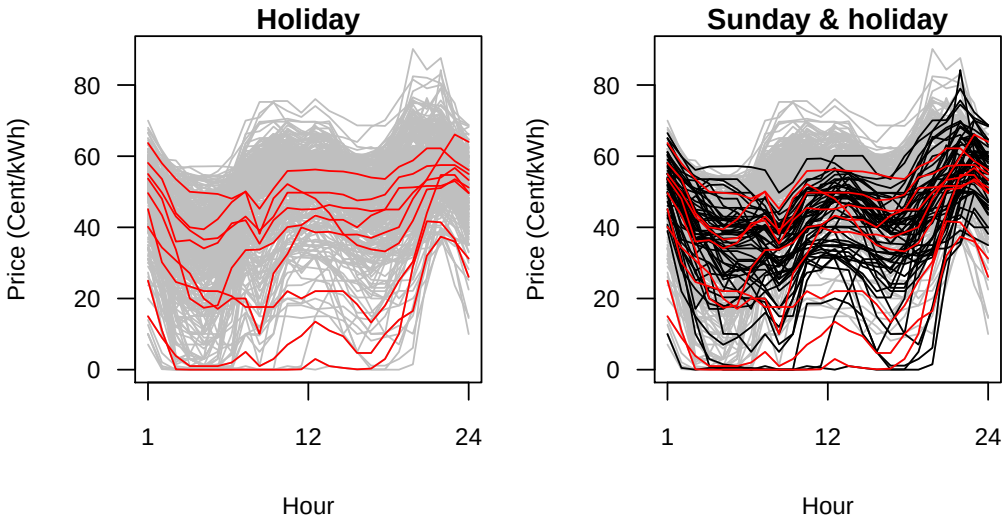


Figure 1.19: Left panel: Price curves for holidays in red. Right panel: Price curves for holidays (in red) and for Sundays (in black).

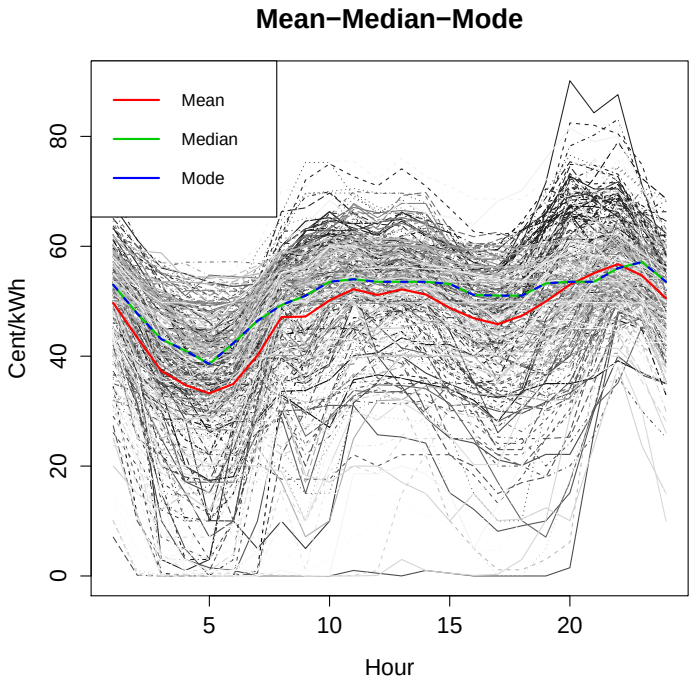


Figure 1.20: Functional mean, median and mode for the electricity price in Spain in 2012.

1.5 Additional data

This Section presents the additional data, used in Chapter 3 as external covariates for some of the prediction methods, that improves the demand and price forecasts. There are many options in the literature about which external covariates influence this kind of predictions, most of them related to weather conditions, Electricity Market operations and economic factors. Obviously, the more accurate the data is, the better predictions one can obtain.

Available data is not always enough to cover all the needed information to build a good prediction model. Sometimes this data is not free and it is difficult to record or to implement in the models. For that reason, only two external sources of information are taken into account: temperature and wind power production. The first one covers an important part of the weather influence over the demand, while wind power production explains most of the price reductions, as the hours with zero price mentioned in Subsection 1.4.2.

Both sources of additional data are analysed in the following two subsections, temperature data in Subsection 1.5.1 and the wind power production in Subsection 1.5.2.

1.5.1 Weather information: Temperature

Temperature has a high influence in the electricity demand and so, it can contribute to improve its predictions. This relation can be attributed, among other reasons, to the use of climate systems. When the weather is cold, electricity demand rises due to the use of heating. However, when the temperature is very high the use of air conditioning also contributes to an increase in the demand. This happens typically during winter and summer, whereas in other periods with warm weather, the temperature does not have influence over the electricity demand. Depending also on the place, the climate changes and the influence of the low and high temperatures may not be equal.

Other weather variables could be taken into account, as the humidity, the sun hours at each day, the amount of rain, the pressure, etc. However, the temperature summarizes very well the changes in the climate that are related to energy consumption. This study considers the maximum daily temperature (°C) in Spain. AEMET (Agencia Estatal de Meterología) provides the maximum daily temperature for each province of the country. By population-weighted average, the corresponding maximum daily temperature for Spain was built. Population data were collected from INE (Instituto Nacional de

Estadística).

It is worthy to highlight the nonlinear effect of meteorological variables on the electricity demand. In Figure 1.21, one can see that the effect of the maximum daily temperature over the daily mean demand is U-shaped. This nonlinear relation has to do with the use of the heating, when the temperature is low, and the use of air conditioning when it is high. In both cases, the use of these climatization systems increases the demand as it was already explained above. There is also a "comfort zone", estimated between 20 and 24°C, with no effect. It corresponds to the days when neither heating nor air conditioning is needed. See Pardo et al. (2002) for a detailed analysis of the relation between these temperature variables and the Spanish electricity demand and also Cancelo and Espasa (1996), who estimated the "comfort zone" between the indicated values of 20 and 24°C.

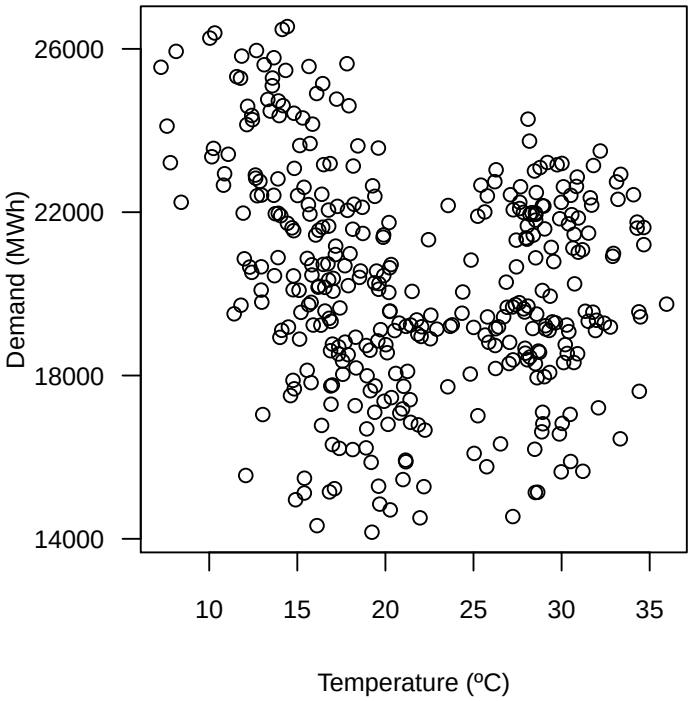


Figure 1.21: Maximum daily temperature against daily mean demand in Spain, year 2012.

In the application to electricity demand, only exogenous variables with linear effect will be considered. Therefore, a transformation of the tempera-

ture data is needed. Two new variables are built, HDD (Heating Degree Days) and CDD (Cooling Degree Days), that are a measurement of the amount of energy needed to heat/cool a building. Specifically, they are defined as

$$HDD(t) = \max\{20 - T(t), 0\} \tag{1.12}$$

$$CDD(t) = \max\{T(t) - 24, 0\}. \tag{1.13}$$

where $T(t)$ denotes the maximum daily temperature in Spain at day t.

Through this transformation, one obtains two variables that summarize all the temperature information but with linear effect over the demand (see Figure 1.22).

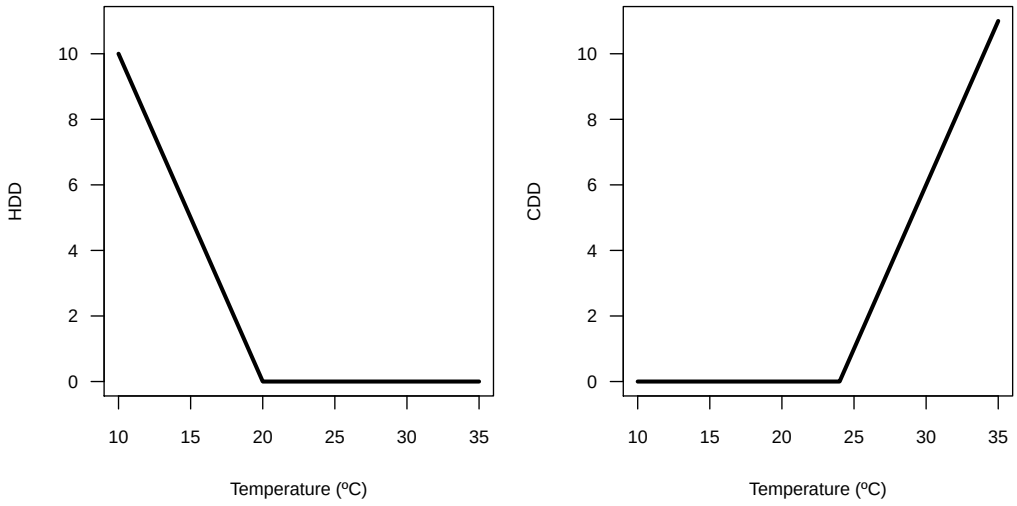


Figure 1.22: HDD (left panel) and CDD (right panel) variables.

1.5.2 Wind Power Production

In the same way as to predict the electricity demand it is used information about the temperature, to predict electricity price information about the wind power production (MWh) will be used.

Section 1.4.2 summarized the particular feature of the zero-price days. This reduction in the price at some moments is directly related to the increase in the renewable energies production, mainly wind power, as it was also pointed out in Section 1.2.

Specialized press and other sources of information, as well as the dossiers from the System Operator in the Spanish Electricity Market, Red Eléctrica de España (REE) report about this fact. Also Geidel and Zareipour (2013) included the wind power in the price forecasting within the Spanish Electricity Market. In the following paragraph, some examples are cited about the increase, at some particular moments, of the wind power. Those days correspond to zero-price periods in our dataset, mentioned in Subsection 1.4.2:

El País (2011/11/7): The contribution of wind energy to the grid sets new record. Wind turbines covered 59% of total demand in early Sunday to coincide with the storm

Cinco Días Journal (2012/04/16): Wind energy sets record by producing 60% of electricity

EUROPA PRESS (2012/09/24): Wind power beats a new record by providing 64% of electricity this morning.

In this study, data about wind power production was obtained from REE. In their web page, they monitor the demand and the generation structure at each moment. Then, one can know the amount of demand covered by wind power during each ten minutes period of the year and, therefore, calculate the corresponding value for each day.

Relation between daily mean wind power production and daily mean price is plotted in Figure 1.23. Contrary to the temperature influence over the demand, now the relation between wind power production and price is linear. As long as the wind power production increases, the electricity price decreases.

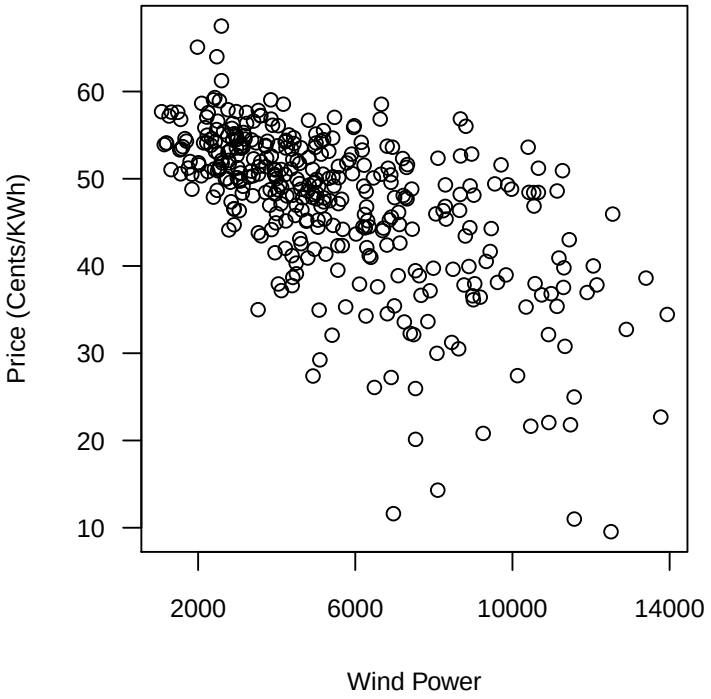


Figure 1.23: Daily mean wind power production against daily mean price in Spain, year 2012.

Chapter 2

Outlier detection in Functional Time Series

2.1 Introduction

In every statistical process that involves prediction, among other analysis, it is important to take into account the presence of outliers. An outlier, in the classical context of univariate data, can be defined as an observation that is distant from the remaining data or that lies outside the overall pattern of a distribution. This kind of "abnormal" observations can disturb the result of the statistical methodologies applied to a dataset. Electricity data is not an exception and the presence of outliers is actually one of its main features, as it was pointed out in Section 1.4.

The most classical tool to detect univariate outliers is the boxplot introduced in Tukey (1970) and Tukey (1977). This is a descriptive procedure to figure out which observations, within a group of numerical data, are outliers. It employs the data quantiles.

Pointwise boxplots can be applied to the electricity demand and price data, considering each hour of the day separately instead of a curve. In that way, one can observe which days, for a fixed hour, are outliers. Figure 2.1 plots the 24 hourly boxplots for the electricity demand along the year 2012. One can see a boxplot for each one of the 24 hours of the day. There are not many outliers present in this graph. At the bottom of the figure one cannot see any outlier and, even in the upper part, is difficult to find them. They are collected in the first hours of the day, being all of them outliers because of their unusually high values.

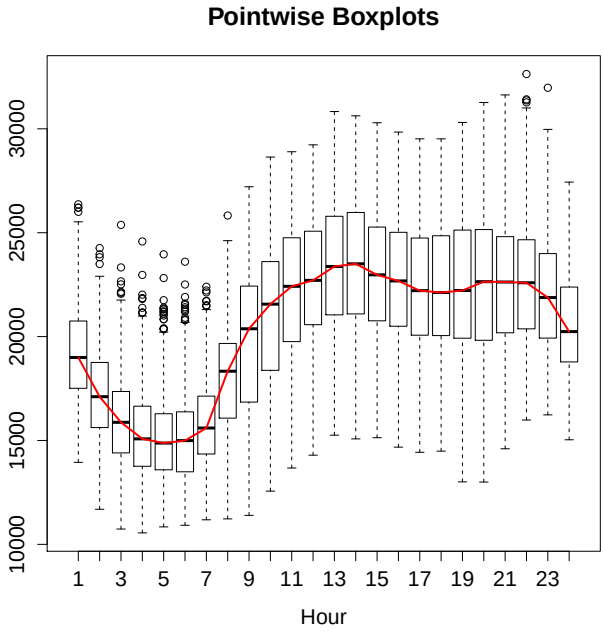


Figure 2.1: Hourly boxplots for the electricity demand.

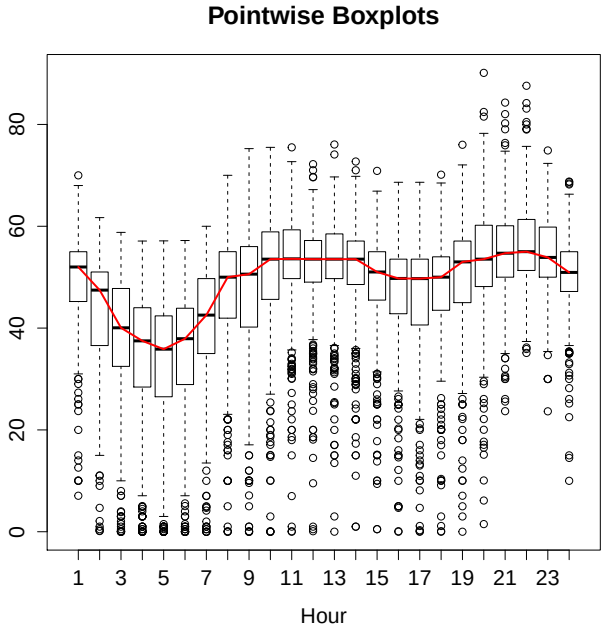


Figure 2.2: Hourly boxplots for the electricity price.

In the case of the price (Figure 2.2) the situation is almost the opposite, one can see several outliers, being the majority at the bottom of the graph. They are also spread along the 24 hours of the day. It is clear that many of the outliers correspond to unusual low recorded prices, including all the zero-price hours.

This kind of tools can give general ideas about the presence of outliers in the electricity data. For example, one may expect outliers in the price to have unusual low values. Meanwhile, one cannot say that one day, within all the year, is an outlier just because any hour, between the 24 hours of the day, is outlier individually. It is then needed a tool which allows to determine if this day, considering it as the whole curve, as a functional datum, is an outlier itself. That is, to detect outliers taking into account the functional nature of the data. This is not an easy task, as functional data is not always easy to visualize or to order. As a matter of fact, a rigorous definition of functional outlier needs to be proposed.

Febrero, Galeano and González-Manteiga (2008) and Hyndman and Shang (2010) consider that "a curve is an outlier if it has been generated by a stochastic process with a different distribution than the rest of the curves, which are assumed to be identically distributed", while Gervini (2012) considers that "an observation can be seen as an outlier if it is far from most of the other observations". In any case, the presence of outliers has serious adverse effects on most statistical approaches. Along this study, the definition given by Febrero, Galeano and González-Manteiga (2008) is used and applied to Functional Time Series, which are valid for the context of electricity data dealt in this memory.

In the general context of functional data, we can distinguish two kinds of outliers: the "magnitude outliers", that arise when they lie outside the range of the majority of the data, and the "shape outliers", that can be in the same range but differ from the rest of the data in shape. Also combinations of these two types can give another new type.

Some statistical methods to detect outliers in the field of functional data can be found in the literature. Some of them are based on descriptive analysis or the ordering given by the functional depths while other are based, for instance, on FPCA. To the best of our knowledge, none of these methods take into account the temporal dependence that exits in the functional time series defined in Section 1.3. Daily curves of electricity demand and price conform a functional time series as they are composed of several observa-

tions that can be temporally ordered, day after day. Meanwhile, within each curve, also data is ordered hour after hour taking into account the time. In summary, this study works with observations from one entire year, breaking it in intervals of time that are homogeneously distributed along the time.

Taking into account the dependence of the functional time series prevents outlier detection methods to give misleading results and, over all, enables to detect outliers hidden by the dependence structure of the data. This can be seen in Figure 2.3. Looking at the set of curves in the right panel, one can easily detect two outlying curves that arise with abnormal high values; they are magnitude outliers. Meanwhile there is another curve plotted in bold line that is totally within the range of the rest of the data. Apparently, there is no reason to suspect that this curve is an outlier, neither because of magnitude nor because of shape. However, one may look at the left panel, in which data is represented as a functional time series plotting one curve after the other, following the temporal ordering. It is clear that the behaviour of the time series abruptly changes at three different points. Two of these leaps correspond to the observed magnitude outliers, meanwhile the third corresponds to the bold line that is between the rest of the curves. Therefore, this third curve is also a magnitude outlier overlapped by the rest of the curves, due to the dependence structure of the functional time series. It can only be found taking into account the dependence in the outlier detection methods. Following this argument, three different methods to detect outliers in functional time series are proposed, which are, as far as we know, the first methods specifically addressing this problem.

The rest of the chapter is organized as follows: Section 2.2 contains a brief review of some methods present in the literature to detect outliers in functional data. Proposed methods to detect outliers in the context of functional time series are presented in two parts. First, Section 2.3 includes one proposal based on depths. Then, Section 2.4 includes two different proposals, both based on FPCA. Each one of these two sections comprehends simulations in order to study the proposed procedures behaviour. Finally, an application to the electrical dataset, for both demand and price, is given in Section 2.5.

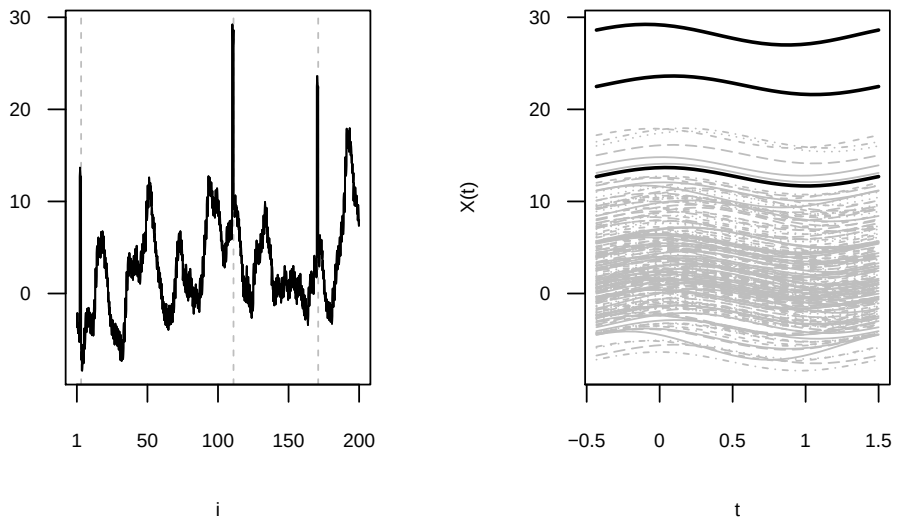


Figure 2.3: Example of hidden functional outliers. Left: Functional Time Series. Right: Set of curves in grey and outliers in black.

2.2 Overview of outlier detection in Functional Data

This section includes a review of some outlier detection methods, proposed in the context of functional data, which are present in the statistical literature. Most of them, in the same way as the study of the functional data itself, are very recent. They were proposed in the last ten years. This fact gives an idea of the novelty of the problem and that there are not many options to deal with it. Also, some of these methods will establish the base from which the proposals presented in this chapter emerge.

Outlier detection methods usually make use of graphical and descriptive tools, once you have an order within the analysed group of data. In the case of functional data, this ordering will be given by functional depths.

Depths of functional data were introduced to measure how deep (central) or outlying an observation is, with respect to a functional dataset. This allows to order a set of functional data from the centre-outward, so that the

most central or interior data in the sample will have higher depth. Thus, functional depths may indicate which observations (if any) can be considered as outliers: those with unusually low depths. Along this chapter, three different functional depths are used: h-modal depth (MD), Band depth (BD) and Modified band depth (MBD), which were introduced in Section 1.3.

Febrero et al. (2007 and 2008) were the first authors dealing with this outlier detection problem in the context of functional data. They proposed a procedure based on the bootstrap. Also Hyndman and Ullah (2007) included a tool to detect outliers within their method for robust estimation of functional principal components. Hyndman and Shang (2010) constructed functional bagplots and functional highest density region boxplots, based also on principal component analysis. An extension of the classical boxplot was developed by Sun and Genton (2011) using functional depths. More recently, other proposals arise in this topic, based on principal components (see Sawant, Billor and Shin (2012) or Yu, Zou and Wang (2012)), based on functional depths, like in Gervini (2012) or the outliergram proposed by Arribas and Romo (2014), which is specifically address to the detection of shape outliers, or based on random projections (see Fraiman and Svarc, 2013).

What all these mentioned methodologies have in common is that they do not take into account the dependence present in functional time series, even in the case in which they are applied to some datasets where temporal dependence is present. Sun and Genton (2012) adapted their Functional Boxplot to deal with dependent data, mainly space dependence. However, it does not include the kind of "hidden" outliers pointed out in Section 2.1, which are the goal of this study.

The following paragraphs describe briefly some of these methodologies, in particular, the ones that play a main role in the development of the new proposals.

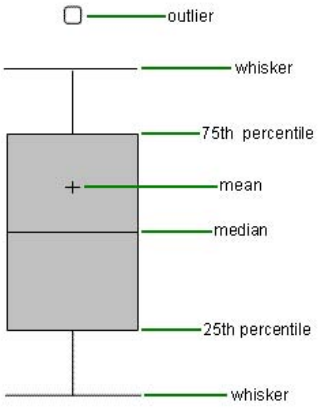
2.2.1 Functional boxplot

Functional boxplot is an extension of the classical boxplot to the context of functional data. That is, instead of working with "points" in the graph, it works with a sample of curves but keeping the same background. First of all, an ordering within the group of curves is needed. For that purpose, one can use any functional depth. Sun and Genton (2011) used the Modified Band Depth. Once the curves are ordered, one can build a band for the central

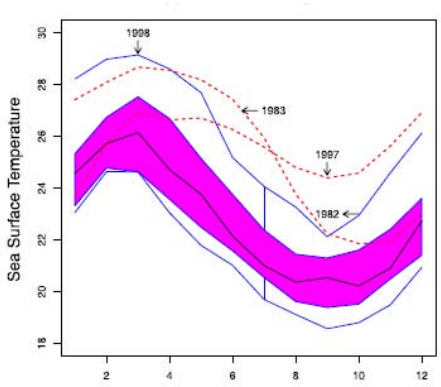
region. This is just an envelope of the 50% deepest curves of the sample and is analogous to the box present in classical boxplot and to the Inter-Quartile Rage (IQR). Once you have this band for the central region, one needs to figure out the "whiskers" that indicates which observations are outliers. Again, it follows the same idea as the classical boxplot in which any observation that arise above or below the whiskers is an outlier. To place this limits in the graph, the IQR is obtained and multiplied by 1.5. This means that it extends the envelope of the 50% deepest curves by 1.5 times the range of these curves. Now, instead of having two points as upper and lower limits of the graph, one has two lines as the "fences" of the graph.

It only rests to indicate the procedure for outlier identification with the functional boxplot. In this case, one curve will be considered as outlier if it is outside the envelope given by the "whisker curves". Even if the curve only exits the fence at one point, all this functional datum will be considered outlier.

In Figure 2.2.1 one can compare the classical boxplot (in the left panel) with the functional boxplot (in the right panel) and see their similarities. In summary, this functional boxplot is a simple descriptive method to detect outliers that conforms the natural extension of the classical boxplot to the context of functional data.



(a) Boxplot.
Source:
https://statsmethods.wordpress.com



(b) Funcional boxplot.
Source: Sun and Genton (2011)

2.2.2 Depth-based trimming

Febrero et al. (2008) were the first in presenting a specific method to detect outliers in functional data. It is based on developing a hypothesis test for each curve on a sample, in order to know if it is an outlier or not, using the order given by a functional depth. Knowing that the outlier curves correspond to low depths, they are intended to estimate a cutoff to say when a curve has low enough depth to be considered outlier. A brief description of the procedure is given below. It was implemented in the R library *fda.usc*.

The depth-based trimming procedure takes into account the fact that an outlier should have a depth that is significantly low. These authors proposed the following general procedure, given a functional sample $\mathcal{S} = \{\chi_i\}_{i=1}^n$, a functional empirical depth $D_S(\cdot)$ and a constant cutoff $C > 0$:

1. Obtain the functional depths of the curves in $\mathcal{S}$: $\{D_{\mathcal{S}}(\chi_i), \chi_i \in \mathcal{S}\}$.
2. Identify as outliers the curves whose depths are below C.
3. Repeat Step 1 and 2 until no more outliers are detected.

This procedure performs multiple tests and depends on the functional depth $D_S(\cdot)$ and the cutoff C, which indicates when a depth is considered significantly low. The key point is the determination of C. They propose two versions of their method, one based on trimming and the other based on weighting, but here only the first option will be considered. This method will be the base for the first proposal (Depth-based trimming for dependent data) that consists in adapting this procedure to deal with functional time series. It will be described in detail in Section 2.3.

2.2.3 Integrated squared error

This method to detect functional outliers, based on Integrated Squared Error (ISE), is extracted from the robust FPCA given by Hyndman and Ullah (2007), which is explained in detail in Section 1.3. They first propose estimating the functional principal components by means of the functions $\widehat{\phi}_k(\cdot)$ that maximize the variance of the scores

$$z_{i,k} = w_i \int_a^b \phi_k(t)\chi_i(t)dt \tag{2.1}$$

subject to the constraints:

$$\int_a^b \phi_k^2(t)dt = 1 \text{ and } \int_a^b \phi_k(t)\phi_j(t)dt = 0 \ (k \neq j).$$

The weights w_i are computed as

$$w_i = \begin{cases} 1 & \text{if} \quad v_i < S + \lambda\sqrt{S} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

where

$$v_i = \int_a^b (\chi_i(t) - \sum_{k=1}^{K} \widetilde{\beta}_{i,k}\widetilde{\phi}_k(t))^2 dt \tag{2.3}$$

with $\widetilde{\phi}_k(\cdot)$ being initial (highly robust) projection-pursuit estimates of $\phi_k(\cdot)$ obtained from the RAPCA algorithm (see Hubert, Rousseeuw and Verboven 2002) considering equal weights w_i in (2.1), while $\widetilde{\beta}_{i,k} = \int_a^b \widetilde{\phi}_k(t)\chi_i(t)dt$. In addition, S is the median of $\{v_1, \ldots, v_n\}$ and $\lambda > 0$ is a tuning parameter to control the degree of robustness. Once the robust estimates $\widehat{\phi}_k(\cdot)$ are obtained, the coefficients corresponding to the curve χ_i are constructed as

$$\widehat{\beta}_{i,k} = \int_a^b \widehat{\phi}_k(t)\chi_i(t)dt. \tag{2.4}$$

As a by-product of this robust FPCA procedure, Hyndman and Ullah (2007) proposed an outlier detection method (the ISE method): the curve χ_i is detected as outlier if $w_i = 0$. This is equivalent, using (2.2), to the following sentence: given the values of v_i, an outlier is the observation with ISE greater than a threshold $(s + \lambda\sqrt{s})$, with $\lambda = 3.29$.

2.2.4 Functional highest density region boxplot

The functional highest density region boxplot (HDR) is based on the bivariate HDR boxplot (Hyndman 1996) applied to the first two robust principal component scores. It was proposed by Hyndman and Shang, (2010). It looks alike the functional boxplot, in the sense of plotting a band with a fence which indicates the outliers as all the curves outside this fence. This can be seen in Figure 2.4 extracted from Hyndman and Shang, (2010).

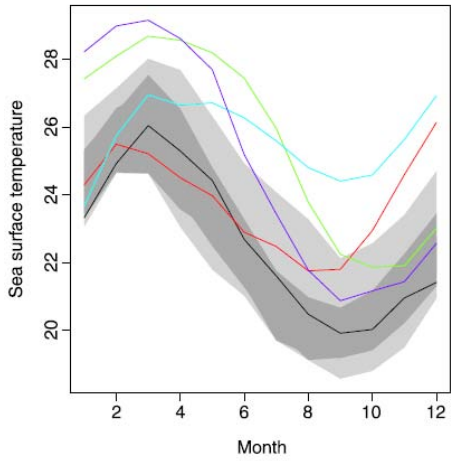


Figure 2.4: Example of Funtional HDR boxplot. Source: Hyndman and Shang, (2010).

Using a bivariate kernel density estimation, the HDR is defined as the region, with coverage probability $(1 - \alpha)$, where all the points within the region have higher density estimate than any of the points outside. Bivariate HDR boxplot displays the mode (highest density point), along with the 55% inner and 99% outlier highest density regions. All the points excluded from the outer HDR are outliers. Functional HDR boxplot is just a mapping of the bivariate HDR boxplot of the first two principal component scores to the functional curves. Thus, it displays the modal curve (curve with highest density), and the inner and outer regions which corresponds to the grey bands in Figure 2.4. The curves that arise out of these grey regions at any point are detected as outliers (see the coloured curves in Figure 2.4 as an example).

2.3 Depth-based trimming for Functional Time Series

This section presents the first proposal to detect outliers in FTS: Depth-based trimming for FTS (DBT). This approach is based on the method by Febrero et al. (2008), which was briefly explained in Subsection 2.2.2, changing some steps along the procedure and mainly adapting it to take into account the dependence in FTS.

More in detail, DBT follows the next general procedure (summarized in Subsection 2.2.2), given a functional sample $\{\chi_i\}_{i=1}^n$, a functional empirical

depth $D_S(\cdot)$ and a constant cutoff $C > 0$.

General procedure:

- **Step a**. Set $\mathcal{S} = \{\chi_i\}_{i=1}^n$ and $\mathcal{O}=\varnothing$.
- **Step b**. Obtain the functional depths of the curves in $\mathcal{S}$: $\mathcal{D}_\mathcal{S} = \{D_\mathcal{S}(\chi_i), \chi_i \in \mathcal{S}\}$.
- **Step c**. Set $\mathcal{S} = \mathcal{S} \setminus \mathcal{A}$ and $\mathcal{O} = \mathcal{O} \cup \mathcal{A}$, where $\mathcal{A}$ denotes the set of curves whose depths obtained in Step b are below C.
- **Step d**. Repeat Step b and Step c until $\mathcal{A} = \varnothing$.
- **Step e**. Establish the curves in $\mathcal{O}$ as outliers.

(Note that Step d is introduced to avoid masking effects). This procedure performs multiple tests, and depends on the functional depth $D_S(\cdot)$ and the cutoff C, which indicates when a depth is considered significantly low. The key point is the determination of C.

Febrero et al. (2008) proposed to select C in such a way that, in the absence of outliers,

$$P(D_\mathcal{S}(\chi_i) \leq C) = \alpha_1, \ i = 1, \ldots, n,$$

where $\mathcal{S} = \{\chi_i\}_{i=1}^n$; that is, the signification level of each individual test is α_1. Given that the distribution of the functional depth is unknown, they used bootstrap-based methods to estimate C (note that C is the quantile of order α_1 of such distribution when no outliers exist).

Then, the general procedure to detect outliers in functional time series (DBT for dependent data) follows the steps given above, in which the cutoff C is selected as follows.

Selection of the cutoff C

- **Step 1**. Detect outliers in $\mathcal{S} = \{\chi_i\}_{i=1}^n$ using a graphical method, and define $\mathcal{S}_1$ as the subset of $\mathcal{S}$ without these outliers.
- **Step 2**. Use a bootstrap method applicable to dependent data to obtain B bootstrap samples of size n from $\mathcal{S}_1$. These bootstrap samples are denoted by $\mathcal{S}^b = \{\chi_i^b\}_{i=1}^n$ $(b = 1, \ldots, B)$.

- **Step 3**. Obtain C^b as the empirical quantile of order α_1 of the distribution of the depths $\{D_{\mathcal{S}^b}(\chi_i^b),\ \chi_i^b \in \mathcal{S}^b\}$.
- **Step 4**. Take C as the median of $\{C^b\}_{b=1}^B$.

Note that the bootstrap procedure is applied on the set $\mathcal{S}_1$ instead of $\mathcal{S}$; that is, the method is based on trimming the sample of suspicious curves. Obviously, trimming is carried out to obtain a robust estimate of C.

Based on the same procedure, one can develop different methodologies changing the functional depth, the way to obtain $\mathcal{S}_1$ and the bootstrap resampling. Febrero et al. (2008) proposed to obtain $\mathcal{S}_1$ in this way:

$$\mathcal{S}_1 = \{\chi_i \in \mathcal{S}, D_{\mathcal{S}}(\chi_i) > q_{\alpha_2}\},$$

where $\mathcal{S} = \{\chi_i\}_{i=1}^n$ and q_{α_2} is the quantile of order α_2 of the functional depths $\mathcal{D}_{\mathcal{S}}$. That is, they fix the quantity of "suspicious outliers" and remove them from the sample to get robust estimates of C. In our proposal, a graphical method to detect outliers is used: the functional boxplot described in Subsection 2.2.1. This election allows not to fix the quantity of "suspicious outliers" in advance.

Concerning the election of the functional depth, Febrero et al. (2008) tried three different depths, giving the recommendation of the modal depth (MD). Also the band depth (BD) and the modified band depth (MBD) will be added to the study. Those three depths have been already described in Section 1.3. Obviously, other functional depths can be chosen instead.

The last choice is the bootstrap method used in the Step 2 of the procedure described above. Here, four different ways to obtain the bootstrap samples S^b are presented. The first one does not take into account the dependence in the data but it is used to compare the original method by Febrero et al. (2008) with our modified version and also with the following options for dependent data. The other three are bootstrap techniques designed to resample real observations that are stationary and weakly dependent.

1. Standard Smoothed Bootstrap on Data (SmBoD)
   It follows the same bootstrap used in Febrero et al. (2008), that is, it is designed for independent data but it is useful to compare the behaviour of the proposals for dependent data. In this case, B standard bootstrap resamples of size n from $\mathcal{S}_1$ are obtained. The bootstrap resamples are denoted by $\{\chi_i^b\}_{i=1}^n$ $(b = 1, \ldots, B)$ and they are smoothed

by adding normal perturbations, z_i^b, with mean $\mathbf{0}$ and covariance matrix $\gamma\widehat{\Sigma}_\chi$, where γ is a smoothing parameter and $\widehat{\Sigma}_\chi$ denotes the sample covariance matrix of the discretized curves $\{(\chi_i(t_1),\ldots,\chi_i(t_m))\}_{\chi_i\in\mathcal{S}_1}$. Let $\mathcal{S}^b=\left\{\zeta_i^b\right\}_{i=1}^n$ be the corresponding smoothed bootstrap samples, where $\zeta_i^b=\chi_i^b+z_i^b$ $(b=1,\ldots,B)$.

2. Moving Blocks Bootstrap (MBBo)
   A block length is chosen to generate a bootstrap realization of a time series; all possible contiguous blocks of this length are considered. The blocks are then sampled with replacement, and pasted together to form the bootstrap time series. Let l and n denote the length of the block and the sample size, respectively, and write each block as

$$B_{i,l}=\{\chi_i,\chi_{i+1},\ldots,\chi_{i+l-1}\},\ i=1,2,\ldots,n-l+1.$$

   In addition, let $I_1,I_2,\ldots,I_k$ (where k is the smallest integer that is greater than or equal to n/l) be a sequence of i.i.d. random variables having discrete uniform distribution on $\{1,\ldots,n-l+1\}$, and consider a realization $i_1,i_2,\ldots,i_k$. The bootstrap time series is then obtained by pasting $B_{i_1,l},B_{i_2,l},\ldots,B_{i_k,l}$ and removing the last $kl-n$ observations. This procedure, proposed in Künsch (1989), generates non-stationary time series.

3. Stationary Bootstrap (StBo)
   Politis and Romano (1994) modified the moving blocks bootstrap by resampling blocks of random lengths, and thereby proved that this procedure generates stationary time series. The blocks are constructed as

$$B_{i,l}=\{\chi_i,\chi_{i+1},\ldots,\chi_{i+l-1}\},\ i=1,2,\ldots,n,$$

   where, in the case $j>n$, χ_j is defined as χ_i, where $i=j(\mathrm{mod}\ n)$ and $\chi_0=\chi_n$. In addition, let $I_1,I_2,\ldots$ and $L_1,L_2,\ldots$ be sequences of i.i.d. random variables from a discrete uniform distribution on $\{1,\ldots,n\}$ and a geometric distribution with parameter p, respectively, and consider their realizations $i_1,i_2,\ldots$ and $l_1,l_2,\ldots$ The bootstrap time series is constructed as above using the blocks $B_{i_1,l_1},B_{i_2,l_2},\ldots$ instead of $B_{i_1,l},B_{i_2,l},\ldots,B_{i_k,l}$.

4. Standard Smoothed Bootstrap on Residuals (SmBoR)
   This option suggests a model-based procedure following the steps explained below, that are a bit different of the main structure of the method. Since in the three later options, the variations only affect the bootstrap procedure used in Step 2, in this case all the procedure to

select the cutoff is modified. The idea is to remove (or attenuate) dependence by extracting the autoregressive component in the data (see Step 2 below) so that the impact of the outliers (if any) on this extraction is low (see Step 1 below). Then, smoothed standard bootstrap is used as an intermediate step (see Step 3 and Step 4 below) to obtain bootstrap samples that maintain the dependence structure (see Step 5 below) and are only slightly influenced by the presence of outliers. These bootstrap samples are finally used to estimate C (see Step 6 and Step 7 below).

The procedure to select the cutoff C (Step 1–4) is substitute, when dealing with SmBoR, by the next algorithm:

- **Step 1**. Detect outliers in $\mathcal{S} = \{\chi_i\}_{i=1}^n$ using a graphical method, and define $\mathcal{S}_1$ as the subset of $\mathcal{S}$ without these outliers. Let $1 \leq i_1 < i_2 < \cdots < i_J \leq n$ the subscripts corresponding to the curves in $\mathcal{S}_1$ so that $\{\chi_{i_j-d}, \ldots, \chi_{i_j-1}, \chi_{i_j}\} \subset \mathcal{S}_1$ (note that $J \leq \#\mathcal{S}_1$).
- **Step 2**. Fit the autoregressive model

$$\chi_{i_j} = m(\chi_{i_j-d}, \ldots, \chi_{i_j-1}) + a_{i_j} \ (j = 1, \ldots, J) \tag{2.5}$$

  and construct the set of the corresponding residuals: $\mathcal{S}_a = \left\{\widehat{a}_{i_j}\right\}_{j=1}^J$, where $\widehat{a}_{i_j} = \chi_{i_j} - \widehat{m}(\chi_{i_j-d}, \ldots, \chi_{i_j-1})$ with $\widehat{m}(\cdot)$ being an estimator of the autoregressive functional $m(\cdot)$.
- **Step 3**. Obtain B standard bootstrap resamples of size J from $\mathcal{S}_a$. These bootstrap resamples are denoted by $\{\widehat{a}_{i_j}^b\}_{j=1}^J$ $(b = 1, \ldots, B)$.
- **Step 4**. Smooth the bootstrap resamples obtained in Step 3 by adding normal perturbations, z_j^b, with mean $\mathbf{0}$ and covariance matrix $\gamma\widehat{\Sigma}_a$, where γ is a smoothing parameter and $\widehat{\Sigma}_a$ denotes the sample covariance matrix of the discretized curves $\{(\widehat{a}_{i_j}(t_1), \ldots, \widehat{a}_{i_j}(t_m))\}_{j=1}^J$. Let $\left\{e_{i_j}^b\right\}_{j=1}^J$ $(b = 1, \ldots, B)$ denote the corresponding smoothed bootstrap resamples, where $e_{i_j}^b = \widehat{a}_{i_j}^b + z_j^b$.
- **Step 5**. Construct B bootstrap resamples $\mathcal{S}^b = \left\{\zeta_{i_j}^b\right\}_{j=1}^J$ $(b = 1, \ldots, B)$, where

$$\zeta_{i_j}^b = \widehat{m}(\chi_{i_j-d}, \ldots, \chi_{i_j-1}) + e_{i_j}^b.$$

- **Step 6**. Obtain C^b as the empirical quantile of order α_1 of the distribution of the depths $\{D_{\mathcal{S}^b}(\zeta_i^b), \zeta_i^b \in \mathcal{S}^b\}$.

- **Step 7**. Take C as the median of $\{C^b\}_{b=1}^B$.

Note that an estimator for the autoregressive functional $m(\cdot)$ in (2.5) must be constructed. For instance, in the case of nonparametric autoregression of order 1 ($d = 1$), the kernel regression estimator

$$\widehat{m}(\chi) = \frac{\sum_{j=1}^J K(h^{-1}d(\chi, \chi_{i_j-1}))\chi_{i_j}}{\sum_{j=1}^J K(h^{-1}d(\chi, \chi_{i_j-1}))}, \tag{2.6}$$

proposed in Ferraty et al. (2011) within a setting of independent data, could be used (in (2.6), $h > 0$ and $d(\cdot, \cdot)$ denote a bandwidth parameter and a semi-metric, respectively). See Horváth and Kokoszka (2012) for the estimation of parametric autoregressive models.

All the functional depths and bootstrap options presented in this section can be combined, resulting in different methods denoted by "depth"-"bootstrap" for each of the considered combinations. For instance, MD-MBBo refers to the election of the modal depth and the Moving Block Bootstrap.

2.3.1 Simulation study

A simulation study was carried out in order to compare the behaviour of the several proposals emerged from DBT method, depending on the different choices for depths and bootstrap. Approaches designed for independent data were also included in the comparison to illustrate the interest of taking into account dependence in the sample.

Simulated Functional Time Series

Functional time series were generated as follows. First, a functional time series (FTS) was obtained from the uncontaminated model

$$\zeta_i(t) = \begin{cases} \cos(\pi t) & \text{if} \quad i = -n+1, \\ \cos(\pi t)(1-\rho) + \rho\zeta_{i-1}(t) + a_i(t) + b_i & \text{if} \quad -n+1 < i \leq n, \end{cases}$$

where $t \in [-0.5, 1.5]$, $a_i(t) = X_i \sin(\pi t)$ with X_i being i.i.d. Gaussian variables with mean 0 and standard deviation 0.3, and $\{b_i\}$ is a scalar Gaussian AR(1) process with correlation coefficient ρ and standard deviation $(1-\rho^2)^{-1/2}$. Three outliers were then introduced at random in $\{\zeta_i\}_{i=1}^n$ to construct the contaminated time series. The analysed FTS were:

- Model with magnitude outliers:

$$\chi_i(t) = \zeta_i(t) + k1_{\{i\in\{I_1,I_2,I_3\}\}},\ 1 \leq i \leq n, \tag{2.7}$$

  where k is a contamination size constant, and I_j $(j = 1, 2, 3)$ are i.i.d. random variables from the discrete uniform distribution on $\{1, \ldots, n\}$.

- Model with shape outliers:

$$\chi_i(t) = \zeta_i(t) + k\cos(3\pi t)1_{\{i\in\{I_1,I_2,I_3\}\}},\ 1 \leq i \leq n. \tag{2.8}$$

- Model partially contaminated:

$$\chi_i(t) = \zeta_i(t) + k1_{\{\{i\in\{I_1,I_2,I_3\}\}\cap\{t\geq T_i\}\}},\ 1 \leq i \leq n, \tag{2.9}$$

  where T_i is a random variable with uniform distribution on $[-0.5, 1.5]$.

The curves were discretized on a grid of 30 equispaced points in $[-0.5, 1.5]$, the sample size was $n = 200$ and the correlation coefficient of the AR(1) process $\{b_i\}$ was $\rho = 0.8$. Values for the contamination size were $k = 10, 15, 20, 25$ for models (2.7) and (2.9), and $k = 4, 5, 6, 7$ for model (2.8). Figure 2.5 shows a simulated time series from each model (2.7), (2.8) and (2.9) considering the minimum value for k.

The implemented procedures and their tuning parameters

Along this simulation study, our proposed procedure (DBT) will be compared with the Functional Boxplot (Fbox) and the origin of our proposal, which is the method by Febrero el al. (2008), which is DBT for independent data (DBT-indep). Within DBT procedure, three functional depths were considered: the h-modal depth (MD), the band depth (BD) and the modified band depth (MBD) and also the four bootstrap procedures: Standard Smoothed Bootstrap on Data (SmBoD), Moving Blocks Bootstrap (MBBo), Stationary Bootstrap (StBo) and Standard Smoothed Bootstrap on Residuals (SmBoR).

The functional depths depend on several tuning parameters that were selected as follows: for the modal depth (MD), following the recommendation of Febrero et al. (2008), it is considered the L_2 norm $\|\chi_i\| = \left(\int_a^b \chi_i^2(t)dt\right)^{1/2}$ and the truncated Gaussian kernel $K(t) = \frac{2}{\sqrt{2\pi}}\exp(-t^2/2)1_{\{t\geq 0\}}$, while the bandwidth h was the 15^{th} percentile of the empirical distribution of $\{\|\chi_i - \chi_j\|\,; i, j = 1, \ldots, n\}$.

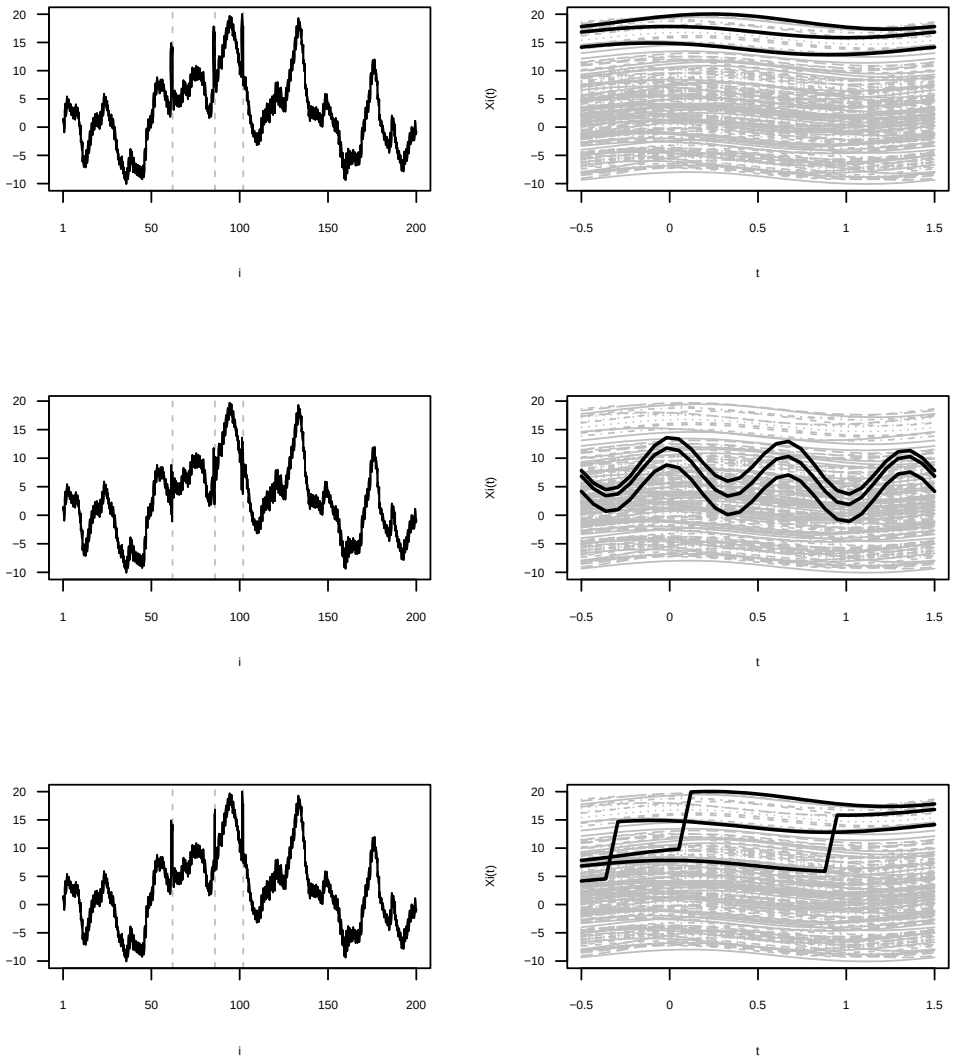


Figure 2.5: Left panels: from top to bottom, functional time series (i denotes the temporal index) generated from models (2.7), (2.8) and (2.9), respectively; the vertical dashed lines indicate the positions where the outliers emerged. Right panels: the corresponding curves $\chi_i(t)$ (the black curves are the outliers).

Actually, h plays a main role in the nonparametric estimation of the density. Nevertheless, as noted in Febrero et al. (2008), here the interest lies in the values around the center of the distribution, which are not very sensitive to the choice of the bandwidth. In fact, the only requirement to obtain good estimates is that the bandwidth should not be very small (see Cuevas and Fraiman, 1997; Cuevas et al., 2001). The smoothness of the curves justifies the choice of the L_2 norm (see Ferraty and Vieu, 2006). Finally, the kernel is known to have a low impact on density estimation (see Wand and Jones, 1995).

For the band and modified band depth $J = 2$ is used, as in most of the published applications (see Sun and Genton, 2011; Sun and Genton, 2012; Arribas-Gil and Romo, 2014). The order of the curves induced by these depths is known to be very stable in $J \geq 2$ (see López-Pintado and Romo, 2009). Thus, minimal value is usually considered in order to avoid computational issues.

The considered bootstrap methods also depend on tuning parameters. These parameters were selected as follows: for MBBo, the considered length of the blocks was $l = 4$. The value used for the parameter of the geometric distribution in StBo was $p = 0.1$. When dealing with SmBoR, a nonparametric autoregression of order $d = 1$ was fitted using the kernel regression estimator (2.6). The bandwidth h was selected using the cross-validation criterion and the Epanechnikov kernel

$$K(t) = 3/4(1 - t^2)1_{\{0 \leq t \leq 1\}}$$

was considered. Finally, the parameter γ used for smoothing the bootstrap samples was $\gamma = 0.05$ (recommended and used by Febrero et al., 2008).

Because of the "curse of dimensionality", orders $d > 1$ were not considered. Regarding the bandwidth, h, here the setting is different to the one noted in the case of the MD. Now the bandwidth needs to provide a very good fit to the regression function, so that the corresponding residuals mimic the random errors. Thus, h should be chosen by a data-driven selector that gives an asymptotically optimal value (see Rachdi and Vieu, 2007, for the case of the cross-validation selector). Note that the kernel has a low impact on the estimates and the Epanechnikov kernel is the most commonly used kernel for the estimation of nonparametric regression functions.

Note that the implemented SmBoD method differs from the proposal in Febrero et al. (2008) only in the first filtrate (that is, in the way $\mathcal{S}_1$ is con-

structed in Step 1): in Febrero et al. (2008), the $\alpha_2 n$ least deepest curves are deleted, while in SmBoD the curves removed are the ones detected as outliers by the functional boxplot.

All in all, the simulation study compares the behaviour of the three versions of the procedure proposed in this thesis (MBBo, StBo and SmBoR) which take dependence in the data into account, with other three approaches (FBox, DBT-indep and SmBoD) that do not.

The results

The methods for detecting outliers listed in the previous section were applied on $M = 100$ samples drawn from each model (2.7), (2.8) and (2.9). The signification level was $\alpha_1 = 0.01$ and the number of bootstrap samples was $B = 200$.

The percentage of correctly detected outliers, p_c, and the percentage of falsely detected outliers, p_f, were estimated for each one of the procedures using the mean of the corresponding empirical values, $\widehat{p}_c$ and $\widehat{p}_f$, respectively. See Tables 2.1, 2.2 and 2.3 for models (2.7), (2.8) and (2.9), respectively.

Several conclusions can be drawn from Tables 2.1–2.3. First, one may take a look at the empirical power, $\widehat{p}_c$; that is, the percentage of correctly detected outliers. In general, the power of the tests proposed in this paper (MBBo, StBo and SmBoR) is greater than the power of the tests corresponding to FBox, DBT-indep and SmBoD; so, the dependence in the data must be taken into account. In fact, the increase in the power is particularly drastic when the band depth (BD) is used. Furthermore, the power depends on both the depth and the kind of outliers considered: power achieves maximum values when band depth is considered in the case of magnitude outliers (Table 2.1), while h-modal depth (MD) is the most appropriate in detecting shape outliers (Table 2.2). Note that this agrees with the findings in the setting of functional outlier detection for independent data, where the usual recommendation is to use density-based approaches to detect shape outliers; see, for instance, Sun and Genton (2011).

Table 2.1: Percentages $\widehat{p}_c$ and $\widehat{p}_f$ for model (2.7).

	$k = 10$		$k = 15$		$k = 20$		$k = 25$	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
Depth MD								
FBox	20.67	1.88	50.33	1.87	72.00	1.86	88.00	1.85
DBT-indep	33.33	0.79	59.67	0.53	83.33	0.17	93.33	0.09
SmBoD	34.00	1.38	63.00	1.35	87.00	1.21	97.33	1.22
MBBo	38.00	2.31	65.00	2.40	90.33	2.52	98.00	2.64
StBo	36.00	2.10	65.67	2.29	90.00	2.38	97.67	2.51
SmBoR	33.33	1.11	60.00	1.11	83.67	1.55	95.00	2.24
Depth BD								
FBox	14.67	0.48	43.67	0.47	68.33	0.46	87.67	0.46
DBT-indep	38.33	10.87	62.00	13.82	79.00	16.26	86.67	17.72
SmBoD	36.67	10.43	45.00	10.30	50.33	10.38	49.67	10.09
MBBo	66.33	19.19	88.33	19.16	95.33	18.57	95.67	18.18
StBo	66.00	19.00	88.33	19.09	95.33	18.49	96.67	18.13
SmBoR	51.00	10.94	80.00	13.61	91.67	11.60	96.33	7.92
Depth MBD								
FBox	22.00	1.26	52.33	1.24	75.00	1.24	92.33	1.24
DBT-indep	35.00	2.34	58.67	1.94	81.33	1.98	87.67	1.76
SmBoD	35.33	2.46	60.33	2.19	83.33	2.28	90.33	2.23
MBBo	42.67	5.03	68.33	4.70	91.33	5.01	99.00	5.39
StBo	43.67	5.08	69.00	5.01	91.67	5.08	98.67	5.48
SmBoR	35.33	2.60	60.33	2.33	87.00	2.88	95.33	3.29

In the case of partial outliers in Table 2.3, the best depth is, again, the h-modal (actually, partial outliers can be seen as a combination of magnitude and shape outliers; so, in our particular setting, the selection of the h-modal depth suggests major importance of the shape component). As expected, when the modified band depth (MBD) is used, observed power is less than the corresponding power associated to the band depth (note that, by construction, depths obtained from the MBD are generally larger than those obtained from the BD). Interestingly, the performance of the empirical power seems alike for both of the block-bootstrap-based procedures (MBBo and StBo). However, the performance of the autoregression-based method (SmBoR) differs on the type of outliers included in the uncontaminated model; this could confirm that, contrary to SmBoR, neither MBBo nor StBo are model-based procedures. Moreover, a misspecification in the autoregression adds bias to the residuals; the effect of this bias seems to depend on the type

of outliers in the final data.

Second, let us focus on the empirical nominal level, $\widehat{p}_f$; that is, the percentage of falsely detected outliers. The study shows that the depth greatly influences the performance of each method. Good results are observed when the h-modal depth is used (values of $\widehat{p}_f$ close to the nominal level 1%). Nevertheless, the nominal level of the proposed tests deteriorates when one of the other two depths are considered. As in the case of the power, the empirical nominal levels of the block-bootstrap and the autoregression-based methods differ. Third, the study allows to suggest the use of the h-modal depth to achieve a trade-off between type I and II errors (that is, values of $\widehat{p}_f$ close to 1% together with high $\widehat{p}_c$ values). This recommendation runs in line with the suggestion in Febrero et al. (2008), established within the context of independent data.

Table 2.2: Percentages $\widehat{p}_c$ and $\widehat{p}_f$ for model (2.8).

	$k=4$		$k=5$		$k=6$		$k=7$	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
Depth MD								
FBox	12.33	1.85	15.00	1.85	17.00	1.85	24.67	1.85
DBT-indep	34.33	0.93	74.67	0.66	92.67	0.24	98.33	0.12
SmBoD	34.00	1.36	71.00	0.74	89.33	0.32	97.67	0.20
MBBo	40.33	1.95	70.00	0.83	91.67	0.45	99.33	0.44
StBo	37.67	1.87	68.33	0.77	90.67	0.40	99.00	0.41
SmBoR	51.67	2.36	82.67	1.60	93.33	0.97	97.33	0.51
Depth BD								
FBox	3.67	0.49	5.67	0.50	7.67	0.51	11.67	0.50
DBT-indep	4.33	1.90	4.33	1.43	1.67	0.62	1.00	0.36
SmBoD	4.33	1.94	3.67	0.88	2.33	0.66	0.00	0.00
MBBo	49.00	15.28	56.67	14.11	61.67	12.49	59.67	10.07
StBo	47.33	14.54	51.67	13.00	55.67	11.28	52.67	8.87
SmBoR	19.33	2.14	25.00	2.15	30.67	2.13	38.67	2.13
Depth MBD								
FBox	6.33	1.27	10.00	1.29	14.00	1.25	17.67	1.24
DBT-indep	0.33	2.98	0.33	3.04	0.33	3.14	0.33	3.18
SmBoD	0.33	2.81	0.33	2.68	0.33	2.59	0.00	2.50
MBBo	0.33	5.54	0.33	5.22	0.33	4.91	0.00	4.61
StBo	0.33	5.48	0.33	5.31	0.33	5.17	0.00	4.96
SmBoR	0.33	4.10	0.33	4.01	0.33	4.14	0.00	4.22

Besides, some remarks on the contamination size k and the behaviour of the tests in comparable situations are added. Of course, the power of the tests depends on k; so, low, medium and high values for k are considered to evaluate their effect. Power generally increases with k, and the proposed methods work well for medium and high k values when an appropriate depth is used. Even when k is low, although the power is also low, it is clearly greater when dependence is taken into account. It is worth noting that the methods work better for model (2.9) (partial magnitude outliers) than they do for model (2.7) (complete magnitude outliers); see the results corresponding to the h-modal depth, which achieves a trade-off between type I and II errors. This shows that our methods are sensible to partial modifications in the outliers.

Table 2.3: Percentages $\widehat{p}_c$ and $\widehat{p}_f$ for model (2.9).

	$k=10$		$k=15$		$k=20$		$k=25$	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
Depth MD								
FBox	20.33	1.84	48.67	1.85	71.33	1.85	87.33	1.85
DBT-indep	79.00	0.32	94.33	0.05	96.00	0.04	97.00	0.03
SmBoD	76.00	0.46	94.00	0.72	96.00	1.09	98.33	1.56
MBBo	77.67	0.82	95.00	1.38	96.33	1.99	99.00	2.53
StBo	76.33	0.71	94.33	1.24	96.33	1.85	99.00	2.41
SmBoR	81.00	0.88	94.00	1.31	96.33	1.89	99.00	2.43
Depth BD								
FBox	14.00	0.49	42.33	0.49	68.67	0.49	88.33	0.49
DBT-indep	1.67	0.51	0.00	0.00	0.00	0.00	0.00	0.00
SmBoD	1.33	0.62	1.00	0.20	1.00	0.20	0.00	0.00
MBBo	62.67	14.10	83.00	12.58	88.00	11.39	90.33	14.12
StBo	61.67	13.58	77.67	11.89	84.33	10.90	91.67	14.17
SmBoR	35.33	2.12	59.67	2.52	79.67	2.93	92.67	3.39
Depth MBD								
FBox	20.33	1.27	48.68	1.26	67.67	1.26	83.33	1.26
DBT-indep	3.33	3.48	4.67	4.16	5.33	5.00	8.67	5.68
SmBoD	2.33	2.58	2.33	2.77	3.33	2.70	4.00	2.16
MBBo	4.67	4.80	3.67	4.09	4.33	3.90	6.00	3.35
StBo	4.67	4.90	3.33	4.30	4.00	3.97	6.00	3.52
SmBoR	3.00	3.91	3.00	3.60	4.00	3.12	6.67	2.62

Regarding the comparison between DBT-indep and DBT using SmBoD Bootstrap, some differences among them are observed. Remember that neither of these models take dependence into account and that they differ on the first filtrate to obtain robust estimation of the cutoff. The biggest improvement of the DBT-SmBoD with respect to DBT-indep is obtained with model (2.7), with magnitude outliers, with MD or MBD depth. With shape outliers (model (2.8)), $\widehat{p}_c$ is slightly worse and it varies with partial magnitude outliers (model (2.9)). Note that BD and MBD work very poor with this partial magnitude outliers. That is, even if SmBoD does not improve DBT-indep in every situation, it is a good choice in most of them. Obviously, as was mentioned before, any other of the approaches taking dependence into account must be selected instead when dealing with functional time series.

Finally, the recommendation when dealing with unknown kinds of outliers, is to select DBT method with h-modal depth and MBB Bootstrap. This combination has shown the best global performance regarding the kind of outlier and its size, considering both $\widehat{p}_c$ and $\widehat{p}_f$ percentages.

Results obtained in this section are published in Raña, Aneiros and Vilar (2015).

2.4 Outlier detection in Functional Time Series using Functional Principal Components Analysis

Based on the use of functional principal components, two methods to detect outliers in functional time series were developed. Both make use of the robust FPCA proposed in Hyndman and Ullah (2007), which has been described in Section 1.3 as it is also the starting point for the ISE method to detect functional outliers.

The two proposals are described separately in Subsection 2.4.1 and 2.4.2. Thereafter, a new simulation study is presented in Subsection 2.4.3. These new simulations are intended not only to deeply analyse the behaviour of the two new proposals based on FPCA, but also to compare them to the DBT methods developed in the previous section. Thus, it pretends to extend and enlarge the simulations carried out in Subsection 2.3.1, to give a more general vision of the involved methods.

2.4.1 Projections-based method

This section presents the projections-based method (PB), which is a procedure to detect outliers in functional time series, based on robust FPCA proposed in Hyndman and Ullah (2007). This proposal detects outliers on the first K robust principal component scores, by applying techniques to identify outliers in scalar time series, and then map the detected outliers into the functional space.

Specifically, this method based on projections proposes to detect outliers in functional time series with the following algorithm:

- Step 1. Perform the robust FPCA proposed in Hyndman and Ullah (2007) and construct the series of coefficients $\{\widehat{\beta}_{1,k}, \ldots, \widehat{\beta}_{n,k}\}$, $k = 1, \ldots, K$ (see (2.4)).
- Step 2. Use univariate robust ARIMA models (for details, see Section 11.2 in Cryer and Chan 2008) to detect outliers in each series $\{\widehat{\beta}_{1,k}, \ldots, \widehat{\beta}_{n,k}\}$, $k = 1, \ldots, K$.
- Step 3. Establish the set of outliers as $\mathcal{O} = \{\chi_i : i \in \mathcal{I}\}$, where $\mathcal{I} = \{i : \{\widehat{\beta}_{i,1}, \ldots, \widehat{\beta}_{i,K}\}$ contains some of the outliers detected in Step 2$\}$.

Note that the key points in this method are the use of robust FPCA together with procedures to detect outliers in univariate time series. Given that the estimated functional principal components $\widehat{\phi}_k$ are not affected by the outliers, the corresponding projections $\widehat{\beta}_{i,k}$ reflect the main features of the datum χ_i. Thus, if a curve is an outlier, its projection on at least some of the directions of maximum variance (the first principal components) will also be an outlier. Note that, because principal component scores, $\widehat{\beta}_{i,k}$ and $\widehat{\beta}_{i,l}$, are uncorrelated for $k \neq l$, Hyndman and Ullah (2007) suggest that each univariate time series $\{\widehat{\beta}_{1,k}, \ldots, \widehat{\beta}_{n,k}\}$, $k = 1, \ldots, K$, can be studied independently.

2.4.2 Errors-based method

This section presents a new method based also in robust FPCA, as the PB method presented above. Unlike the previous method (PB), our second procedure based on FPCA takes the whole of each curve into account. Using techniques for robust forecasting in functional time series, it constructs a

non-contaminated version for each curve, which is compared with the corresponding original curve. A curve is considered an outlier if it is substantially different from its uncontaminated version.

Specifically, this method proposes to detect outliers in functional time series with the following algorithm:

- Step 1. Perform robust FPCA proposed in Hyndman and Ullah (2007) and construct the series of coefficients $\{\widehat{\beta}_{1,k}, \ldots, \widehat{\beta}_{n,k}\}$, $k = 1, \ldots, K$ (see (2.4)).

- Step 2. Fit univariate robust ARIMA models (for details, see Section 11.2 in Cryer and Chan 2008) for each series $\{\widehat{\beta}_{1,k}, \ldots, \widehat{\beta}_{n,k}\}$, $k = 1, \ldots, K$.

- Step 3. Obtain the fitted values $\{\widetilde{\widehat{\beta}}_{1,k}, \ldots, \widetilde{\widehat{\beta}}_{n,k}\}$, $k = 1, \ldots, K$, from the models constructed in Step 2.

- Step 4. Construct the fitted curves $\widehat{\chi}_i = \sum_{k=1}^{K} \widetilde{\widehat{\beta}}_{i,k} \widehat{\phi}_k$ and compute the L_2-norm of the corresponding prediction errors:

$$u_i = \sqrt{\int_a^b (\chi_i(t) - \widehat{\chi}_i(t))^2 dt}, \ i = 1, \ldots, n.$$

- Step 5. Establish the set of outliers as $\mathcal{O} = \{\chi_i : i \in \mathcal{J}\}$, where $\mathcal{J} = \{i : u_i > q_{0.75} + 1.32(q_{0.75} - q_{0.25})\}$ (q_p denotes the quantile of order p of $\{u_1, \ldots, u_n\}$).

As in the method based on projections, robust FPCA and robust ARIMA models play the main roles. Note that, because the forecasts obtained from the procedure explained in Cryer and Chan (2008) are not contaminated by the outliers, $\widehat{\chi}_i$ can be seen as the "expected value" of the functional time series at time i when no contamination is present. Thus, a "high value" for u_i suggests that χ_i is an outlier. To decide if u_i is high enough to correspond to an outlier, the rule given by the classical boxplot is used; that is, under normality, the probability of detecting no outliers is 0.993, when no outliers are actually present (note that the usual constant factor 1.5 was changed to 1.32 because low values are not considered outliers). Finally, let point out the main difference between our proposal and the proposal suggested in Hyndman and Ullah (2007) (χ_i is detected as outlier if $v_i \geqslant s + \lambda\sqrt{s}$; see (2.3)): our procedure takes the dependence among the sample into account

(see Step 3) to construct the coefficients associated to each functional data χ_i ($\widetilde{\widetilde{\beta}}_{i,k}$ in Step 4 above), while the method in Hyndman and Ullah (2007) does not do so (see $\widetilde{\beta}_{i,k}$ in (2.3)).

2.4.3 Simulation study

Simulated FTS

Three main models were constructed to generate functional time series. This is an extension of the simulations carried out in Subsection 2.3.1. Since the proposed methods here are easier to implement (because they do not have so many options as DBT regarding depth and bootstrap elections), we can analyse them in a deeper way on different dependence scenarios. The simulation models are the superposition of a deterministic signal and random noise. Noise in main Models 1, 2, and 3 was the superposition of a scalar AR(1) process and functional AR(1)-, MA(1)- and ARMA(1,1)-type processes, respectively. On the other hand, another main model (Main Model 0) was constructed in the same way, but considering independent noise instead of dependent one. Note that main Models 1, 2 and 3 are favourable to methods that take dependence in the sample into account, while Main Model 0 is favourable to methods designed for independent data. From each main model, two contaminated models were constructed by randomly adding either three magnitude outliers or three shape outliers.

More specifically, the following main models were considered:

- Main Model 0:

$$\zeta_i(t) = \cos(\pi t)(1-c) + a_i(t) \quad \text{if} \quad -n+1 \leq i \leq n.$$

- Main Model 1:

$$\zeta_i(t) = \begin{cases} \cos(\pi t) & \text{if} \quad i = -n+1 \\ \cos(\pi t)(1-c) + \rho\zeta_{i-1}(t) + a_i(t) + b_i & \text{if} \quad -n+1 < i \leq n. \end{cases}$$

- Main Model 2:

$$\zeta_i(t) = \cos(\pi t)(1-c) + \theta a_{i-1}(t) + a_i(t) + b_i \quad \text{if} \quad -n+1 \leq i \leq n.$$

- Main Model 3:

$$\zeta_i(t) = \begin{cases} \cos(\pi t) & \text{if} \quad i = -n+1 \\ \cos(\pi t)(1-c) + \rho\zeta_{i-1}(t) + \theta a_{i-1}(t) + a_i(t) + b_i & \text{if} \quad -n+1 < i \leq n. \end{cases}$$

In the processes above it is denoted $a_i(t) = X_i \sin(\pi t)$ with X_i being i.i.d. Gaussian variables with mean 0 and standard deviation 0.3, while $\{b_i\}$ is a scalar Gaussian AR(1) process with correlation coefficient $d = 0.8$ and standard deviation $(1 - d^2)^{-1/2}$. $c = 0.8$ and $t \in [-0.5, 1.5]$ were considered.

Values ρ and θ manage the dependence strength in the functional time series. Two options were considered, one with low dependence (LD, $\rho = 0.5$ and $\theta = -0.5$) and other with high dependence (HD, $\rho = 0.8$ and $\theta = 0.8$).

Then, given the series of each main model, ζ_i, several methods were applied to detect outliers on the following contaminated models:

- Contaminated model with magnitude outliers:

$$\chi_i(t) = \zeta_i(t) + k1_{\{i \in \{I_j\}\}},\ 1 \leq i \leq n.$$

- Contaminated model with shape outliers:

$$\chi_i(t) = \zeta_i(t) + k\cos(3\pi t)1_{\{i \in \{I_j\}\}},\ 1 \leq i \leq n.$$

Note that k is a contamination size while $1_{\{.\}}$ and I_j denote the indicator function and i.i.d. random variables with discrete uniform distribution on $\{1, \ldots, n\}$, respectively. The curves χ_i were discretized on a grid $\{t_j\}$ of 30 equispaced points in $[-0.5, 1.5]$. Comparing these simulations with the ones carried out in Subsection 2.3.1, only the two principal kinds of outliers are considered: magnitude and shape outliers. Partial-magnitude outliers were removed from the simulations for the sake of the brevity, as they behave as a mix between magnitude and shape outliers.

In the simulation process, the generated curves correspond to the double of the sample size n. That is, the curves $\{\zeta_i(t)\}, -n + 1 \leq i \leq n$ were simulated, but only the last half of the curves were used, $\{\zeta_i(t) : 1 \leq i \leq n\}$, for the contaminated models. The first n realizations are not used in order to avoid the impact of the initial values. The number of outliers introduced in the models was $j = 0.02n$ (that is, 2% of the curves). The value of k was 0.75 for contaminated Model 0, in which dependence does not affect, and 5 for contaminated Models 1, 2 and 3. It is worth noting here that the contamination size, k, considered in this study is low compared with other existing simulation studies (see, e.g. Sun and Genton 2011) and also with the simulations in Subsection 2.3.1.

Figure 2.6 shows curves simulated from these four contaminated models. The first row corresponds to Model 0 (no dependence), and the other three rows to Models 1, 2 and 3 (functional time series), respectively. The last three models are shown for the HD case. One can see in the figure the difference between the data simulated from Model 0 and from Models 1, 2 and 3: in the case of functional time series, outliers are almost always hidden within the rest of the curves.

The following four procedures were applied on each generated sample, which have been already described in Section 2.2, in addition to the proposed projections-based (PB) and errors-based (EB) methods.

- Functional highest density region boxplot (HDR). The routine *fboxplot*, available in the R package *rainbow* was used with $\alpha_0 = 0.01$.

- Integrated squared error (ISE). The routine *foutliers*, available in the R package *rainbow* was used, where λ and K were chosen following the suggestion given in Hyndman and Ullah (2007): $\lambda = 3.29$ and K being the value minimizing the ISFE.

- Depth-based trimming for independent data (DBT-indep), implemented in the routine *Outliers.fdata*, available in the R package *fda.usc*, with $\alpha_1 = \alpha_2 = 0.01$, $\gamma = 0.05$ and $B = 200$, while the functional depth was the h-modal depth

- DBT for dependent data, with functional h-modal depth and Moving Blocks Bootstrap (MD-MBB).

Note that the methods HDR, ISE and DBT-indep are designed to detect outliers in samples of independent curves, even if they were also applied to functional time series. However, DBT MD-MBB, PB and EB are specifically designed to deal with the problem of outlier detection in the context of functional time series. Along this simulation study, the performance of the cited methods was compared in situations of both independent and dependent data.

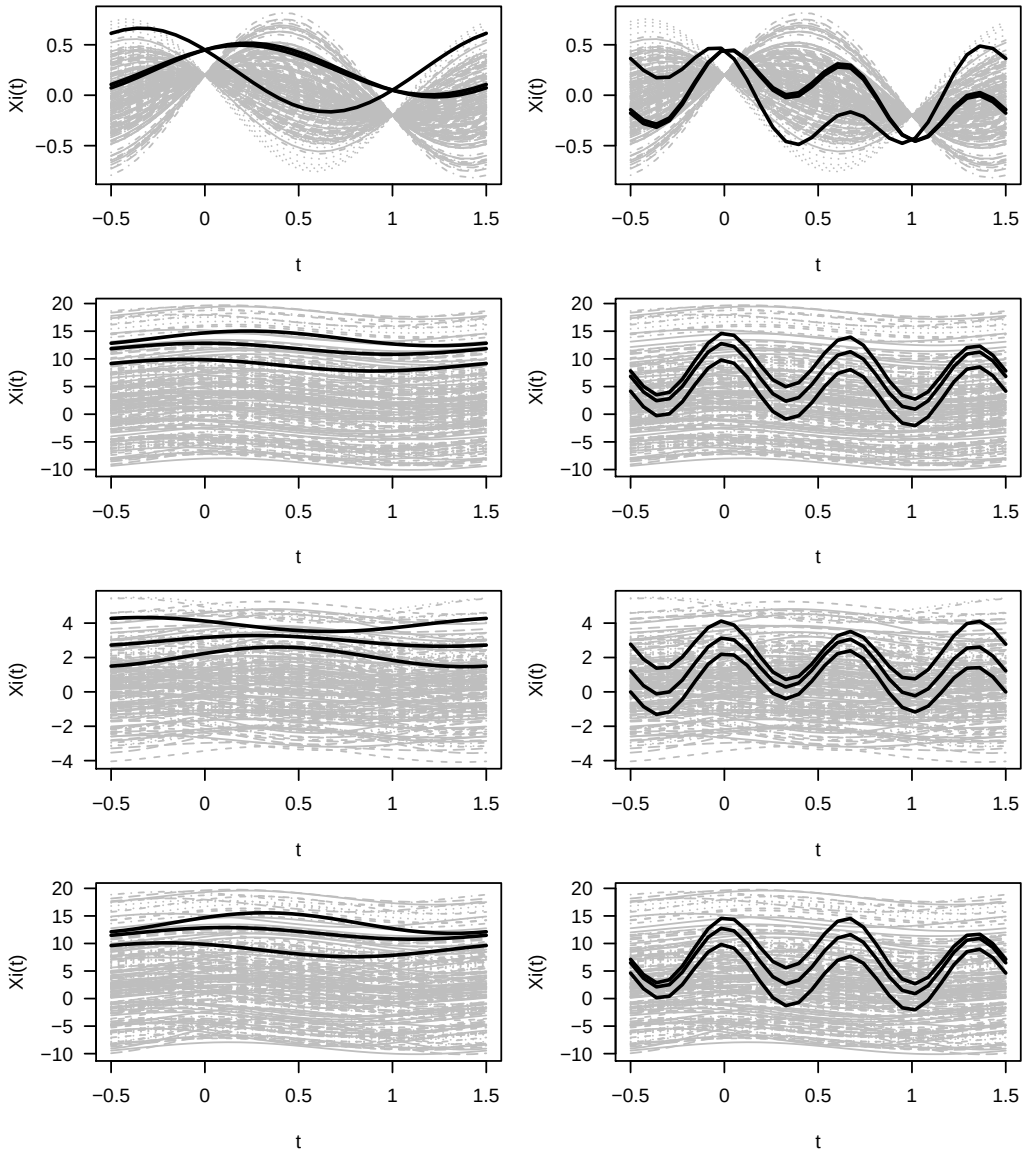


Figure 2.6: Left panels: from top to bottom, curves $(\chi_i(t))$ generated from contaminated Models 0, 1, 2 and 3, respectively (the black curves are magnitude outliers). Right panels: from top to bottom, curves $(\chi_i(t))$ generated from contaminated Models 0, 1, 2 and 3, respectively (the black curves are shape outliers).

Tuning parameter

As is common to all FDA procedures using FPCA, the proposed methods depend on the quantity of principal components considered, K. In practice, the value of K must be specified. Hyndman and Ullah (2007) suggest choosing K to minimize the mean integrated squared forecast error, while Hyndman and Booth (2008) find that the forecasts are insensitive to the choice of K, provided K is large enough. Sensitivity studies were made for the values of K in our methods, using the dependent simulated data considered in Models 1 and 2.

On the one hand, the findings agree with the general suggestion given in Hyndman and Booth (2008): to consider a larger than necessary value K (for instance, a value explaining at least 98% of the variability). On the other hand, to detect "shape outliers" (that arise when they are within the range of the rest of the data but differ from them in shape; see Hyndman and Shang 2010) by means of the method based on projections, the recommendation is to select a value K even higher (for instance, explaining at least the 99.9% of variability). To justify this very high value it is argued that (i) the method based on projections only uses scores (and not the whole curve), (ii) the first scores inform on the possible presence of "magnitude outliers" and (iii) the higher order scores inform on the possible presence of "shape outliers". The number of principal components for the ISE method is selected in the same way, allowing also to take a higher number of principal components when detecting "shape outliers" in dependence models.

Specifically, $K = 1$ was chosen for magnitude outliers (for both methods PB and EB). In the case of shape outliers, it was chosen $K = 3$ for the PB method and $K = 1$ for the EB. This election explains more than 98% of the variability (in some cases, even with only the first component, it explains around 99.5%), increasing until 99.9% when we use PB method to detect shape outliers. This choice agrees with the guidelines about the requirement of more components when dealing with shape outlier detection and the PB method. In the case of Model 0, due to the simplest performance of the data, it is enough to take $K = 1$ for the EB and $K = 2$ for the PB method and both kind of outliers. The significance level used to detect scalar outliers in the PB method (Step 2) was $\alpha_3 = 0.01$. In the case of the norm to be used in Step 4 of the EB procedure, both the L_1-norm and the L_2-norm were considered. Because similar results were obtained, it is only shown the corresponding ones to the L_2-norm.

The results

$M = 500$ simulations were run for each model. The percentage of correctly identified outliers p_c (100 times the number of correctly identified outliers over the number of outliers in the sample, or sensitivity) and the percentage of false positives p_f (100 times the number of wrongly identified outliers over the number of non-outlying curves in the sample, or false detection percentage) were computed for each simulation and for each method considered.

The first simulation study employs $n = 200$ and the results are reported in Tables 2.4, 2.5 and 2.6. These tables show the mean and standard deviation of the values of both p_c and p_f obtained from the two proposed procedures (PB and EB) and the other four considered methods (HDR, ISE, DBT-indep and DBT MD-MBB) when they are applied to the different contaminated models. In Table 2.4 Model 0 is considered (independent data), which is contaminated with magnitude or shape outliers. In Tables 2.5 and 2.6, Models 1, 2 and 3 (dependent data) and the two cases of dependence (low and high dependence) are considered (see Table 2.5 for contamination with magnitude outliers and Table 2.6 for the case of shape outliers).

Table 2.4: Mean and standard deviation (in parentheses) of the percentage of correctly and falsely identified outliers in Model 0 contaminated with magnitude or shape outliers.

Model 0				
	Magnitude outliers		Shape outliers	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	40.60 (14.21)	0.19 (0.29)	40.50 (14.24)	0.19 (0.29)
ISE	100.00 (0.00)	0.00 (0.02)	100.00 (0.00)	0.00 (0.00)
DBT-indep	87.00 (26.12)	0.64 (0.48)	84.15 (27.46)	0.61 (0.47)
DBT MD-MBB	99.80 (4.47)	2.83 (1.44)	99.80 (4.47)	2.83 (1.44)
PB	95.40 (10.44)	0.02 (0.10)	95.15 (10.98)	0.02 (0.10)
EB	100.00 (0.00)	2.14 (1.23)	95.75 (15.61)	2.10 (1.22)

Table 2.5: Mean and standard deviation (in parentheses) of the percentage of correctly and falsely identified outliers in Models 1-2-3 (with low or high dependence) contaminated with magnitude outliers.

Model 1				
	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	16.50 (15.91)	0.68 (0.32)	9.85 (14.04)	0.82 (0.29)
ISE	22.00 (22.01)	15.33 (2.69)	25.25 (22.74)	21.07 (3.65)
DBT-indep	26.30 (23.28)	1.19 (0.87)	10.55 (15.67)	1.14 (1.17)
DBT MD-MBB	30.45 (24.04)	2.07 (1.78)	13.10 (16.93)	2.38 (2.24)
PB	70.65 (35.95)	0.31 (0.45)	62.05 (38.65)	0.56 (0.59)
EB	88.55 (17.12)	3.71 (1.24)	84.10 (20.81)	4.07 (1.26)
Model 2				
	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	28.25 (16.92)	0.44 (0.35)	30.25 (17.08)	0.40 (0.35)
ISE	24.30 (23.73)	14.60 (2.51)	27.00 (23.39)	16.64 (2.79)
DBT-indep	66.75 (25.49)	0.81 (0.69)	67.60 (24.83)	0.76 (0.69)
DBT MD-MBB	73.15 (24.58)	1.60 (1.17)	73.40 (24.00)	1.55 (1.24)
PB	67.60 (37.26)	0.07 (0.20)	68.40 (37.24)	0.07 (0.18)
EB	91.60 (14.83)	3.17 (1.29)	91.60 (15.08)	3.23 (1.28)
Model 3				
	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	14.70 (15.31)	0.72 (0.31)	10.75 (14.27)	0.80 (0.29)
ISE	19.60 (21.78)	13.31 (2.72)	30.65 (24.04)	28.96 (3.40)
DBT-indep	26.30 (23.55)	1.20 (0.86)	10.65 (15.44)	1.13 (1.15)
DBT MD-MBB	30.30 (24.35)	2.08 (1.67)	12.85 (16.78)	2.35 (2.03)
PB	69.30 (36.30)	0.33 (0.47)	60.95 (38.36)	0.61 (0.64)
EB	88.45 (16.76)	3.63 (1.24)	84.20 (20.28)	3.81 (1.17)

Table 2.6: Mean and standard deviation (in parentheses) of the percentage of correctly and falsely identified outliers in Models 1-2-3 (with low or high dependence) contaminated with shape outliers.

Model 1				
	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	16.15 (19.20)	0.69 (0.39)	13.90 (18.77)	0.74 (0.38)
ISE	100.00 (0.00)	14.58 (2.74)	100.00 (0.00)	20.15 (3.59)
DBT-indep	95.75 (16.47)	0.21 (0.41)	64.25 (38.55)	0.61 (0.91)
DBT MD-MBB	99.40 (6.69)	0.70 (1.16)	56.05 (37.45)	0.58 (1.06)
PB	95.20 (10.71)	0.04 (0.17)	95.00 (11.19)	0.04 (0.24)
EB	100.00 (0.00)	2.58 (1.26)	100.00 (0.00)	2.59 (1.29)
Model 2				
	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	10.75 (15.20)	0.80 (0.31)	9.10 (14.23)	0.83 (0.23)
ISE	100.00 (0.00)	13.85 (2.51)	100.00 (0.00)	15.95 (2.67)
DBT-indep	96.40 (15.57)	0.38 (0.45)	98.30 (11.06)	0.33 (0.42)
DBT MD-MBB	100.00 (0.00)	1.81 (1.36)	100.00 (0.00)	1.67 (1.35)
PB	95.20 (10.71)	0.03 (0.16)	95.15 (10.63)	0.04 (0.16)
EB	100.00 (0.00)	2.61 (1.38)	100.00 (0.00)	2.64 (1.34)
Model 3				
	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	17.35 (19.26)	0.67 (0.39)	4.65 (11.40)	0.93 (0.23)
ISE	100.00 (0.00)	12.52 (2.62)	100.00 (0.00)	27.86 (3.23)
DBT-indep	94.40 (19.15)	0.22 (0.42)	49.60 (38.54)	0.79 (1.04)
DBT MD-MBB	99.40 (6.69)	0.47 (1.23)	41.80 (34.50)	0.94 (1.32)
PB	95.05 (10.82)	0.05 (0.20)	94.95 (11.45)	0.14 (0.31)
EB	100.00 (0.00)	2.51 (1.26)	100.00 (0.00)	2.06 (1.19)

Several conclusions can be drawn from these results. First of all, one may look at Contaminated Model 0 in Table 2.4, which considers independent data. Under that situation, ISE method gets the best result for both kinds of outliers. On the contrary, HDR presents poor results with the lowest sensitivity, but also its false detection rate is low. Looking at the pair of DBT-indep and DBT MD-MBB method, one can see an improvement with the second option, even if dependence is not affecting this data. p_c is much better for the DBT MD-MBB method, compared to the DBT-indep, but also p_f is higher.

Note that DBT MD-MBB not only adapts DBT-indep to work with functional time series (by taking dependence into account), but also improves the method itself by changing some other aspects. This is why one can see different results even when they are applied to independent data. Our both proposals, PB and EB, are very competitive in this situation, even compared to methods designed to work with independent data. They maintain high and low values for p_c and p_f, respectively. Their sensitivity is greater than 95% and there is no big difference between magnitude and shape outliers. One can see that PB detects less outliers than EB but also its false detection rate is lower.

Now, we focus on the simulated models that include dependence structure; that is, contaminated Models 1, 2 and 3. The role of this analysis is two-fold: to illustrate the performance of the two proposed procedures and to show the need to take into account the dependence in the functional time series. Restrict first to the magnitude outliers under both situations of low dependence (LD) and high dependence (HD), which results are given in Table 2.5. In general not major differences are observed in the behaviour of the proposed methods (PB and EB) when the dependence scenario changes (LD or HD), and one can see that the best results are achieved by the methods that take into account dependence (DBT MD-MBB, PB and EB). Results are analysed below in a deeper way.

HDR and ISE methods lose their effectiveness in detecting outliers when dealing with dependent data. It is highlighted the large p_f (around 20%) of the ISE method, indicating a high volatility in its behaviour. Looking now at the pair DBT-indep and DBT MD-MBB methods (remember that DBT MD-MBB adapts DBT-indep to work with functional time series), it is true that DBT MD-MBB gets always higher p_c, which clearly indicates that taking dependence in the data into account is outstanding. Both methods are also better than HDR and ISE in most of the cases. Despite of getting worse

p_c than ISE when dealing with Models 1 and 3 under high dependence, they get significantly lower p_f. Both methods (DBT-indep and DBT MD-MBB) also show a sharp difference between dependence scenarios for Models 1 and 3, in which the outlier detection rate decreases as the dependence structure becomes more relevant.

All the methods analysed above are overcome by our two proposals PB and EB. Both options achieve high sensitivity, greater than DBT MD-MBB (except for Model 2 in which DBT MD-MBB overcomes PB) and far away from the other methods considered that not take into account dependence. PB method holds lower sensitivity than EB, but also lower false detection rate. To obtain a trade-off between high sensitivity and low false detection rate, in general, the proposed EB seems to be a good choice for magnitude outlier detection under the considered dependence scenarios.

Table 2.6 shows the results when the models are contaminated with shape outliers. HDR still performs very similar to the magnitude outliers case, however ISE method shows an improvement by detecting all the shape outliers (at the expense of a large false detection rate). DBT-indep and DBT MD-MBB behave also similarly to the magnitude outliers case, with a remarkable difference in the levels of p_c. They achieve now very high sensitivity with low dependence (around $95-100\%$) but under high dependence they provide low values, around $40-60\%$ for Models 1 and 3.

The proposed methods PB and EB show high sensitivity (95% and 100%, respectively) and low false detection rate (0.05% and 2.5%), being very stable for the three simulated models. As in Table 2 for magnitude outliers, also with shape outliers there are no major differences between both dependence scenarios (LD and HD). In summary, even if both methods obtain very good results for shape outlier detection under dependence, EB seems to be a better choice due its great success detecting all the outliers.

A second simulation study is developed in order to study the influence of the sample size (n) over the analysed methods for outlier detection. In this case we restrict ourselves to Models 1, 2 and 3 (simulated functional time series) contaminated with magnitude outliers.

Table 2.7: Mean of the percentage of correctly and falsely identified outliers in Models 1-2-3, with high dependence, contaminated with magnitude outliers and for $n = 100, 200, 300$ and 400.

Model 1								
	n=100		n=200		n=300		n=400	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	12.30	0.77	9.85	0.82	7.43	0.87	7.28	0.87
ISE	26.10	22.44	25.25	21.07	22.57	21.10	23.13	20.94
DBT-indep	12.10	0.34	10.55	1.14	9.23	1.60	9.63	1.86
DBT MD-MBB	12.80	1.13	13.10	2.38	11.80	2.99	12.03	2.98
PB	29.20	0.56	62.05	0.56	80.23	0.42	87.90	0.28
EB	58.80	4.52	84.10	4.07	90.83	3.99	93.60	4.00
Model 2								
	n=100		n=200		n=300		n=400	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	34.20	0.33	30.25	0.40	28.57	0.44	26.50	0.48
ISE	32.60	17.10	27.00	16.64	22.50	16.51	21.50	16.45
DBT-indep	63.70	0.47	67.60	0.76	68.63	0.90	68.87	1.01
DBT MD-MBB	73.60	1.66	73.40	1.55	72.30	1.52	72.08	1.54
PB	34.50	0.10	68.40	0.07	84.93	0.04	89.43	0.04
EB	73.70	3.60	91.60	3.23	95.70	3.35	96.25	3.27
Model 3								
	n=100		n=200		n=300		n=400	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	15.80	0.70	10.75	0.80	8.70	0.84	8.40	0.85
ISE	31.90	29.08	30.65	28.96	30.10	28.74	29.80	28.84
DBT-indep	12.00	0.32	10.65	1.13	9.37	1.61	9.58	1.85
DBT MD-MBB	12.70	1.18	12.85	2.35	11.17	2.78	11.40	2.77
PB	29.80	0.60	60.95	0.61	81.10	0.40	87.00	0.29
EB	54.90	3.68	84.20	3.81	90.67	3.78	93.18	3.81

Table 2.7 shows the mean of the percentage of correctly and falsely identified outliers (p_c and p_f, respectively) when the sample size varies within the values $n = 100, 200, 300$ and 400. These results are obtained under the scenario of high dependence (HD) and the number of outliers introduced in each sample follows the same rule as the previous results (including $j = 0.02n$ outliers; that is, 2% of the curves).

Results given by the two proposed methods (PB and EB) in Table 2.7 clearly overcome the rest of the methods included in the comparison (HDR, ISE, DBT-indep and DBT MD-MBB) in almost all the situations (except when one considers $n = 100$ in Model 2). That is, for the three contaminated models and the different values of the sample size n (except the combination Model 2, $n = 100$), PB and EB get the best performance. On the one hand, HDR, ISE, DBT-indep and DBT MD-MBB show poor results with very low p_c and also, in the case of ISE method, very high false detection rate. DBT MD-MBB gets always better results than DBT-indep, showing again the importance of taking dependence in the data into account. Actually, both DBT-indep and DBT MD-MBB are very competitive for the Model 2, specially with the lowest sample size $n = 100$, but they are overcome by PB and EB as n increases.

On the other hand, also HDR, ISE, DBT-indep and DBT MD-MBB remain stable when the sample size varies. Indeed, one can see a slight decrease in the p_c and increase in p_f as long as the sample size n increases. On the opposite, for the three contaminated models, the proposed methods PB and EB clearly improve the sensitivity (p_c) while the false detection rate (p_f) decreases slightly as n increases. The reason for this is that PB and EB methods are based on fitting univariate time series (of the coefficients given by FPCA) as a previous step to outlier detection. Therefore, by increasing the sample size, n, the fit of the univariate time series is improved and, accordingly, also the outlier detection with PB and EB methods improves.

Results obtained in this section have been published in Vilar, Raña and Aneiros (2016).

2.5 Applications in the electricity market

It is well known that the presence of outliers in a dataset affects the accuracy of forecasts obtained from regression models, and electrical data is not an exception. Thus, outlier detection represents a first step prior to any type of modeling.

In the next two sections, the proposed methods for detecting outliers in functional time series are applied on the daily curves of electricity demand and price, from the dataset described in Section 1.4. It is important to remember that in the Spanish Electricity Market, renewable energy sources (as wind power among others) enter in the pool without cost. This causes that, in specific days with high production of wind power, all the demand is covered by this energy source and the price decreases even until zero.

Note that, to not overextend the results, not all the methods/options mentioned in the previous sections are shown. The comparison is then reduced to the DBT using the combinations MD-MBBo and MD-SmBoR, together with the two proposals based on FPCA: PB and EB. In order to compare their results with the methods for independent data, also DBT-indep and HDR are considered. It allows to see the difference of taking into account the dependence in the particular case of our electrical dataset.

2.5.1 Case study: electricity demand

The application focus first in outlier detection in time series of electricity demand curves. Data collect hourly electricity demand in the Spanish mainland Electricity Market on Mondays, ..., Fridays in the year 2012. These hourly data present a trend. Thus, by subtracting the trend (estimated by means of kernel regression) the corresponding detrended hourly series is obtained. The functional dataset under analysis is composed of the $n = 261$ daily demand curves obtained from this detrended hourly series, measured in Megawatt-hour (MWh). The number of functional principal components considered was $K = 9$. These K principal components explained, at least, 98% of the variance.

The functional time series and the corresponding daily curves are shown in Figure 2.7. Higher demands are observed in the interval 10:00h–22:00h while lower ones correspond to the interval 3:00h–5:00h.

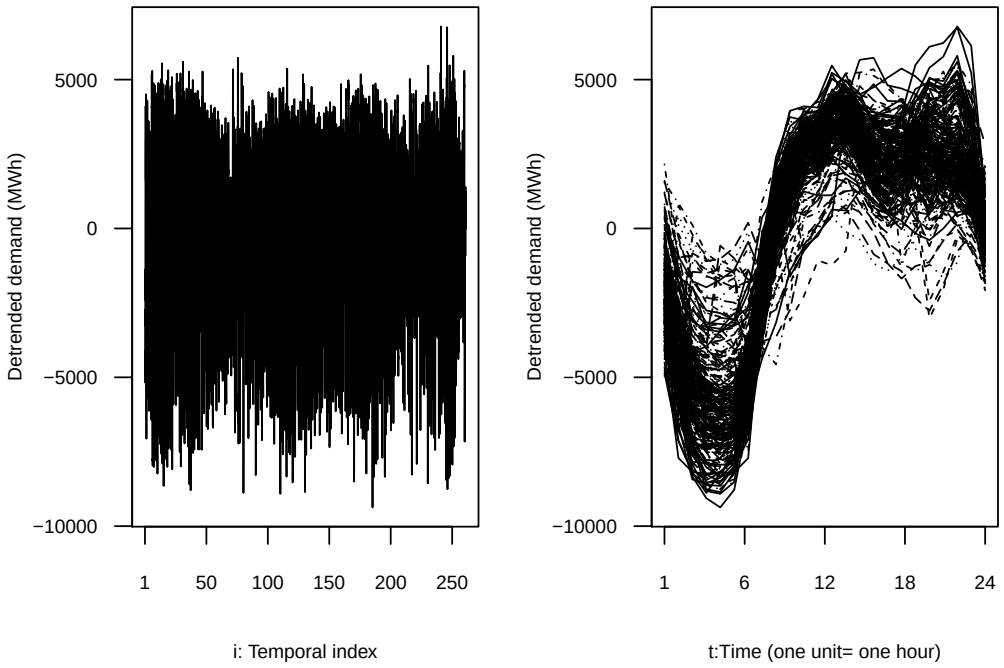


Figure 2.7: Left panel: time series of electricity demand. Right panel: daily electricity demand curves.

The outliers identified from the proposed procedures are listed in Table 2.8. Possible causes for most of these abnormal curves can be found. For example, April 16 and 19, November 1 and 2 and also December 24 correspond to days with zero price hours. During some hours in these days, the overproduction of wind power decreased the electricity price fixed by daily market. This have to do with the different taxations of this "green energies" because, as the wind power production increases, the electricity price decreases. As a result, if the wind power production covers and abnormally high percentage of the electricity demand, the price can drop even until zero during a period of time (this being the case of the cited days). We find also as outliers some previous or posterior days to these "zero price days", such as April 24 and 26, which are also affected by the disturbance in the price.

Some of the outliers correspond to nonworking days in which people usually behave in a different way than the rest of the regular days (simply because most of the economical and industrial activities stop during these days), affecting the electrical consumption and, as a consequence, also the demand. This is the case of May 1 (Labour Day), August 15 (Assumption Day), October 12 (National holiday in Spain), November 1 (All Saints Day), December

6 (Constitution Day in Spain) and 25 (Christmas). December 24 (Christmas Eve) is also a special day, even if it is not officially nonworking day. Friday, November 2, besides being a zero-price day, is situated also in the middle of a long weekend caused by All Saints Day, in which a lot of people took some holidays. Finally, November 14 was a Strike Day in Spain, which clearly affects electrical consumption, as it can be considered in some sense as a nonworking day.

Table 2.8: Outliers in electricity demand.

	DBT					
Method	indep	MD-MBBo	MD-SmBoR	HDR	PB	EB
January 6	X	X	X	X		
February 13				X		
” 14					X	
March 5				X		
April 16					X	X
” 19				X	X	X
” 24				X		X
” 26					X	
May 1	X	X	X	X	X	X
” 4					X	
August 15	X	X	X	X	X	X
October 12	X	X	X			X
November 1			X			X
” 2			X	X		X
” 7			X	X		X
” 8				X		
” 14			X		X	X
December 3				X	X	
” 6			X		X	X
” 10				X		X
” 21				X		X
” 24					X	
” 25	X		X			X
” 27				X		
” 28						X
Num. outliers	5	4	10	14	11	15

It is easy to see the big difference in the outliers detected by the considered methods. For example, DBT-indep and MD-MBBo work in a very similar way in this case, contrary to the MD-SmBoR that detects more outliers. EB detects more outliers than any other method, being some of them coincident with PB and also HDR.

Finally, it is worth pointing out that electricity demand curves observed at days April 16 (price zero), November 1 (holiday and price zero), November 14 (strike) and December 6 (holiday) are detected as outliers by, at least, two of our proposed methods, but none of these curves is identified as an outlier by neither the HDR nor DBT-indep procedures (remember that neither HDR nor DBT-indep take dependence in the data into account). Actually, as can be seen in Figure 2.8, these four curves have features that can, to say the least, be considered suspicious: demand curves observed at April 16 and November 1 take high values throughout the first hours (possibly because the electricity price at some hours was zero); demand curve corresponding to November 14 (strike day) maintains low values from 7:00h, this being the typical behaviour of demand curves corresponding to nonworking days; December 6 is a nonworking day.

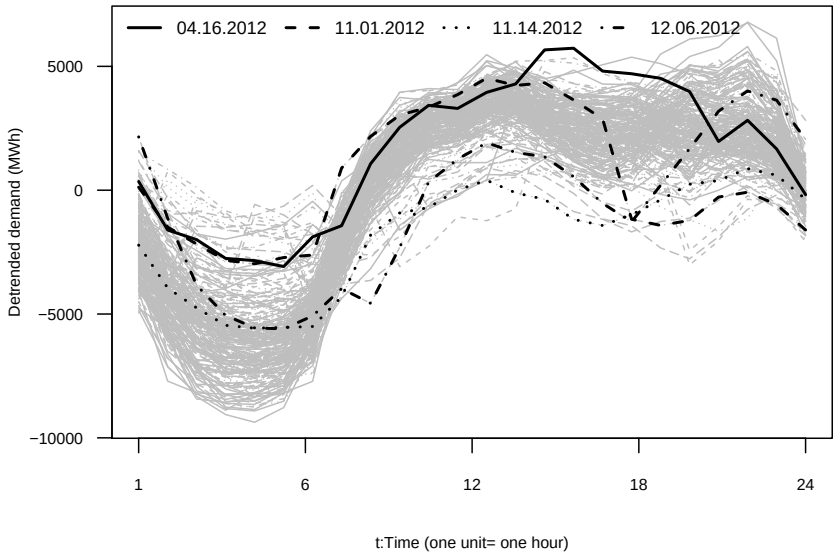


Figure 2.8: Outliers detected in the demand curves with some of the proposed procedures, but not detected by methods designed for independent data.

2.5.2 Case study: electricity price

A similar study is conducted in this section, applied to electricity price. Prices were available for the same period as demand: weekdays in 2012. Unlike the previous case, there was no trend in the data. The quantity of functional principal components considered was $K = 8$, which explained, at least, 98% of the variance.

Figure 2.9 displays the functional time series of electricity prices and the associated daily curves. Note that periods of low and high prices roughly correspond with periods of low and high demand, respectively.

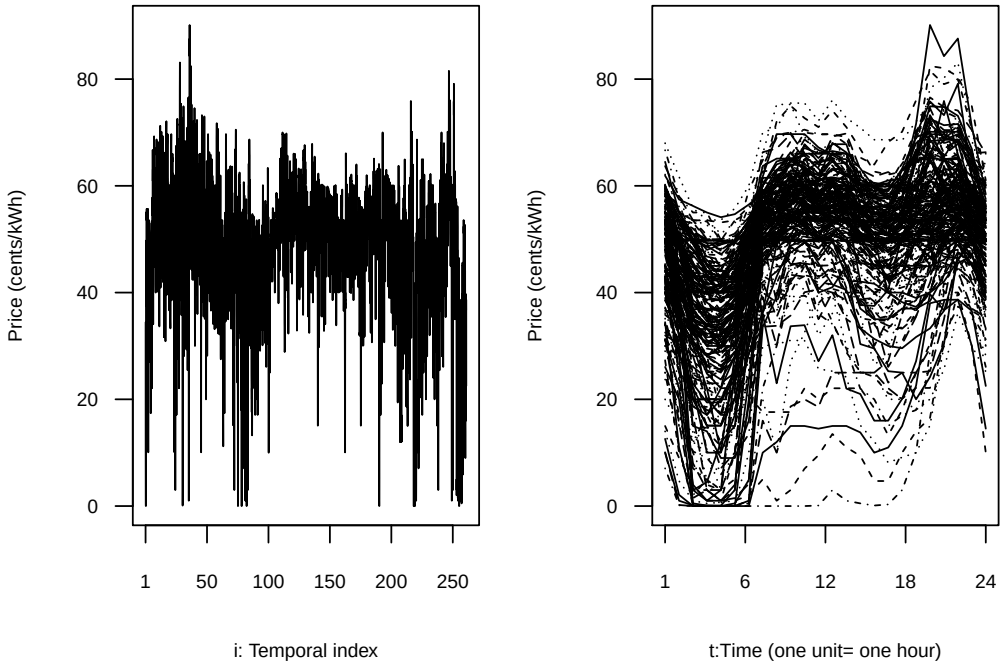


Figure 2.9: Left panel: time series of electricity price. Right panel: daily electricity price curves.

The outliers identified by the proposed procedures are listed in Table 2.9. Note that a total of 29 observations are detected as abnormal curves, 20 of them were detected by the proposals based on FPCA (13 from the PB method and 15 from the EB method). In addition, 10 of the 29 days corresponding to such outliers were days when demand curves were also identified as outliers (compare Tables 2.8 and 2.9).

Table 2.9: Outliers in electricity price.

	DBT					
Method	indep	MD-MBBo	MD-SmBoR	HDR	PB	EB
February 3				X		
" 13			X		X	
" 20		X	X	X		
" 21						X
April 6				X	X	
" 10		X	X		X	X
" 11					X	X
" 18		X	X	X		
" 19	X	X	X	X	X	
" 25	X	X	X	X	X	X
May 1	X	X	X		X	X
" 4			X			
" 8						X
June 11					X	
August 15		X	X		X	X
" 16						X
September 24			X	X		X
October 1						X
" 24						X
November 1	X	X	X		X	X
" 2	X	X	X	X	X	
" 6				X		
December 12				X		
" 14						X
" 21			X	X		
" 24	X	X	X	X	X	X
" 25	X	X	X		X	X
" 26		X	X	X		
" 31	X	X	X	X		
Num. outliers	8	13	17	14	13	15

Following the classical rules of any kind of market, it is usual that demand and price are highly interconnected, this being also the case of electricity markets and the reason why some of the outlying curves in demand are repeated as outliers for electricity price. As in the previous application, one can argue causes for most of the abnormal curves of electricity price, being some of them already cited in the study of outliers in electricity demand. Some of

the outliers correspond to zero-price days, as April 19 and 25, September 24, November 1 and 2 or December 24 or days with a period close to zero price (February 13). Nonworking days have also some kind of influence over electricity prices, as April 6 (Good Friday) or May 1, August 15, November 1 and December 25. Finally, also some special days can be found related to other holidays or linked to nonworking, this being the case of February 21 (Carnival, nonworking day in part of Spain), August 16 (posterior to a nonworking day), November 2 (in the middle of a long weekend) or December 24 (Christmas Eve).

Finally, again as in the case of the demand, four price curves are detected as outliers by, at least, two of our proposed methods, but none of these curves is identified as an outlier by neither the HDR nor DBT-indep procedure; this is referring to the curves corresponding to February 13, April 10 and 11, and August 15, see Figure 2.10. It seems that makes sense considering them as outliers: the price was zero in some hours on February 13, very low in the second half of the day on April 10 and the first half of the following day, April 11; August 15 is a nonworking day, and the pattern of the corresponding price curve of this day is different from the working days pattern.

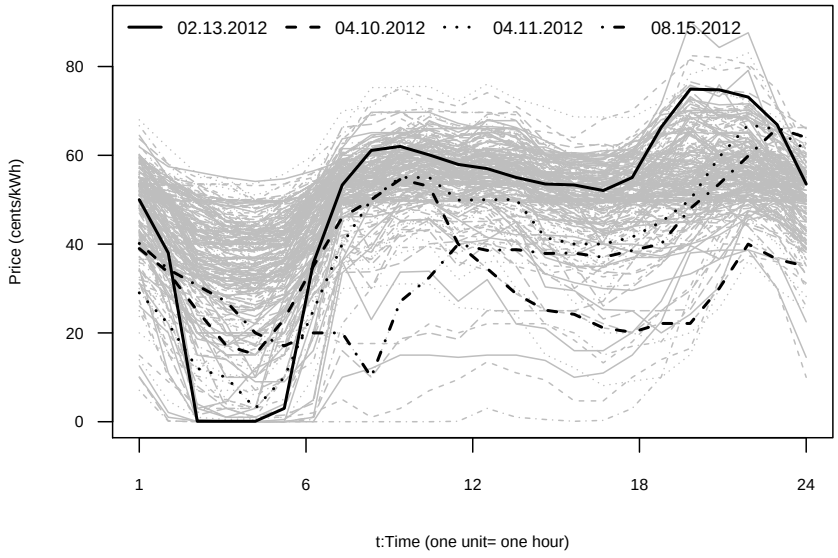


Figure 2.10: Outliers in the price curves detected using procedures for dependent data, but not detected when applying a method designed for independent data.

2.6 Conclusions

This chapter proposes three methods to detect outliers in functional time series. The first one is an extension of the method proposed in Febrero et al. (2008), in which one can make different choices for the election of the functional depth and the bootstrap method, resulting in different combinations. This proposal can be found in Raña, Aneiros and Vilar (2015). It is a contribution derived from this thesis in which the proposed method is presented, together with an analogous simulation study and also an application to a real dataset. In this case, the dataset employed contains temperature and NO_3 emissions data, instead the electrical dataset used here.

The projections-based (PB) and the errors-based (EB) methods use robust functional principal component analysis (FPCA), following the proposal of Hyndman and Ullah (2007). Also these two proposals can be found in Vilar, Raña and Aneiros (2016).

As far as we know, these three proposals are the first methods dealing with the problem of outlier detection, specifically addressed for functional time series. The importance of taking into account the dependence present in this kind of structure has been pointed out along all the chapter.

Simulations were carried out to show the performance of the proposed procedures, comparing them with the correspondent methods for independent data. Results show an improvement in the models with dependent data, specially in the case of magnitude outliers in which the masking effect of the functional time series is more outstanding.

The PB method has very low false detection rate (p_f) while the sensitivity (p_c) of the EB approach is very high. Although in our simulation study small contamination sizes have been considered, both methods show acceptable trade-offs between p_c and p_f. Results for the DBT methods are more difficult to figure out because of the different combinations between depth and bootstrap. In summary, it seems that the best election is the use of the modal depth, specially for the shape outliers, and the MBBo or SmBoR. MBBo and StBo have a very similar behaviour with better correctly detection rate, but also higher false detection rate.

The practical usefulness of this methodology has been illustrated on the daily curves of electricity demand and price in Spain.

It should be emphasized that no universally optimal method for detecting outliers currently exists (not even in the simplest univariate observation setting). It is thus recommendable to apply different approaches and subsequently carry out an accurate analysis of the outliers detected by each (if any) of these approaches.

Chapter 3

Electricity demand and price prediction

3.1 Introduction

Nowadays, in many countries all over the world, the production and sale of electricity is traded under competitive rules in free markets. The agents involved in this market (namely, system operators, regulatory agencies, producers and consumers) are greatly interested in the study of electricity load and price. Since electricity cannot be stored, the demand must be satisfied instantaneously and producers need to anticipate future demands to avoid overproduction. Thus, disposing of good forecasting of electricity demand is important for the agents in the market. On the other hand, if reliable predictions of electricity price are available to producers and consumers, they can develop their bidding strategies and establish a pool-bidding technique to achieve a maximum benefit. Consequently, the prediction of electricity demand and price pose significant concerns to this sector, as pointed out in Section 1.2.

The problem of electricity demand and price forecasting has been widely studied in the literature, most of them from an engineering point of view. Statistics has also a big presence in this field but the use of functional data has not spread far for the moment.

Electricity demand and price predictions can be classified in different classes. First, regarding the prediction horizon, it can be divided in short-term forecasts (this is 1 day ahead hourly forecasts), mid-term forecasts (several days ahead predictions for daily data) and long-term forecasts (one or

more years ahead). There is no consensus about the exact limits of time between these temporal classes, but our study can be included in short-term forecasting, as it is focused on next-day prediction horizon.

A wide range of methodologies and models for forecasting are given in the literature and so, regarding the methodologies employed for this problem, a new classification can be given:

- Statistical approaches: including time series, exponential smoothing, regression analysis, naïve methods, etc.

- Artificial intelligence techniques: neural networks, support vector machines, etc.

Over the last years, artificial intelligence techniques have become a popular tool in any prediction issue. Computational improvements clearly contributed to this fame. The main advantage of these tools is their flexibility: they can deal with complex problems including non-linearity. Among these algorithms, artificial neural networks have received the most attention, as it does not require previous modelling experience to obtain accurate results with no human supervision. This property is also a lack, as it cannot incorporate specific relations into the models.

Some examples of prediction in this field can be found in the literature, for instance methods based on statistical models (regression in Kim et al., 2002; time series in Weron and Misiorek, 2008, etc.) while other ones are based on computational intelligence models (neuronal networks in Singhal and Awarup, 2011; fuzzy neural networks in Chang et al., 2011; support vector machines in Zhao et al., 2008, etc.). See the book by Weron (2006) for a nice monograph on electricity demand and price forecasting. See also Suganthi and Samuel (2012) and Weron (2014) for reviews on electricity demand and electricity price forecasting, respectively.

Focussing on the Spanish Electricity Market, one can find in Cancelo and Espasa (1996) the model used for prediction of electricity demand in Red Eléctrica de España, the system operator. In their model they include outliers treatment and the influence of temperature. Also Cancelo et al. (2008) deal with the problem of demand forecasting. Regarding the electricity price, one can find several studies, most of them based on time series analysis as Nogales et al. (2002), Contreras et al. (2003), Conejo et al. (2005) or García-Martos et al. (2007). For instance, Cruz et al. (2011) analysed the effect of

wind generation and weekdays on the price forecasting.

Most of the works in the literature related to model and forecast electricity demand and price take information from scalar variables; that is the case of researches referred in the previous paragraph. This chapter focuses on some statistical approaches, proposing to use functional data techniques in this field. In recent years, one can find some studies in which this problem has been addressed from a functional perspective. On the one hand, focusing on forecasting daily curves of electricity demand, the reader will find in Anthoch et al. (2010) a parametric model to predict electricity consumption curves in Sardinia; Paparoditis and Sapatinas (2013) introduced a novel functional time series methodology that is applied to historical daily curves of load in Cyprus; Cho et al. (2013) proposed a hybrid approach which was applied to French demand curves. See also Aneiros et al. (2013) for the case of residual demand curves in Spain. On the other hand, when the interest is to forecast scalar values (not curves) from functional data, the reader can see Liebl (2013), where the case of hourly electricity price forecasting is dealt, or Vilar et al. (2012) where, in addition, hourly electricity demand forecasting is considered. See also Goia et al. (2010) for peak load forecasting.

The aim of this study is next-day forecasting of daily curves of electricity demand and price. Three approaches, taking information from a single endogenous functional covariate, are considered, as well as models including, in addition, exogenous scalar covariates. Combined forecasts are also implemented. Forecasts for the corresponding daily curves in the market of mainland Spain, year 2012, are obtained. It can be seen as an extension in two ways of the proposal given in Vilar et al. (2012). The first extension is related to the response in the models: Vilar et al. (2012) considered both FNP and SFPL models with scalar response (hourly electricity demand or price) and functional covariates, while this study deals, among others, with those same models but considering functional response (daily electricity demand or price curve). The second extension concerns the information managed by the partial linear model: while Vilar et al. (2012) did not consider exogenous covariates, this study will take into account information from weather variables and wind power production.

The rest of the chapter is organized as follows: Section 3.2 presents the methods used to predict electricity demand and price. It starts with non-functional approaches, as the naïve method (which is a reference in this field) and the use of ARIMA models, handling the problem from a univariate time series point of view. Also, a prediction method based on FPCA is included.

Finally, two functional regression models are given: FNP regression model and SFPL regression model. After that, Section 3.3 contains a comparative study of all the presented methods in practice, applying them to forecast electricity demand in Subsection 3.3.1 and also electricity price in Subsection 3.3.2. The chapter finishes with Section 3.4, giving the main conclusions of this prediction study.

3.2 Some approaches to forecast electrical data

Both electricity demand and price are assumed to be continuous time stochastic processes, and the same notation is used to refer any of them: $\{\boldsymbol{\chi}(t)\}_{t\in R}$ (units for t are hours). Because $\{\boldsymbol{\chi}(t)\}_{t\in R}$ is a seasonal process with seasonal length $\tau = 24$, and considering that such process is observed on the interval $(a, b]$ with $b = a + N\tau$, the observed daily curves (of electricity demand or price) and the curve to predict can be written as $\{\boldsymbol{\chi}_i\}_{i=1}^{N}$ and $\boldsymbol{\chi}_{N+1}$, respectively, where

$$\boldsymbol{\chi}_i(t) = \boldsymbol{\chi}(a + (i-1)\tau + t), \text{ with } t \in (0, \tau] = (0,\ 24],$$

as the definition given in (1.11) for a Functional Time Series. Then, to forecast the curve $\boldsymbol{\chi}_{N+1}$, information from the previous 365 curves will be used and this will be done for each day in 2012.

In the next subsections, the approaches to predict functional time series are presented. Although such approaches could be useful to predict general functional time series, they are written for the particular case where the curve to predict is the daily electricity demand or price curve $\boldsymbol{\chi}_{N+1}$.

In this setting, the dynamic of the curves depends on the type of day where they are observed: Sundays, weekdays or Saturdays (see Section 1.4). Thus, the considered approaches should take this fact into account.

First, the following notation is introduced, which will be useful to distinguish the various scenarios:

$$\mathcal{I}_0 = \{N-364, N-363, \ldots, N-1, N\},$$
$$\mathcal{I}_{Sat} = \{j \in \mathcal{I}_0 \text{ such that } \boldsymbol{\chi}_j \text{ is a Saturday}\},$$
$$\mathcal{I}_{Sun} = \{j \in \mathcal{I}_0 \text{ such that } \boldsymbol{\chi}_j \text{ is a Sunday}\}$$

and

$$\mathcal{I}_w = \{j \in \mathcal{I}_0 \text{ such that } \boldsymbol{\chi}_j \text{ is a weekday}\}.$$

In addition, defining

$$\mathcal{I}'_w = \{1, 2, \ldots, \#(\mathcal{I}_w)\},$$

one denotes, for $j \in \mathcal{I}'_w$,

$$\boldsymbol{\chi}'_j = \boldsymbol{\chi}_{(j)},$$

where $(j) = \#\{i \in \mathcal{I}_w : i \leq j\}$.

3.2.1 The naïve method

The naïve method, or similar-day method, consists in forecasting the electricity demand or price curve of a given day by means of the electricity demand or price curve in the previous day with the same characteristics. As noted above, in this study three classes of days are considered: Sundays, weekdays and Saturdays. Thus, a Tuesday is similar to the previous Monday, and the same rule applies for Wednesdays, Thursdays and Fridays. A Monday is similar to the Friday of the previous week, a Saturday is similar to the Saturday of the previous week and the same rule applies for Sundays.

As Weron points out (see Weron (2006), Section 3.4.1, pg. 79), this method is very simple but in some cases can be surprisingly powerful and can be used as a benchmark for more sophisticated models.

3.2.2 ARIMA models

The problem of prediction for electricity demand and price has been traditionally addressed using time series. Several authors contributed to this topic applying techniques for time series, mainly ARIMA models, as in Contreras et al. (2003) or Conejo et al. (2005) in the Spanish Electricity Market. For a review about prediction with ARIMA models see, for instance, Brockwell and Davis (1996), Shumway and Stoffer (2006) or Hyndman and Athanasopoulos (2013).

ARIMA models will be considered, even if they are not designed to predict functional time series, in order to compare the accuracy of the presented methods with this popular procedure. In fact, they will be applied separately for 24 univariate time series, one for each hour of the day.

The procedure is computed as follows: from the historical time series of electricity demand, 24 separate univariate time series are built, one for each hour of the day. An ARIMA model of this type $(p, 0, q) \times (P, 1, Q)_7$, is fitted for each time series. This means that, within each of the 24 hourly time

series, the seven days of the week are computed jointly, fitting for them a seasonal ARIMA model of order 7. Through this ARIMA model, considering the previous 365 days, one can compute the prediction for the fixed hour of the next day. In order to rebuild the prediction for the entire day, it only rests to repeat the procedure 24 times, obtaining the whole prediction. At each new day, the coefficients of the ARIMA model, p, q, P and Q are uploaded again.

3.2.3 Robust functional principal component analysis

Hyndman and Ullah (2007) and Hyndman and Shang (2009) proposed an approach to forecast functional time series based on both robust FPCA, which is explained in detail in Section 1.3, and univariate time series prediction (see Aue et al. (2015) for a modified version). First, such approach is presented in the particular case where the curve to predict $\boldsymbol{\chi}_{N+1}$ corresponds to a Saturday. Because the procedure is designed for stationary functional time series forecasting, only information from curves observed in previous Saturdays must be used.

Let us assume that the discretized observed curves can be written as

$$\boldsymbol{\chi}_i(j) = f_i(j) + \sigma_i(j)\varepsilon_{i,j},\ i \in \mathcal{I}_{Sat},\ j = 1, \ldots, 24, \tag{3.1}$$

where $\{\varepsilon_{i,j}\}$ are i.i.d. standard normal random variables, $\sigma_i^2(\cdot)$ is the conditional variance and $f_i(\cdot)$ is a smooth function. The forecasting method proceeds as follows:

- Obtain a nonparametric estimate $\widehat{f}_i(\cdot)$ of $f_i(\cdot)$ from the sample $\{(j, \boldsymbol{\chi}_i(j))\}_{j=1}^{24}$ $(i \in \mathcal{I}_{Sat})$.
- Decompose the fitted curves using Roburst FPCA:

$$\widehat{f}_i(x) = \mu(x) + \sum_{k=1}^{K} \beta_{i,k}\phi_k(x) + e_i(x),\ i \in \mathcal{I}_{Sat}, \tag{3.2}$$

  where $\mu(x)$ is a localization measure of $f_i(x)$, $(\phi_k(x))_{k=1}^{K}$ is a set of functional principal components and $e_i(x)$ is noise with distribution $N(0, v(x))$.

- For each $k = 1, \ldots, K$, fit an ARIMA model to the series of coefficients $\{\beta_{i,k}\}_{i \in \mathcal{I}_{Sat}}$, and use such model to predict $\beta_{N+1,k}$ by $\hat{\beta}_{N+1,k}$.

- Predict the curve $\boldsymbol{\chi}_{N+1}$ by means of

$$\hat{\boldsymbol{\chi}}_{N+1} = \mu + \sum_{k=1}^{K} \hat{\beta}_{N+1,k}\phi_k.$$

Of course, if one wishes to predict the behaviour of the electricity demand or price curve corresponding to a Sunday or a weekday instead of a Saturday, all that must be done is to replace $\mathcal{I}_{Sat}$ in the previous algorithm by $\mathcal{I}_{Sun}$ or $\mathcal{I}_w$, respectively.

This method depends on the number of principal components K in (3.2). As recommended in Hyndman and Shang (2009), it is considered $K = 6$ (a value "larger than any of the components really require").

3.2.4 Functional nonparametric model

An approach based on nonparametric regression with functional both response and covariate, was proposed in Ferraty et al. (2011) in a setting of independent curves. Recently, Aneiros et al. (2013) modified such proposal to be applied on functional time series; in fact, Aneiros et al. (2013) used such model to predict residual demand curves. Also Vilar et al. (2012) employed that model, but considering scalar response. Here, both models will be considered to predict $\boldsymbol{\chi}_{N+1}$ considering scalar or functional response. That is, when considering functional response, it predicts directly the daily curve, while with scalar response it will predict each hour of the day separately. The procedure will be explained first considering functional response and then, indications to adapt it to scalar response are given.

The idea is to construct three regression models, one for each type of day to predict. On the one hand, to forecast a curve corresponding to either Sunday or Saturday information from the previous curve (i.e. from the curve observed on previous Saturday or Friday, respectively) will be used. On the other hand, if one wishes to forecast a curve corresponding to a weekday, information will be taken from the curve observed on the previous weekday (note that the previous weekday to a Monday is a Friday). Specifically, the FNP model in Aneiros et al. (2013), particularized to next-day forecasting of the daily curve of electricity demand or price $\boldsymbol{\chi}_{N+1}$ corresponding to a Saturday, can be written as

$$\boldsymbol{\chi}_{i+1} = m\left(\boldsymbol{\chi}_i\right) + \varepsilon_{i+1},\ i+1 \in \mathcal{I}_{Sat}, \tag{3.3}$$

where $m(\cdot)$ is an unknown smooth functional and ε_{i+1} is the random functional error with zero mean. Then, a prediction $\widehat{\boldsymbol{\chi}}_{N+1}$ for the curve $\boldsymbol{\chi}_{N+1}$ can be obtained by estimating $m\left(\boldsymbol{\chi}_N\right)$ in (3.3). For that, the Nadaraya-Watson type estimator

$$\widehat{m}_h^{FNP}\left(\boldsymbol{\chi}_N\right) = \sum_{i \ / i+1 \in \mathcal{I}_{Sat}} w_h(\boldsymbol{\chi}_N, \boldsymbol{\chi}_i)\boldsymbol{\chi}_{i+1} \tag{3.4}$$

could be used. In (3.4), the weights $w_h(\cdot,\cdot)$ are constructed as

$$w_h(\boldsymbol{\chi}_N, \boldsymbol{\chi}_i) = \frac{K\left(d(\boldsymbol{\chi}_N, \boldsymbol{\chi}_i)/h\right)}{\sum_{j \ / j+1 \in \mathcal{I}_{Sat}} K\left(d(\boldsymbol{\chi}_N, \boldsymbol{\chi}_j)/h\right)}, \tag{3.5}$$

where $K : [0,\infty) \to [0,\infty)$ is a kernel function, $h > 0$ is a smoothing parameter and $d(\cdot,\cdot)$ denotes a semi-metric. The interested reader can find in Ferraty et al. (2011) asymptotic properties of the estimator (3.4) under independence conditions.

Finally, if one wishes to predict the curve $\boldsymbol{\chi}_{N+1}$ corresponding to a Sunday or a weekday, then $\mathcal{I}_{Sat}$ in expressions (3.3)-(3.5) should be replaced by $\mathcal{I}_{Sun}$ or $\mathcal{I}_w^*$, respectively. In addition, for the case of the weekday, $\boldsymbol{\chi}^*$ should be used in (3.3)-(3.5) instead of $\boldsymbol{\chi}$.

As mentioned above, the last procedure is devoted to predict the entire daily curve using functional response within the regression model. However, it will be applied also considering scalar response, predicting each hour of the day separately. Specifically, the functional response $\boldsymbol{\chi}_{i+1}$ is substituted by the scalar one $\boldsymbol{\chi}_{i+1}(j)$, $j = 1, \ldots, 24$. It will be referred as "FNP sc".

This method depends on the bandwidth, h, the semi-metric, $d(\cdot,\cdot)$, and the kernel function, $K(\cdot)$. To choose h, $d(\cdot,\cdot)$ and $K(\cdot)$, the suggestions given in Aneiros et al. (2013) were followed: cross-validation ideas for selecting h and $d(\cdot,\cdot)$ (for details, see Aneiros et al. 2013) and the Epanechnikov kernel $K(t) = 3/4(1-t^2)1_{\{0\leq t\leq 1\}}$.

3.2.5 Semi-functional partial linear model

As noted in Section 1.5, there exist exogenous variables that could improve the predictions of the daily curves of electricity demand and price obtained from models that only consider endogenous variables. In this way, extensions of the procedures given in the previous Sections 3.2.3 and 3.2.4 can be seen in Aue et al. (2015) and Aneiros et al. (2013), respectively. This section

focus on the extension given in Aneiros et al. (2013), applying it also with both functional and scalar response.

The SFPL model proposed in Aneiros et al. (2013) generalizes the previous FNP model (3.3) by incorporating in the regression function a linear component with p exogenous scalar variables. Focusing, again, on the case where the daily curve of electricity demand or price to predict, $\boldsymbol{\chi}_{N+1}$, corresponds to a Saturday, the SFPL model is constructed as

$$\boldsymbol{\chi}_{i+1} = \boldsymbol{X}_{i+1}^T \boldsymbol{\beta} + m\left(\boldsymbol{\chi}_i\right) + \varepsilon_{i+1},\ i+1 \in \mathcal{I}_{Sat}, \tag{3.6}$$

where $\boldsymbol{X}_{i+1}^T = (x_{i+1\ 1}, \ldots, x_{i+1\ p})$ denotes a vector of p scalar covariates, $m(\cdot)$ is a unknown smooth functional, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown functional parameters and ε_{i+1} is the random functional error with zero mean. Computing estimations $\widehat{\boldsymbol{\beta}}$ and $\widehat{m}\left(\boldsymbol{\chi}_N\right)$ of $\boldsymbol{\beta}$ and $m\left(\boldsymbol{\chi}_N\right)$, respectively, in (3.6), the following forecast of $\boldsymbol{\chi}_{N+1}$ is obtained:

$$\widehat{\boldsymbol{\chi}}_{N+1} = \boldsymbol{X}_{N+1}^T \widehat{\boldsymbol{\beta}} + \widehat{m}\left(\boldsymbol{\chi}_N\right).$$

The estimators for $\boldsymbol{\beta}$ and $m(\cdot)$ proposed in Aneiros et al. (2013) will be considered. Such estimators are based on both least squares and kernel smoothing, and their expressions are

$$\widehat{\boldsymbol{\beta}}_h = (\widetilde{\mathbf{X}}_h^T \widetilde{\mathbf{X}}_h)^{-1} \widetilde{\mathbf{X}}_h^T \widetilde{\boldsymbol{\chi}}_h \tag{3.7}$$

and

$$\widehat{m}_h^{SFPL}\left(\boldsymbol{\chi}\right) = \sum_{i+1 \in \mathcal{I}_{Sat}} w_h(\boldsymbol{\chi}, \boldsymbol{\chi}_i) \left(\boldsymbol{\chi}_{i+1} - \boldsymbol{X}_{i+1}^T \widehat{\boldsymbol{\beta}}_h\right), \tag{3.8}$$

respectively. Note that it is denoted $\widetilde{\mathbf{X}}_h = (\mathbf{I} - \mathbf{W}_h)\mathbf{X}$ and $\widetilde{\boldsymbol{\chi}}_h = (\mathbf{I} - \mathbf{W}_h)\boldsymbol{\chi}$, where $\mathbf{W}_h = (w_h(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j))_{i+1, j+1 \in \mathcal{I}_{Sat}}$, $\mathbf{X} = (x_{i+1j})_{\substack{i+1 \in \mathcal{I}_{Sat} \\ 1 \le j \le p}}$ and $\boldsymbol{\chi} = (\boldsymbol{\chi}_{i+1})_{i+1 \in \mathcal{I}_{Sat}}$.

Similar changes as those referred at the end of the previous Section 3.2.4 should be done if, instead of predicting a curve corresponding to a Saturday, the goal is to forecast a curve related to a Sunday or a weekday.

Again, this procedure will be applied also considering scalar response, predicting each hour of the day separately. Specifically, the functional response ,$\boldsymbol{\chi}_{i+1}$, will be changed by the scalar one, $\boldsymbol{\chi}_{i+1}(j)$, $j = 1, \ldots, 24$. It will be referred as "SFPL sc".

This method depends, as the FNP model, on the bandwidth h, the semi-metric $d(\cdot,\cdot)$ and the kernel function $K(\cdot)$ that are selected following the same guidelines as in the FNP model.

The scalar covariates in the SFPL model

Several authors (see, for instance, Hyde and Hodnett (2015), Taylor and Buizza (2003) and Taylor et al. (2006) discussed the impact of meteorological factors (temperature, day light duration, humidity and cloud cover,...) on the electricity demand. Generally, it is observed a large demand of electrical heating in cold weather. As noted in Section 1.5.1, the effect on the temperature on the demand is not linear and so, to include the temperature as a covariate in the SFPL model, some transformation should be applied. In this chapter, the HDD and CDD variables are considered (see (1.12)-(1.13)), which exert a linear effect on the demand (see Figure 1.22). In summary, the vector of scalar covariates included in the SFPL model to forecast the daily curves of electricity demand is $\boldsymbol{X} = (x_1, x_2)^T = (HDD, CDD)^T$.

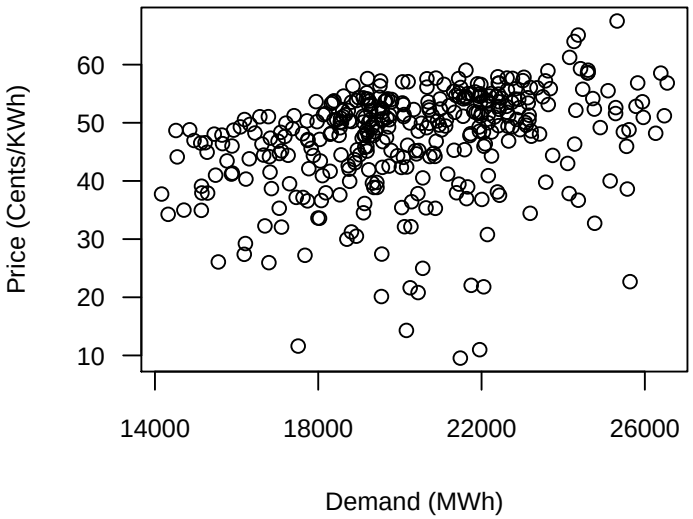


Figure 3.1: Cumulative daily predicted demand against daily mean price in Spain, year 2012.

On the other hand, when the aim is to forecast the daily curves of electricity price, both the predicted daily demand (obtained from the predicted daily curve of demand by using the previous SFPL model) and the wind power production (say $\boldsymbol{X} = (X_1, X_2)^T = (D, WPP)^T$) will be introduced as

scalar covariates. Note that Figure 1.23 suggests a linear effect of the wind power production on the electricity price. Same happens with demand-price as can be seen in Figure 3.1.

At this moment, it is important to highlight the fact that the covariates included in the SFPL models are non-observed (temperature and wind power production corresponding to the day to forecast). Thus, to put in practice our SFPL procedure, one needs to have at hand good next-day forecasts for the values of such covariates. On the other hand, it is known that there exist sophisticated meteorological models giving very good forecasts for temperature and wind power. Nevertheless, neither such models nor forecasts are public. To not mask the predictive power of the SFPL model, which largely deteriorates when not very good forecasts of these covariates are included in it, the decision was to incorporate the ideal forecasts given by the values themselves.

The case of a single SFPL model

The regression function in the SFPL model (3.2.5) includes a linear part. Thus, it can be adapted to model the behaviour of the electricity demand or price curves by means of a single model (instead of a model for each type of day, as in the previous sections). For that, it suffices to change $\mathcal{I}_{Sat}$ and $\boldsymbol{X} = (X_1, X_2)^T$ in (3.6) by $\mathcal{I}_0$ and $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)^T$, respectively, where it is denoted $X_3 = 1_{Saturday}$, $X_4 = 1_{Sunday}$ and $X_5 = 1_{Monday}$ (i.e., the indicators of a Saturday, a Sunday and a Monday, respectively). This modification will be applied considering only functional response due to the computational time cost.

3.3 Forecasting in action

This section reports results related to next-day forecasting of daily curves of electricity demand and price using the approaches described in the previous Section 3.2. In addition, new predictors obtained from combinations of such approaches will be considered.

Remember that electricity demand and price daily curves must be predicted for each day in year 2012, using information from the 365 previous days.

Treatment of outliers

The outlier detection methods presented in Chapter 2 were developed with the aim of identify those days that may disturb the predictions based on electricity demand and price. For that purpose, the outliers detected in Section 2.5 will be considered in this prediction problem.

In order to attenuate the influence of the outliers in the prediction methods, they will be replaced, using weighted moving average, by the surroundings observations. Thus, if $\boldsymbol{\chi}$ denotes an outlier, it will be substituted by the weighted mean of the two previous and two following observations with the corresponding weights indicated in the expression:

$$\boldsymbol{\chi}_{-2}^{(0.2)} \longleftarrow \boldsymbol{\chi}_{-1}^{(0.3)} \longleftarrow \boldsymbol{\chi} \longrightarrow \boldsymbol{\chi}_{+1}^{(0.3)} \longrightarrow \boldsymbol{\chi}_{+2}^{(0.2)}$$

That is, the two closest observations will be averaged with weight 0.3 for each one, and the other two using weight 0.2.

This transformation will be applied within each group of days: weekdays, Saturday and Sunday. Therefore, if $\boldsymbol{\chi}$ corresponds to a Saturday or a Sunday, the observations included in the average will be the two previous and two following Saturdays or Sundays, respectively. In the case of the weekdays, only the observations belonging to the same week will be considered. Thus, if $\boldsymbol{\chi}$ corresponds to a Wednesday, the observations included in the average will be the previous Mondays and Tuesdays and the following Thursday and Friday. In other case, the number of days included will be reduced to those from the same week as the outlier and the weight will be recalculated in proportion.

Finally, as the behaviour of the outliers may differ from the rest of the days, even if its effect is attenuated, they will be also removed from the prediction error computation. Thus, the outliers are not included in the prediction errors shown below.

3.3.1 Forecasting electricity demand

This section focuses on the demand case. To compare the accuracy of each considered model and obtained forecast, $\widehat{\boldsymbol{\chi}}_{N+1}$, from the predictors presented in Section 3.2, the integrated absolute percentage error (IAPE) will be used:

$$IAPE_{N+1} = \frac{1}{24}\int_0^{24} APE_{N+1}(t)dt \approx \frac{1}{24}\sum_{j=1}^{24} APE_{N+1}(j),$$

where

$$APE_{N+1}(t) = 100 \times \left| \frac{\widehat{\boldsymbol{\chi}}_{N+1}(t) - \boldsymbol{\chi}_{N+1}(t)}{\boldsymbol{\chi}_{N+1}(t)} \right|.$$

Note that the IAPE is a generalization of the APE to the continuous setting, the APE being a measure widely used when one forecasts scalar values of electricity demand (for instance, hourly values). Note also that the fact that the electricity demand takes values significantly higher than zero plays a main role in the popularity of this measure.

Comparing approaches to forecast demand

Table 3.1 shows the mean IAPE of each one of the methods presented in Section 3.2. In the case of the SFPL predictor, SFPL1 and SFPL2 refer to the cases where three models (one for each type of day to predict) and a single model are considered, respectively.

Table 3.1: Mean of the IAPE for the electricity demand curves by type of day, week, quarter and year.

	Weekday					Saturday				
Method	Q1	Q2	Q3	Q4	Year	Q1	Q2	Q3	Q4	Year
Naïve	**4.43**	5.91	4.46	6.36	5.24	7.29	8.06	5.21	13.62	8.41
ARIMA	4.66	6.14	4.45	7.02	5.52	4.97	6.23	4.42	10.00	6.30
RFPCA	4.45	**5.20**	**4.16**	**6.21**	**4.96**	6.50	7.67	4.67	13.33	7.89
FNPsc	5.08	6.09	4.78	6.59	5.60	6.11	6.74	4.28	11.08	6.94
FNP	5.26	6.06	4.69	6.40	5.58	5.82	6.06	4.29	10.85	6.64
SFPL1sc	4.58	5.91	4.34	6.40	5.27	6.02	6.15	4.31	10.30	6.59
SFPL1	4.61	5.68	4.28	6.52	5.23	5.47	5.49	**4.10**	**9.98**	6.15
SFPL2	5.26	6.17	4.41	7.36	5.76	**4.43**	**5.18**	4.21	10.14	**5.86**
	Sunday					Week				
Naïve	9.08	5.62	8.94	16.50	10.10	5.50	6.18	5.21	8.84	6.39
ARIMA	**6.13**	6.37	6.78	10.16	**7.37**	**4.91**	6.18	4.78	7.90	5.90
RFPCA	9.81	**5.20**	7.78	14.02	9.25	5.51	**5.55**	4.75	8.35	6.00
FNPsc	9.66	6.24	6.30	11.55	8.44	5.88	6.21	4.92	7.94	6.20
FNP	8.86	6.39	**5.74**	10.27	7.80	5.85	6.10	4.79	**7.59**	6.05
SFPL1sc	10.87	6.47	6.41	11.61	8.84	5.69	6.02	4.63	7.70	5.97
SFPL1	9.60	5.77	5.94	11.36	8.17	5.45	5.67	**4.49**	7.70	**5.78**
SFPL2	8.28	6.31	6.78	**8.89**	7.57	5.57	6.05	4.72	7.98	6.03

Several conclusions can be drawn from Table 3.1. On the one hand, in Saturdays and, specially, in Sundays the errors are bigger than in the weekdays. The effect of the quarter of the year is also remarkable and it agrees with the descriptive analysis in Section 1.4. Note that the worst predictions are obtained in the last quarter, in which a bigger variance makes prediction even more difficult.

On the other hand, focusing on the prediction methods, one can see that, for the case of the weekdays, the simplest naïve approach is not far away from the others, that are much more complicated. RFPCA and FNP procedures are competitive but, in general, they are overcome by methods SFPL1 and SFPL2, that incorporate exogenous covariates.

Some advantages of implementing curve forecasting against scalar (i.e. hourly) forecasting can be shown. Among the scalar predictors, the ARIMA model gives, in general, the best hourly predictions. Furthermore, the forecasts obtained from models with functional response (FNP and SFPL1) improve, in general, the results of those obtained from the corresponding models with scalar one (FNP sc and SFPL1 sc). Except in the particular case of Sundays, predictions from SFPL1 are also better than the ones obtained from the ARIMA model.

Finally, it is worth to be noted that the computational cost to obtain the 24 forecasts for each day by means of models with functional response is much lower than the consumed time when one uses the corresponding models with scalar response. For instance, the time consumed by the SFPL1 sc method is fivefold the time corresponding to the SFPL1.

Combining forecasts

Various (five) methods to forecast curves have been proposed in Section 3.2 and implemented above. In this situation, many authors have invited to use a method for combining forecasts, this proposal being based on the idea that the forecast combinations could provide better results than the individual methods (see Timmermann (2006) and Wallis (2011)). See, in addition, Taylor and Majithia (2000), Weron (2006), Taylor (2010) and Weron (2014) for studies on combined forecasts in the context of electricity markets.

In this section, two simple methods of combined forecasts are considered. They are as follows:

- CF1: this is the simplest combination. It consists just in averaging the results of all the proposed methods, allowing to offset the excess and default predictions for each day.

- CF2: for each group of days (weekdays, Saturday and Sunday), it averages the results from the two predictors that give the minimum yearly mean IAPE.

Note that, due to the prediction errors obtained and the computational cost, only functional-response methods are considered within the combinations. Table 3.2 reports the combinations considered in the CF2 method, while Table 3.3 shows the mean IAPE of the two combined procedures CF1 and CF2.

Table 3.2: Composition for combined method CF2 (electricity demand curves).

Weekday	Saturday	Sunday
RFPCA + SFPL1	SFPL1 + SFPL2	FNP + SFPL2

Table 3.3: Combining forecasts. Mean of the IAPE for the electricity demand curves by type of day, week, quarter and year.

	Weekday					Saturday				
Method	Q1	Q2	Q3	Q4	Year	Q1	Q2	Q3	Q4	Year
CF1	**4.29**	5.36	**3.99**	**6.00**	**4.87**	4.96	5.43	**3.94**	10.90	6.17
CF2	4.35	**5.25**	4.10	6.18	4.93	**4.75**	**4.81**	4.04	**10.00**	**5.77**
	Sunday					Week				
CF1	**7.74**	**5.19**	6.34	11.35	7.68	**4.88**	5.34	**4.32**	7.46	5.46
CF2	8.47	5.90	**5.90**	**9.12**	**7.35**	4.99	**5.28**	4.35	**7.15**	**5.40**

Table 3.3 suggests that, from the point of view of the weekly mean, it is convenient to combine forecasts (note that both CF1 and CF2 methods improve the accuracy of each of the individual procedures considered, in each quarter as well as in the whole year). Such improvement is frequent in the weekdays, while Sunday is the type of day in which is least advisable to make

combinations.

Figure 3.2 shows a visual comparison of the daily errors (IAPE) of the methods naïve, SFPL1 and CF2, while Figure 3.3 displays the curves forecasted from these methods for the ninth week in each quarter (from Monday to Sunday).

Figure 3.2 suggests comparable results for the SFPL1 and CF2 methods, the accuracy of the CF2 being slightly better. In addition, one can see that, in general, the naïve fails in forecasting the whole year. Figure 3.3 clearly shows the poor behaviour of the naïve method to forecast curves in weekends, this being the reason of its poor global accuracy.

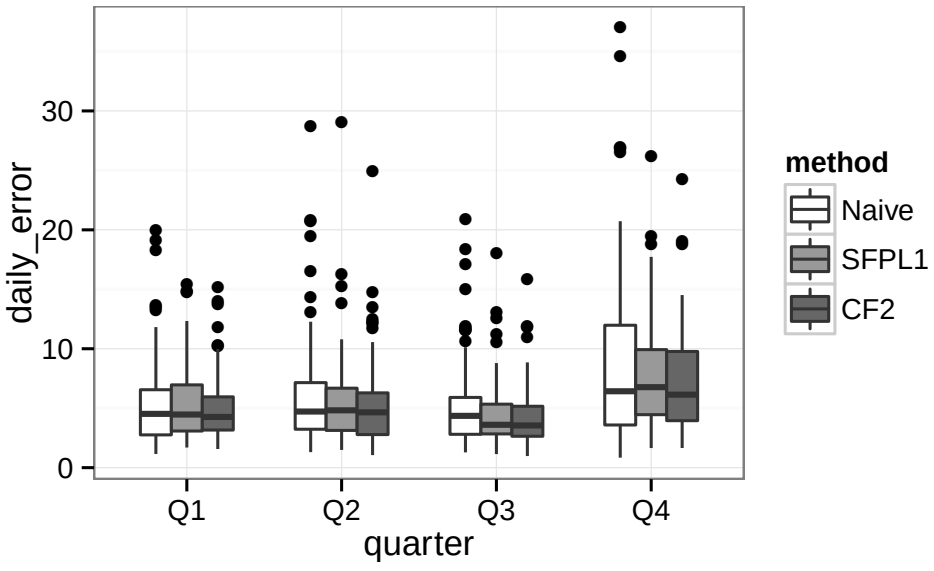


Figure 3.2: Daily errors (IAPE) for electricity demand curves corresponding to Naïve, SFPL1 and CF2.

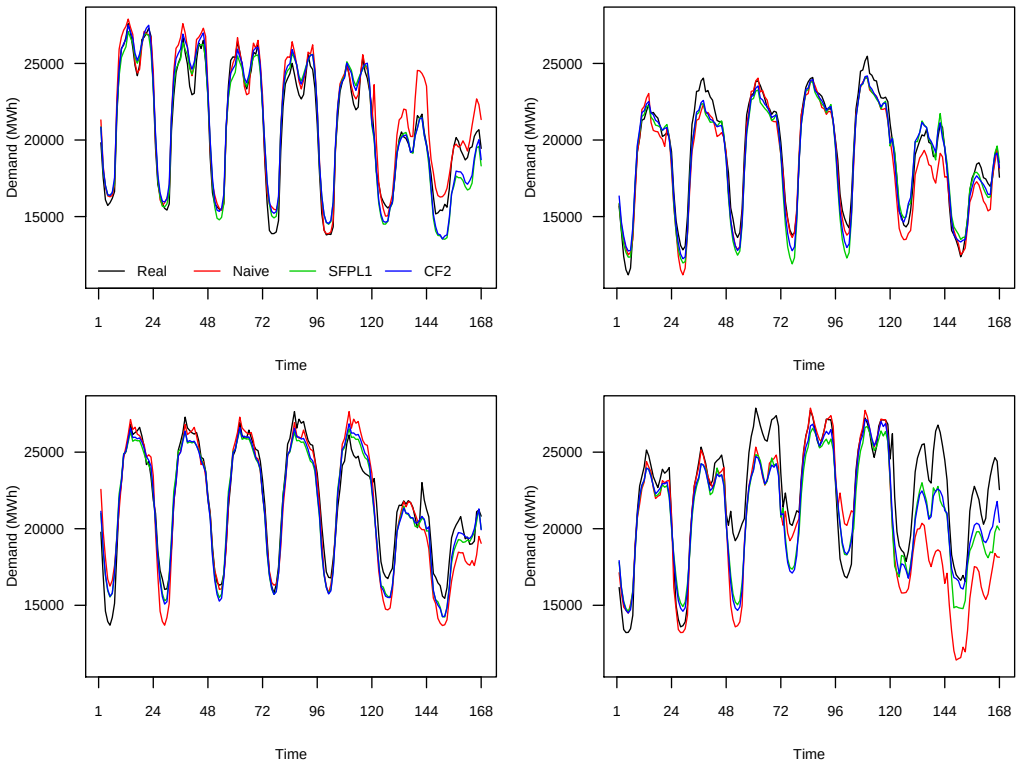


Figure 3.3: Actual demand (black lines) and predicted curves using the naïve method (red lines), SFPL1 (green lines) and CF2 (blue lines) for the weeks: (Up-left panel) February 27–March 4, (Up-right panel) May 28–June 3, (Bottom-left panel) August 27–September 2, and (Bottom-right panel) November 26–December 2.

3.3.2 Forecasting electricity price

A similar study to the previous one, but now focused on price instead of demand, is carried out in this section. In this case, because price takes values close (or even equal) to zero, an absolute measure of accuracy must be considered, the integrated absolute error (IAE):

$$IAE_{N+1} = \frac{1}{24}\int_0^{24} AE_{N+1}(t)dt \approx \frac{1}{24}\sum_{j=1}^{24} AE_{N+1}(j),$$

where

$$AE_{N+1}(t) = \left|\widehat{\boldsymbol{\chi}}_{N+1}(t) - \boldsymbol{\chi}_{N+1}(t)\right|.$$

Comparing approaches to forecast price

Table 3.4 shows the mean IAE of each one of the methods presented in Section 3.2. As in the previous section, SFPL1 and SFPL2 refer to the cases where three SFPL models (one for each type of day to predict) and a single SFPL model are considered, respectively.

Similar conclusions as those obtained in Section 3.3.1 for the demand case are drawn from Table 3.4 for this price case: (i) the best forecasts are obtained for the weekdays, (ii) the more difficult quarter to predict is the last one, (iii) the naïve approach is a simple predictor giving good results for weekdays, and (iv) the RFPCA and FNP predictors are competitive, but they are overcome by the procedures SFPL1 and SFPL2, which take advantage from exogenous covariates.

Again, one can compare functional-response methods and predictors for hourly price (not curve). The same notation as in Section 3.3.1 is used. As noted from Table 3.4, contrary to what happened in the demand case, the ARIMA approach is not better among the scalar predictors. In this case, the method SFPL1 sc, which incorporates exogenous covariates, clearly overcomes the other two approaches. Finally, comparing methods FNP sc and SFPL1 sc with methods FNP and SFPL1, respectively, one observes a slight improvement favourable to the procedures that consider functional response (i.e. FNP and SFPL1). Remember that, in addition, these procedures require minor computational time to predict the 24 hourly prices in a day than the scalar ones.

Table 3.4: Mean of the IAE for the electricity price curves by type of day, week, quarter and year.

	Weekday					Saturday				
Method	Q1	Q2	Q3	Q4	Year	Q1	Q2	Q3	Q4	Year
Naïve	4.96	7.58	4.42	7.07	5.95	5.63	7.52	7.73	12.22	8.13
ARIMA	5.65	6.66	4.61	6.97	5.94	5.16	**5.43**	5.73	8.69	6.16
RFPCA	4.69	6.16	3.95	6.37	5.26	4.88	7.23	6.68	8.93	6.85
FNP sc	5.31	7.57	4.42	6.68	5.96	4.72	6.18	4.75	9.54	6.20
FNP	5.08	7.36	4.41	6.58	5.82	4.61	6.38	5.17	9.94	6.42
SFPL1 sc	4.80	5.63	**3.72**	5.00	4.77	4.38	**5.43**	4.94	7.81	5.56
SFPL1	4.65	**5.50**	3.83	5.08	4.75	**3.86**	5.63	**4.51**	7.55	5.32
SFPL2	**4.45**	5.57	3.74	**4.47**	**4.61**	4.13	5.51	4.67	**7.01**	**5.27**
	Sunday					Week				
Naïve	8.29	6.92	10.35	13.79	9.90	5.53	7.48	5.74	8.76	6.83
ARIMA	6.92	**5.95**	7.66	11.27	7.98	5.76	6.39	5.21	7.83	6.27
RFPCA	8.11	6.95	7.34	11.44	8.47	5.21	6.43	4.83	7.46	5.94
FNP sc	7.73	7.07	7.73	12.90	8.87	5.57	7.30	4.94	7.97	6.41
FNP	7.32	6.63	7.65	14.18	8.97	5.33	7.12	4.98	8.14	6.36
SFPL1 sc	6.01	6.93	6.20	9.33	7.10	4.91	5.78	4.25	6.02	5.22
SFPL1	**5.71**	7.02	5.86	9.33	6.96	4.69	5.74	4.21	6.04	5.15
SFPL2	6.07	6.52	**5.52**	**8.47**	**6.63**	**4.64**	**5.69**	**4.13**	**5.60**	**5.00**

Combining forecasts

In order to improve the accuracy of the several predictors considered in the previous section, the combined forecasts CF1 and CF2 were implemented, constructed in a similar way as in Section 3.3.1. For any type of day, the two predictors included in CF2 were the SFPL1 and SFPL2. Table 3.5 reports the mean IAE of the CF1 and CF2 approaches.

As in the demand case, Table 3.5 suggests the convenience of combining predictors in order to improve the individual accuracy of each predictor. Note, for instance, that combination CF2 gives better yearly accuracy (on each type of day) than any individual predictor considered.

Figure 3.4 shows a visual comparison of the daily errors (IAE) of the methods naïve, SFPL1 and CF2, while Figure 3.5 displays the curves forecasted from these methods for the ninth week in each quarter (from Monday to Sunday).

Table 3.5: Combining forecasts. Mean of the IAE for the electricity price curves by type of day, week, quarter and year.

	Weekday					Saturday				
Method	Q1	Q2	Q3	Q4	Year	Q1	Q2	Q3	Q4	Year
CF1	**4.06**	5.77	**3.52**	5.25	4.62	3.97	5.59	5.09	8.11	5.60
CF2	4.44	**5.46**	3.65	**4.79**	**4.57**	**3.88**	**5.40**	**4.30**	**7.19**	**5.13**
	Sunday					Week				
CF1	6.07	**5.72**	6.62	9.73	7.05	**4.34**	5.74	4.18	6.30	5.11
CF2	**5.84**	6.59	**5.51**	**8.53**	**6.60**	4.56	**5.61**	**4.01**	**5.66**	**4.94**

As expected from the previous study, Figure 3.4 suggests comparable results for the SFPL1 and CF2 methods, the accuracy of the CF2 being slightly better (again). The fail of the naïve method in forecasting the whole year is even more clear than in the demand case. In fact, Figure 3.5 shows as the naïve method fails even is some weekdays.

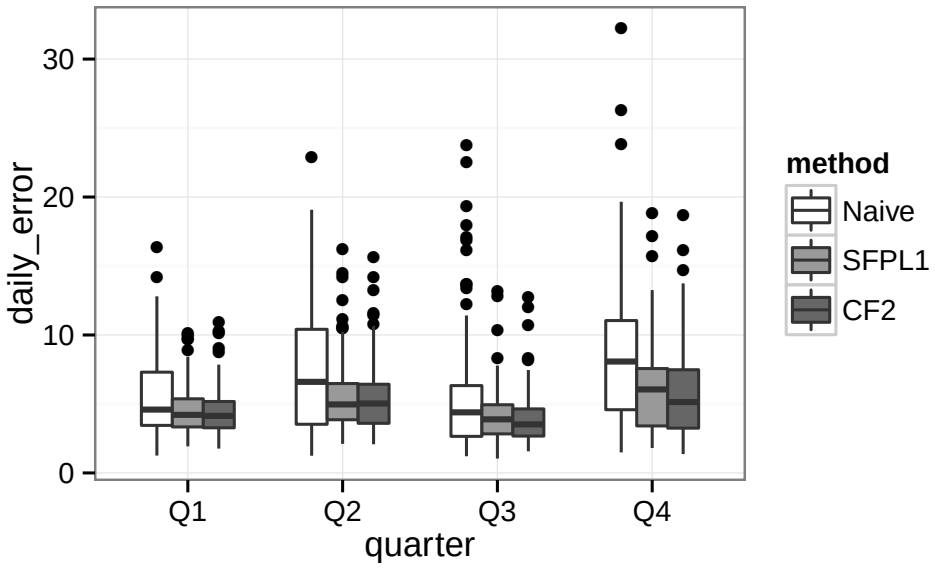


Figure 3.4: Daily errors (IAE) for electricity price curves corresponding to naïve, SFPL1 and CF2.

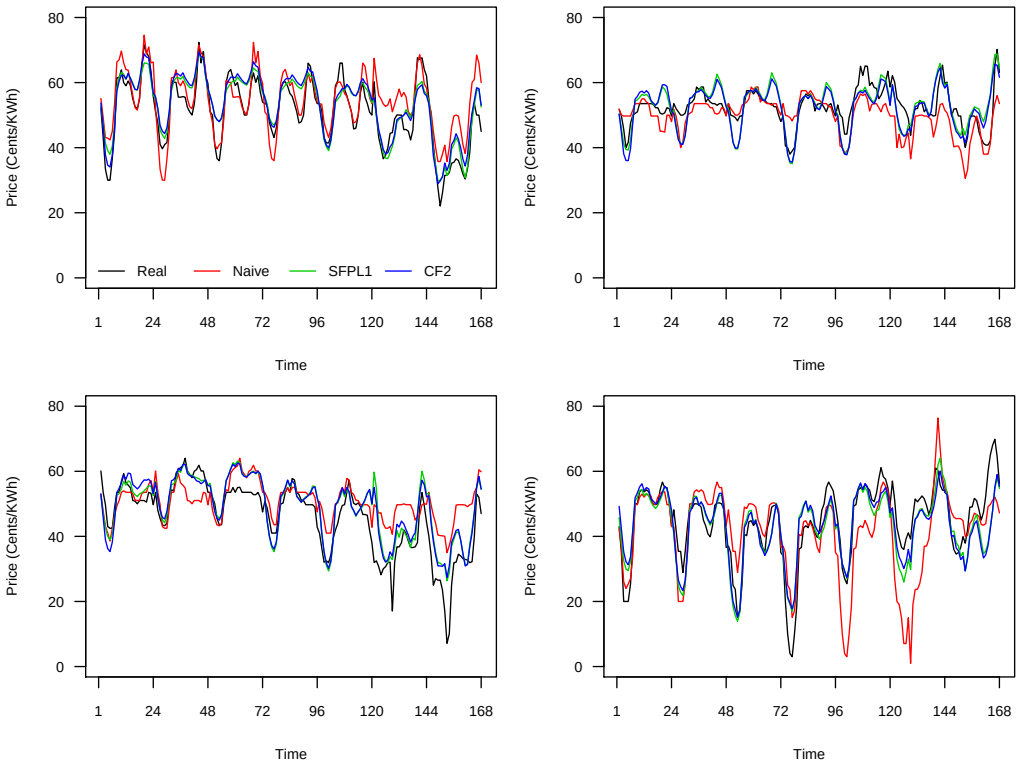


Figure 3.5: Actual price (black lines) and predicted curves using the naïve method (red lines), SFPL1 (green lines) and CF2 (blue lines) for the weeks: (Up-left panel) February 27–March 4, (Up-right panel) May 28–June 3, (Bottom-left panel) August 27–September 2, and (Bottom-right panel) November 26–December 2.

3.4 Conclusions

New methods to address next-day forecasting of daily curves of electricity demand and price have been successfully used for Spanish data. Aneiros, Vilar and Raña (2016) summarize the study developed here. Four main conclusions were obtained: (i) to take information from exogenous covariates improves the forecasts, (ii) combining forecasts is a simple technique giving nice improvements in the individual accuracy of different methods, (iii) hourly predictions obtained from the discretization of the predicted curves are better than the corresponding predictions given by functional models with scalar response, and, interestingly, (iv) the computational time to obtain such hourly predictions from discretization is much lower than the needed time to implement the scalar procedures.

It may be highlighted that prediction in this field is not an easy task and, it is even more difficult in the selected period of 2012. If one compares the results given in Vilar et al. (2012), for the naïve method in the year 2008 (which is exactly the same presented here), one can see lower errors. There are many possible reasons to justify this difference, looking at the economic conditions or market and industry instability. In Figure 3.6, one can see how the demand in 2012 becomes more unstable and, as a consequence, more difficult to be predicted.

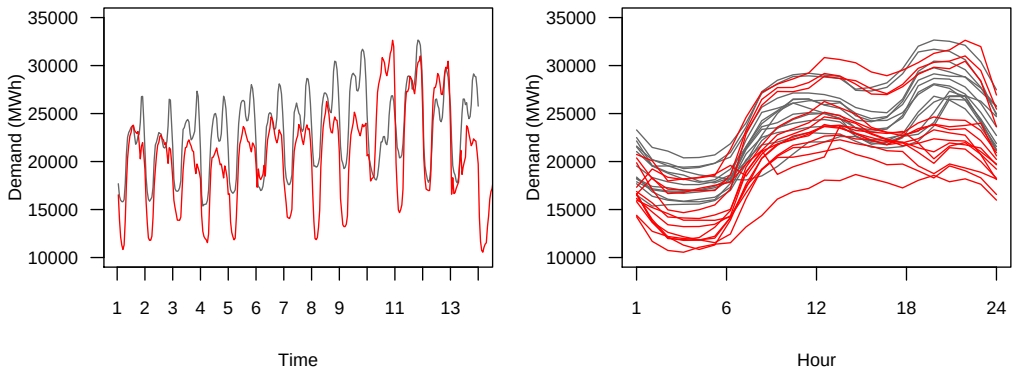


Figure 3.6: Comparison between electricity demand in Mondays for the last quarter of 2008 (black line) and 2012 (red).

Future research includes the extension of the SFPL model to autoregressions with higher order than one and/or to incorporate functional covariates exerting linear effect. In addition, it could be interesting to seek for new informative covariates that can enter in the SFPL model either in parametric

or nonparametric form and also consider other regression models as, for instance, generalized additive models for functional data. On the other hand, this chapter is focused on point forecasts but it would be nice to construct prediction bands from our functional data approaches. It that sense, the following chapters will be focus on bootstrap procedures to build confidence and prediction intervals for the FNP and SFPL regression models.

Finally, it is worth being noted that the simple methods of combining forecasts used in this paper have provided encouraging results. Thus, other interesting future research line is the optimal selection of the weights corresponding to each forecast included in the combination of forecasts.

Chapter 4

Confidence Intervals in Functional Nonparametric Regression

4.1 Introduction

This chapter is devoted to study the FNP regression model, when a functional explanatory variable and scalar response are considered. Naïve and wild bootstrap procedures are proposed to construct pointwise confidence intervals for the nonparametric regression function when the data are dependent. Assuming α-mixing conditions on the sample, the asymptotic validity of the both procedures is obtained. A simulation study shows promising results when finite sample sizes are used, while an application to electricity demand data illustrates its usefulness in practice.

The asymptotic results present in this chapter represent the extension, to the case of dependent data, of the study from Ferraty, Van Keilegom and Vieu (2010). In that paper, they dealt with independent data and proposed also the two bootstrap procedures to construct pointwise confidence intervals for the same regression model.

The chapter is organized as follows: the FNP regression model, together with the two bootstrap procedures are presented in Section 4.2. Section 4.3 involves all the asymptotic results, including the assumptions (Subsection 4.3.1) needed to establish the main result of this chapter: the theorem giving the validity of the two bootstrap procedures, that is stated in Subsection 4.3.2. Within this section, some preliminary results are included in Subsection 4.3.3 and the proofs of the theorems are detailed in Subsection 4.3.3. A simulation study is carried out in Section 4.4, while Section 4.5 includes the

application of the proposed procedures to electrical data from the Spanish Electricity Market. Finally, some conclusions of this chapter are given in Section 4.6.

4.2 The model and the bootstrap procedures

A FNP regression model is developed along this chapter. Since our main interest falls upon functional time series, one considers the model:

$$G(\boldsymbol{\chi}_{i+1}) = m(\boldsymbol{\chi}_i) + \varepsilon_i \ (i = 1, \ldots, n), \tag{4.1}$$

where $G(\cdot)$ is a known operator and $m(\cdot)$ is an unknown operator to be estimated, while $\boldsymbol{\chi}_i$ are functional random variables and ε_i are zero-mean regression errors.

The interest lies on time series prediction from the FNP model (4.1) with scalar response (i.e. assuming that the operator $G(\cdot)$ is real valued), and in which the dependence is controlled by means of some strong dependence condition: namely the covariates $\boldsymbol{\chi}_i$ are identically distributed functional random variables verifying some α-mixing condition. Other less important restrictions on this model are that the ε_i are i.i.d. zero-mean random errors, and that $\boldsymbol{\chi}_i$ are valued in some infinite-dimensional space $\mathcal{H}$, which is endowed with a semi-metric $d(\cdot, \cdot)$.

Although the interest is time series prediction from Model (4.1) with scalar response, asymptotic theory for the more general model

$$Y_i = m(\boldsymbol{\chi}_i) + \varepsilon_i \tag{4.2}$$

is obtained, where the process $\{(\boldsymbol{\chi}_i, Y_i)\}$ is α-mixing and identically distributed as $(\boldsymbol{\chi}, Y)$. In this way, the results will be valid even when the response is exogenous. As indicated at the beginning of this chapter, the response, Y, is scalar while the covariate, $\boldsymbol{\chi}$, is valued in some infinite-dimensional space, $\mathcal{H}$, which is endowed with a semi-metric $d(\cdot, \cdot)$. Finally, $m(\cdot)$ is an unknown smooth real-valued operator and the corresponding random errors $\{\varepsilon_i\}$ are i.i.d. as ε, and it is assumed that $\mathbb{E}(\varepsilon|\boldsymbol{\chi}) = 0$ and $\mathbb{E}(\varepsilon^2|\boldsymbol{\chi}) = \sigma_\varepsilon^2(\boldsymbol{\chi}) < \infty$. Let

$$\mathcal{S} = \{(\boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{\chi}_n, Y_n)\}$$

denote the sample one has at hand.

Given a fixed element χ in the space $\mathcal{H}$, the remainder of this chapter focuses on inference on $m(\chi)$ for Model (4.2). Specifically, the aim is to construct confidence intervals for $m(\chi)$. On the one hand, in the setting of independent data, $\{(\boldsymbol{\chi}_i, Y_i)\}$, this topic was dealt in Ferraty et al. (2007), which obtained the asymptotic normality of a properly standardized estimator, $\widehat{m}_h(\chi)$; then, by estimating the constants involved in the standardized estimator one can construct the corresponding confidence intervals. The main drawback of this procedure is that such constants could be difficult to estimate (for some simple examples, see Proposition 1 in Ferraty et al., 2007). This drawback was overcome in Ferraty, Van Keilegom and Vieu (2010) by means of bootstrapping techniques, by approximating directly the distribution of the estimation error without having to estimate the constants involved in the standardized estimator. On the other hand, some studies exist in the case of dependent data, $\{(\boldsymbol{\chi}_i, Y_i)\}$. For instance, Masry (2005) and Delsol (2009) obtained the asymptotic normality of a properly standardized estimator, $\widehat{m}_h(\chi)$, under α-mixing conditions. The main advantage of the results in Delsol (2009) against the ones in Masry (2005) is the fact that Delsol obtained explicit constants, which is not the case of Masry (2005). As in the setting of independent data, recently referred, there exist situations where the constants given in Delsol (2009) are difficult to estimate, and this drawback could be overcome, again, through implementation of bootstrap techniques. In the following, two bootstrap procedures designed for that are presented.

The same estimator $\widehat{m}_h(\cdot)$ of the regression function $m(\cdot) = \mathbb{E}(Y \mid \boldsymbol{\chi} = \cdot)$ as in Masry (2005), Ferraty et al. (2007), Delsol (2009) and Ferraty, Van Keilegom and Vieu (2010) is considered; that is,

$$\widehat{m}_h(\chi) = \frac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h) Y_i}{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h)}, \tag{4.3}$$

where $K(\cdot)$ is a kernel function and $h > 0$ is a smoothing parameter. In addition, we will focus on both naïve and wild bootstrap procedures, which have been successfully used in the literature related to regression models (for models with scalar variables, see for instance Freedman, 1981, and Mammen, 1993 for linear models, and Cao, 1991, Härdle and Marron, 1991, and Hall, 1992, for nonparametric models; the case of models with functional variables was studied, for instance, in González-Manteiga and Martínez-Calvo, 2011, for linear models while Ferraty, Van Keilegom and Vieu, 2010 and 2012, focused on nonparametric models).

The algorithms for resampling proceed as follows:

Naïve bootstrap.

This procedure is designed for the case where the model is homoscedastic; that is, $\sigma_\varepsilon^2(\boldsymbol{\chi}) = \sigma_\varepsilon^2$.

Step 1: Fix some pilot bandwidth b and construct the residuals $\widehat{\varepsilon}_{i,b} = Y_i - \widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$.

Step 2: Draw n i.i.d random variables $\varepsilon_1^*, \ldots, \varepsilon_n^*$ from the empirical distribution function of $(\widehat{\varepsilon}_{1,b} - \overline{\widehat{\varepsilon}}_b, \ldots, \widehat{\varepsilon}_{n,b} - \overline{\widehat{\varepsilon}}_b)$, where $\overline{\widehat{\varepsilon}}_b = n^{-1} \sum_{i=1}^n \widehat{\varepsilon}_{i,b}$.

Step 3: Obtain $Y_i^* = \widehat{m}_b(\boldsymbol{\chi}_i) + \varepsilon_i^*$, $i = 1, \ldots, n$.

Step 4: Define $\widehat{m}_{hb}^*(\chi) = \dfrac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h) Y_i^*}{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h)}$.

Wild bootstrap.

This procedure allows the possibility of heteroscedasticity, but is also applicable in the case of homoscedasticity. All one must do is to change Step 2 in the naïve bootstrap: define $\varepsilon_i^* = \widehat{\varepsilon}_{i,b} V_i$, $i = 1, \ldots, n$, where $V_1, \ldots, V_n$ are i.i.d. random variables that are independent of the data $\mathcal{S}$ and that satisfy $\mathbb{E}(V_1) = 0$ and $\mathbb{E}(V_1^2) = 1$. The other three steps are maintained.

Remark 1 *As usual, when one deals with asymptotic related to bootstrap procedures in nonparametric regression, two bandwidths are involved in both algorithms. The first bandwidth, b, is used to construct the residuals to resample. Then, a second bandwidth, h, is considered to smooth the bootstrap resample. Our assumptions imposed to obtain the asymptotic validity of the proposed bootstrap procedures require that b must be taken to be larger than h (in the same way as in the independent case considered by Ferraty, Van Keilegom and Vieu, 2010 and 2012; see also Cao, 1991, and Härdle and Marron, 1991, for the scalar case).*

4.3 Asymptotic theory

This section presents the main result: the asymptotic validity of the two proposed bootstrap procedures. First, the considered assumptions are stated and some comments on them are given. Then, the asymptotic result is stated.

4.3.1 Assumptions

Let us start with some notation. For a given fixed element χ of the space $\mathcal{H}$, let denote:

$$\begin{aligned} B(\chi, l) &= \{\chi_1 \in \mathcal{H} \text{ such that } d(\chi_1, \chi) \leq l\}, \\ F_\chi(l) &= P(\boldsymbol{\chi} \in B(\chi, l)) \text{ for } l > 0, \\ \varphi_\chi(s) &= \mathbb{E}(m(\boldsymbol{\chi}) - m(\chi) | d(\boldsymbol{\chi}, \chi) = s) \\ \tau_{h\chi}(s) &= F_\chi(hs)/F_\chi(h) \text{ for } s \in (0, 1] \end{aligned}$$

and

$$\tau_{0\chi}(s) = \lim_{h \downarrow 0} \tau_{h\chi}(s).$$

In addition, let

$$\begin{aligned} M_{0\chi} &= K(1) - \int_0^1 (sK(s))' \tau_{0\chi}(s) ds, \\ M_{1\chi} &= K(1) - \int_0^1 K'(s) \tau_{0\chi}(s) ds, \\ M_{2\chi} &= K^2(1) - \int_0^1 (K^2(s))' \tau_{0\chi}(s) ds \end{aligned}$$

and

$$\Theta(s) = \max\{\max_{i \neq j} P(d(\boldsymbol{\chi}_i, \chi) \leq s, d(\boldsymbol{\chi}_j, \chi) \leq s), F_\chi^2(s)\}.$$

As noted at the beginning of Section 4.3, the asymptotic validity of the two proposed bootstrap procedures will be proved theoretically. For that, it will be obtained that both the standard estimator, $\widehat{m}_h(\chi)$, and the bootstrap version, $\widehat{m}^*_{hb}(\chi)$, (properly standardized) converge to the same distribution, existing, in addition, a third negligible term (for details, see Section 4.3.3). The following set of assumptions guarantees the convergence of $\widehat{m}_h(\chi)$:

$$m(\cdot) \text{ and } \sigma_\varepsilon^2(\cdot) \text{ are continuous on a neighbourhood of } \chi;\ \sigma_\varepsilon^2(\chi) > 0 \quad (4.4)$$

$$F_\chi(0) = 0 \text{ and } \varphi_\chi(0) = 0 \text{ and } \varphi'_\chi(0) \text{ exists} \quad (4.5)$$

$$\forall s \in [0, 1],\ \lim_{n \to \infty} \tau_{h\chi}(s) = \tau_{0\chi}(s) \text{ with } \tau_{0\chi}(s) \neq 1_{[0,1]}(s) \quad (4.6)$$

$$\exists p > 2,\ \exists M > 0 \text{ such that } \mathbb{E}(|\varepsilon|^p|\boldsymbol{\chi}) \leq M \ a.s. \tag{4.7}$$

and

$$\max\{\mathbb{E}(|Y_iY_j|^p|\boldsymbol{\chi}_i,\boldsymbol{\chi}_j),\mathbb{E}(|Y_i|^p|\boldsymbol{\chi}_i,\boldsymbol{\chi}_j)\} \leq M \ a.s.\ \forall i,j \in \mathbb{Z} \tag{4.8}$$

$$h(nF_\chi(h))^{1/2} = O(1) \text{ and } \lim_{n\to\infty} nF_\chi(h) = \infty \tag{4.9}$$

$$\begin{gathered} K(\cdot) \text{ is supported on } [0,1], \text{ has a continuous derivative on } [0,1), \\ K'(s) \ \leq 0 \text{ for } s \in [0,1) \text{ and } K(1) > 0 \end{gathered} \tag{4.10}$$

$$\begin{gathered} \{(\boldsymbol{\chi}_i, Y_i)\}_{i=1}^n \text{ comes from a } \alpha\text{-mixing process with} \\ \alpha\text{-mixing coefficients } \alpha(n) \leq Cn^{-a}, \end{gathered} \tag{4.11}$$

the exponent a being related with both the small ball probabilities $F_\chi(h)$ and the number of moments p in the following way:

$$\begin{gathered} \exists v > 0 \text{ such that } \Theta(h) = O(F_\chi(h)^{1+v}) \text{ with } a > \tfrac{(1+v)p-2}{v(p-2)} \\ (\text{with } p \text{ and } a \text{ introduced in (4.7) and (4.11), respectively}) \end{gathered} \tag{4.12}$$

and

$$\begin{gathered} \exists \gamma > 0 \text{ such that } nF_\chi(h)^{1+\gamma} \to \infty \text{ and } a > \max\left\{\tfrac{4}{\gamma}, \tfrac{p}{p-2} + \tfrac{2(p-1)}{\gamma(p-2)}\right\} \\ (\text{with } p \text{ and } a \text{ introduced in (4.7) and (4.11), respectively}). \end{gathered} \tag{4.13}$$

Remark 2 *Most of the assumptions (4.4)-(4.13) are standard ones in the setting of nonparametric regression with functional data. In fact, all of them were used in Delsol (2009) to obtain the asymptotic distribution of $\widehat{m}_h(\chi)$ under α-mixing conditions (see Theorem 28 in Appendix A). As can be seen in Delsol (2009), assumptions (4.11)-(4.13), related to arithmetic α-mixing coefficients, could be changed by other more general assumptions (but less clear). The interested reader will find in Delsol (2009) justifications of assumptions (4.4)-(4.13). Finally, it is worth being noted that other sets of assumptions to obtain the asymptotic distribution of $\widehat{m}_h(\chi)$ under α-mixing conditions could be considered. For instance, one could use the assumptions presented in Masry (2005). Note that although the drawback of the assumptions given in Masry (2005) (the constants involved in the standardization of $\widehat{m}_h(\chi)$ are not explicit), actually this is not a problem when, as in this chapter, bootstrap procedures are used.*

Now, a second set of assumptions, which should be added to the first one in order to obtain the convergence of $\widehat{m}^*_{hb}(\chi)$ and the negligible nature of the third term to which we have referred previously, is stated:

$$\text{The function } \mathbb{E}(|Y||\boldsymbol{\chi} = \cdot) \text{ is continuous on a neighbourhood of } \chi, \text{ and } \sup_{d(\chi_1,\chi)<\delta} \mathbb{E}(|Y|^q|\boldsymbol{\chi} = \chi_1) < \infty \text{ for some } \delta > 0; \forall q \geq 1 \tag{4.14}$$

$$\forall(\chi_1, s) \text{ in a neighbourhood of } (\chi, 0),\ \varphi_{\chi_1}(0) = 0,\ \varphi'_{\chi_1}(s) \text{ exists, } \varphi'_{\chi_1}(0) \neq 0 \text{ and } \varphi'_{\chi_1}(s) \text{ is uniformly Lipschitz continuous of order } 0 < \alpha \leq 1 \text{ in } (\chi_1, s) \tag{4.15}$$

$$\forall \chi_1 \in \mathcal{H},\ F_{\chi_1}(0) = 0 \text{ and } F_{\chi_1}(t)/F_{\chi}(t) \text{ is Lipschitz continuous of order } \alpha \text{ in } \chi_1,\ \text{uniformly in } t \text{ in a neighbourhood of } 0 \text{ (with } \alpha \text{ introduced in (4.15))} \tag{4.16}$$

$$\forall \chi_1 \in \mathcal{H} \text{ and } \forall s \in [0,1],\ \tau_{0\chi_1}(s) \text{ exists, } \sup_{\chi_1 \in H, s\in[0,1]} |\tau_{h\chi_1}(s) - \tau_{0\chi_1}(s)| = o(1), M_{0\chi} > 0,\ M_{2\chi} > 0,\ \inf_{d(\chi_1,\chi)<\varepsilon} M_{1\chi} > 0 \text{ for some } \varepsilon > 0, \text{ and } M_{k\chi_1} \text{ is Lipschitz continuous of order } \alpha \text{ for } k = 0, 1, 2 \text{ (with } \alpha \text{ introduced in (4.15))} \tag{4.17}$$

$$\forall n\ \exists r_n \geq 1,\ l_n > 0 \text{ and curves } \chi_{1n}, \ldots, \chi_{r_n n} \text{ such that } B(\chi, h) \subset \cup_{k=1}^{r_n} B(\chi_{kn}, l_n), \text{ with } r_n = O(n^{b/h}) \text{ and } l_n = o(b(nF_{\chi}(h))^{-1/2}), \inf_{d(\chi_1,\chi)<\varepsilon} M_{1\chi} > 0 \text{ for some } \varepsilon > 0, \text{ and} M_{k\chi_1} \text{ is Lipschitz continuous of order } \alpha \text{ for } k = 0, 1, 2 \text{ (with } \alpha \text{ introduced in (4.15))} \tag{4.18}$$

$$\max\{b, h/b, b^{1+\alpha}(nF_{\chi}(h))^{1/2}, (F_{\chi}(h)/F_{\chi}(b)) \log n, n^{1/p} F_{\chi}(h)^{1/2} \log n\} = o(1), \max\{bh^{\alpha-1}, F_{\chi}(b)^{-1}h/b\} = O(1) \text{ and } \lim_{n\to\infty} F_{\chi}(b+h)/F_{\chi}(b) = 1 \text{ (with } p \text{ and } \alpha \text{ introduced in (4.15) and (4.7), respectively)} \tag{4.19}$$

$$a > 4.5 \text{ (with } a \text{ introduced in (4.11))}. \tag{4.20}$$

Remark 3 *With the exception of Assumption (4.20),* $n^{1/p}F_\chi(h)^{1/2}\log n = o(1)$ *and* $F_\chi(b)^{-1}h/b = O(1)$ *in Assumption (4.19), all the other assumptions were used in Ferraty, Van Keilegom and Vieu (2010) to obtain the validity of the bootstrap in the independent case. Note that Assumption (4.20) is introduced here to manage the dependence (remember that the case dealt with in Ferraty, Van Keilegom and Vieu, 2010, was that of independent data). Note also that such assumption together with* $n^{1/p}F_\chi(h)^{1/2}\log n = o(1)$ *and* $F_\chi(b)^{-1}h/b = O(1)$ *allow to apply the Lemma 21 in Appendix A, from Aneiros-Pérez and Vieu (2008) (see the last part in the proof of our Theorem). Finally, a special attention has to be given to the part of Assumption (4.18) related to the balls, which is only necessary to make use of the results of uniform consistency of nonparametric regression smoothers. Therefore, it can be weakened by changing it into any other kind of assumptions insuring such uniform consistency (see, for instance, Ferraty, Laksaci, Tadj and Vieu (2010) for alternative assumptions in the case of independent data; although we do not know examples for dependent data, our feeling is that such kind of alternative assumptions should also work in the case of dependent data).*

4.3.2 Asymptotic result

Let $P^{\mathcal{S}}$ denote probability, conditionally on the sample $\mathcal{S}$, and let us suppose that χ is a fixed element of the space $\mathcal{H}$.

Theorem 4 *Under assumptions (4.4)-(4.20), if the model is homoscedastic (i.e.* $\sigma_\varepsilon^2(\cdot) = \sigma_\varepsilon^2$*) for the naïve bootstrap procedure, one has that*

$$\sup_{y\in\mathbb{R}}\left|P^{\mathcal{S}}\left(\sqrt{nF_\chi(h)}(\widehat{m}^*_{hb}(\chi)-\widehat{m}_b(\chi))\le y\right) - P\left(\sqrt{nF_\chi(h)}(\widehat{m}_h(\chi)-m(\chi))\le y\right)\right| \to 0 \ a.s.$$

Theorem 5 *Under assumptions (4.4)-(4.20), for the wild bootstrap procedure, one has that*

$$\sup_{y\in\mathbb{R}}\left|P^{\mathcal{S}}\left(\sqrt{nF_\chi(h)}(\widehat{m}^*_{hb}(\chi)-\widehat{m}_b(\chi))\le y\right) - P\left(\sqrt{nF_\chi(h)}(\widehat{m}_h(\chi)-m(\chi))\le y\right)\right| \to 0 \ a.s.$$

Remark 6 *Theorems 4 and 5 extend Theorem 1 in Ferraty, Van Keilegom and Vieu (2010) from the independent case to the dependent one. Its main practical usefulness is related to the building of confidence intervals for* $m(\chi)$ *in a context of dependent data. As noted above, due to the (most of the times) difficulty in estimating the constants involved in the standardization of* $\widehat{m}_h(\chi)$*, the asymptotic distribution of the true error* $m(\chi)-\widehat{m}_h(\chi)$ *could be useless to construct the desired confidence interval. Nevertheless, from these theorems one has that the* α*-quantile,* $q_\alpha(\chi)$*, of* $m(\chi)-\widehat{m}_h(\chi)$ *can be approximated by the* α*-quantile,* $q^*_\alpha(\chi)$*, obtained from the distribution of the bootstrapped errors* $\widehat{m}_b(\chi)-\widehat{m}^*_{hb}(\chi)$*. Then, because one can generate more and more replicates (say* B *replicates) of such bootstrapped error, the percentile method (for instance) allows to obtain a very good approximation, say* $q^{*,B}_\alpha(\chi)$*, of* $q^*_\alpha(\chi)$*. Finally, one can build confidence intervals for* $m(\chi)$ *approximating the* $(1-\alpha)$*-confidence interval* $I_{\chi,1-\alpha}=(\widehat{m}_h(\chi)+q_{\alpha/2}(\chi),\widehat{m}_h(\chi)+q_{1-\alpha/2}(\chi))$ *by* $I^{*,B}_{\chi,1-\alpha}=(\widehat{m}_h(\chi)+q^{*,B}_{\alpha/2}(\chi),\widehat{m}_h(\chi)+q^{*,B}_{1-\alpha/2}(\chi))$*. See Section 4.4 and Section 4.5 for details on the algorithms developed to construct the confidence intervals with simulated data of real electricity data, respectively.*

4.3.3 Proofs

Preliminary proofs

Before proving our main results (Theorems 4 and 5), some preliminary lemmas are stated to be used in that proof. First, let us denote

$$J_\chi=\frac{\int_0^1 tK(t)dP^{d(\boldsymbol{\chi},\chi)/h}(t)}{\int_0^1 K(t)dP^{d(\boldsymbol{\chi},\chi)/h}(t)} \text{ and } \widehat{m}_h(\chi)=\frac{\widehat{g}_h(\chi)}{\widehat{f}_h(\chi)},$$

where

$$\widehat{g}_h(\chi)=\frac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i,\chi)/h)Y_i}{nF_\chi(h)} \text{ and } \widehat{f}_h(\chi)=\frac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i,\chi)/h)}{nF_\chi(h)}.$$

Furthermore, it is used the notation

$$A_1=-\mathbb{E}\left(\widehat{g}_h(\chi)(\widehat{f}_h(\chi)-\mathbb{E}(\widehat{f}_h(\chi)))\right)$$

and

$$A_2=\mathbb{E}\left((\widehat{f}_h(\chi)-\mathbb{E}(\widehat{f}_h(\chi))^2\widehat{m}_h(\chi)\right).$$

The auxiliary lemmas in Appendix A will be used, coming from different studies available in the literature that establish the background of our proposal, but also the following lemmas.

Lemma 7 *Under assumptions (4.4)-(4.10) and (4.12)one has that*

$$A_1 = O((nF_\chi(h))^{-1}), \text{ and } A_2 = O((nF_\chi(h))^{-1}).$$

Proof. The proof of this lemma is consequence of Lemma 27 in Appendix A. On the one hand, one has that $A_1 = -Cov\left(\widehat{g}_h(\chi), \widehat{f}_h(\chi)\right)$; so, from Lemma 27 one obtains that $A_1 = O((nF_\chi(h))^{-1})$. On the other hand,

$$\begin{aligned}
|A_2| &= \left|\mathbb{E}\left((\widehat{f}_h(\chi) - \mathbb{E}(\widehat{f}_h(\chi))^2 \widehat{m}_h(\chi)\right)\right| \\
&= \left|\mathbb{E}\left(\mathbb{E}\left((\widehat{f}_h(\chi) - \mathbb{E}(\widehat{f}_h(\chi))^2 \widehat{m}_h(\chi)|\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n\right)\right)\right| \\
&= \left|\mathbb{E}\left((\widehat{f}_h(\chi) - \mathbb{E}(\widehat{f}_h(\chi))^2 \mathbb{E}\left(\widehat{m}_h(\chi)|\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n\right)\right)\right| \\
&= \left|\mathbb{E}\left((\widehat{f}_h(\chi) - \mathbb{E}(\widehat{f}_h(\chi))^2 \frac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h) m(\boldsymbol{\chi}_i)}{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h)}\right)\right| \\
&\leq CVar\left(\widehat{f}_h(\chi)\right),
\end{aligned}$$

where the inequality is a consequence of assumptions (4.4) and (4.10). The proof concludes by using, again, Lemma 27. □

Lemma 8 *Under assumptions (4.4)-(4.10) and (4.12) one has that*

$$\mathbb{E}\left(\widehat{m}_h(\chi)\right) - m(\chi) = \varphi'_\chi(0)\frac{M_{0\chi}}{M_{1\chi}}h + O\left(\frac{1}{nF_\chi(h)}\right) + o(h). \tag{4.21}$$

If in addition Assumption (4.13) holds, then

$$Var\left(\widehat{m}_h(\chi)\right) = \frac{1}{nF_\chi(h)}\frac{M_{2\chi}}{M_{1\chi}^2}\sigma_\varepsilon^2 + o\left(\frac{1}{nF_\chi(h)}\right). \tag{4.22}$$

Proof of (4.21). This proof is based on the decomposition

$$\mathbb{E}\left(\widehat{m}_h(\chi)\right) = \frac{\mathbb{E}\left(\widehat{g}_h(\chi)\right)}{\mathbb{E}\left(\widehat{f}_h(\chi)\right)} + \frac{A_1}{\left(\mathbb{E}\left(\widehat{f}_h(\chi)\right)\right)^2} + \frac{A_2}{\left(\mathbb{E}\left(\widehat{f}_h(\chi)\right)\right)^2}. \tag{4.23}$$

In fact, using Lemmas 17, 18 and 7 in Appendix A, the proof of (4.21) is easily obtained following the same steps as those in the proof of (2) in Ferraty et al. (2007). □

Proof of (4.22). In the case of independent data, this term corresponds to (3) in Theorem 1 from Ferraty et al. (2007). The extension to our case, when dealing with dependent data, has been studied in Theorem 7.3.1 in Delsol (2008) giving the same expression for the variance in (4.22). □

Lemma 9 *Under assumptions (4.4)–(4.5) and (4.14)–(4.19), one has that:*

$$\frac{Var^{\mathcal{S}}\left(\widehat{m}^{*}_{hb}(\chi)\right)}{Var\left(\widehat{m}_{h}(\chi)\right)} \longrightarrow 1 \ a.s. \tag{4.24}$$

Proof This is true due to Lemmas 19 and 27 in Appendix A and Lemma 8 above. As can be seen in Ferraty, Van Keilegom and Vieu (2010), when data in $\mathcal{S}$ are independent, the proof of (4.24) corresponds to its Lemma 3 and it is based on both the type of bootstrap procedure used and Lemmas 4 and 5 and Theorem 1 in Ferraty et al. (2007). On the one hand, because the random errors ε_i in our model (4.1) are independent, the same bootstrap procedures as in Ferraty et al. (2007) are considered. On the other hand, Lemmas 19, 27 and 8 give the same results as Lemmas 4 and 5 and Theorem 1 in Ferraty et al. (2007), respectively, but under dependence conditions on $\mathcal{S}$. These facts allow to follow step by step the proof of Lemma 3 in Ferraty, Van Keilegom and Vieu (2010) for the independent case and to conclude that (4.24) holds under our dependence conditions on $\mathcal{S}$.

Proof of Theorem 4

The proof of Theorem 4 follows the same steps as those of Theorem 1 in Ferraty, Van Keilegom and Vieu (2010), where the case of an independent sample $\mathcal{S}$ was dealt. Thus, for the sake of brevity, it will be focused on the issues where the dependence affects.

First, let $\mathbb{E}^{\mathcal{S}}$ and $Var^{\mathcal{S}}$ denote expectation and variance, respectively, conditionally on the sample $\mathcal{S}$, while Φ is the standard normal distribution function. Let us write

$$\begin{aligned} P^{\mathcal{S}}\left(\sqrt{nF_{\chi}(h)}(\widehat{m}^{*}_{hb}(\chi)-\widehat{m}_{b}(\chi))\leq y\right) - \\ P\left(\sqrt{nF_{\chi}(h)}(\widehat{m}_{h}(\chi)-m(\chi))\leq y\right) = \\ T_1(y)+T_2(y)+T_3(y), \end{aligned} \tag{4.25}$$

where

$$\begin{aligned} T_1(y) = P^{\mathcal{S}}\left(\sqrt{nF_{\chi}(h)}(\widehat{m}^{*}_{hb}(\chi)-\widehat{m}_{b}(\chi))\leq y\right) - \\ \Phi\left(\frac{y-\sqrt{nF_{\chi}(h)}\left(\mathbb{E}^{\mathcal{S}}\left(\widehat{m}^{*}_{hb}(\chi)\right)-\widehat{m}_{b}(\chi)\right)}{\sqrt{nF_{\chi}(h)Var^{\mathcal{S}}\left(\widehat{m}^{*}_{hb}(\chi)\right)}}\right), \end{aligned}$$

$$T_2(y) = \Phi\left(\frac{y - \sqrt{nF_\chi(h)}\left(\mathbb{E}^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right) - \widehat{m}_b(\chi)\right)}{\sqrt{nF_\chi(h)Var^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}}\right) - \Phi\left(\frac{y - \sqrt{nF_\chi(h)}\left(\mathbb{E}\left(\widehat{m}_h(\chi)\right) - m(\chi)\right)}{\sqrt{nF_\chi(h)Var\left(\widehat{m}_h(\chi)\right)}}\right)$$

and

$$T_3(y) = \Phi\left(\frac{y - \sqrt{nF_\chi(h)}\left(\mathbb{E}\left(\widehat{m}_h(\chi)\right) - m(\chi)\right)}{\sqrt{nF_\chi(h)Var\left(\widehat{m}_h(\chi)\right)}}\right) - P\left(\sqrt{nF_\chi(h)}(\widehat{m}_h(\chi) - m(\chi)) \leq y\right).$$

Each one of these terms is going to be studied separately.

First, term $T_3(y)$ is analysed. This term deals with the real regression function (m) and its kernel estimation ($\widehat{m}_h$), without any intervention of the bootstrap procedure. Thus, one can apply here some results available in the literature about the FNP regression model under dependence.

Theorem 28 in Appendix A establishes the asymptotic normality of the same estimator, under assumptions (4.4)-(4.13):

$$\frac{M_{1_\chi}}{\sqrt{M_{2_\chi}\sigma^2_\varepsilon}}\sqrt{n\widehat{F}_\chi(h)}(\widehat{m}_h(\chi) - m(\chi) - B_n) \to N(0,1), \tag{4.26}$$

where $B_n = \varphi'_\chi(0)\frac{M_{0\chi}}{M_{1\chi}}h$ and $\widehat{F}_\chi(h) = 1/n\sum_{i=1}^n 1_{[d(\chi,\chi_i),\infty)}(t)$.

Lemma 8, which indicates the bias and variance of the estimator $\widehat{m}_h$, allows to develop the following expression:

$$\frac{\widehat{m}_h(\chi) - \mathbb{E}\left(\widehat{m}_h(\chi)\right)}{\sqrt{Var\left(\widehat{m}_h(\chi)\right)}} = \frac{\widehat{m}_h(\chi) - m(\chi) - B_n + O\left(\frac{1}{nF_\chi(h)}\right) + o(h)}{\sqrt{\frac{1}{nF_\chi(h)}\frac{M_{2\chi}}{M^2_{1\chi}}\sigma^2_\varepsilon + o\left(\frac{1}{nF_\chi(h)}\right)}} =$$

$$\sqrt{nF_\chi(h)}\,\frac{\widehat{m}_h(\chi) - m(\chi) - B_n + O\left(\frac{1}{nF_\chi(h)}\right) + o(h)}{\sqrt{\frac{M_{2\chi}}{M^2_{1\chi}}\sigma^2_\varepsilon + o\left(1\right)}} =$$

$$\frac{\sqrt{nF_\chi(h)}}{\sqrt{n\widehat{F}_\chi(h)}}\sqrt{n\widehat{F}_\chi(h)}\,\frac{\widehat{m}_h(\chi) - m(\chi) - B_n}{\sqrt{\frac{M_{2\chi}}{M^2_{1\chi}}\sigma^2_\varepsilon + o\left(1\right)}} + \sqrt{nF_\chi(h)}\,\frac{O\left(\frac{1}{nF_\chi(h)}\right) + o(h)}{\sqrt{\frac{M_{2\chi}}{M^2_{1\chi}}\sigma^2_\varepsilon + o\left(1\right)}}$$

where, by Lemma 2.6 in Delsol (2009) (which gives $\frac{\sqrt{nF_\chi(h)}}{\sqrt{n\widehat{F}_\chi(h)}} \xrightarrow{d} 1$) and (4.26), the convergence

$$\frac{\sqrt{nF_\chi(h)}}{\sqrt{n\widehat{F}_\chi(h)}}\sqrt{n\widehat{F}_\chi(h)}\ \frac{\widehat{m}_h(\chi)-m(\chi)-B_n}{\sqrt{\frac{M_{2\chi}}{M_{1\chi}^2}\sigma_\varepsilon^2+o(1)}} \xrightarrow{d} N(0,1)$$

holds. In addition from assumption (4.9) one obtains:

$$\sqrt{nF_\chi(h)}\ \frac{O\left(\frac{1}{nF_\chi(h)}\right)+o(h)}{\sqrt{\frac{M_{2\chi}}{M_{1\chi}^2}\sigma_\varepsilon^2+o(1)}} \longrightarrow 0,$$

which allows to conclude:

$$\frac{\widehat{m}_h(\chi)-\mathbb{E}\left(\widehat{m}_h(\chi)\right)}{\sqrt{Var\left(\widehat{m}_h(\chi)\right)}} \xrightarrow{d} N(0,1). \tag{4.27}$$

Therefore,

$$P\left(\sqrt{nF_\chi(h)}(\widehat{m}_h(\chi)-m(\chi))\le y\right)=$$
$$P\left(\frac{\widehat{m}_h(\chi)-\mathbb{E}\left(\widehat{m}_h(\chi)\right)}{\sqrt{Var\left(\widehat{m}_h(\chi)\right)}}\le \frac{y-\sqrt{nF_\chi(h)}(\mathbb{E}\left(\widehat{m}_h(\chi)\right)-m(\chi))}{\sqrt{nF_\chi(h)Var\left(\widehat{m}_h(\chi)\right)}}\right)$$

and thus,

$$\begin{aligned}
T_3(y) &= \Phi\left(\frac{y-\sqrt{nF_\chi(h)}\left(\mathbb{E}\left(\widehat{m}_h(\chi)\right)-m(\chi)\right)}{\sqrt{nF_\chi(h)Var\left(\widehat{m}_h(\chi)\right)}}\right)- \\
&\quad P\left(\sqrt{nF_\chi(h)}(\widehat{m}_h(\chi)-m(\chi))\le y\right)= \\
&\quad \Phi\left(\frac{y-\sqrt{nF_\chi(h)}\left(\mathbb{E}\left(\widehat{m}_h(\chi)\right)-m(\chi)\right)}{\sqrt{nF_\chi(h)Var\left(\widehat{m}_h(\chi)\right)}}\right)- \\
&\quad P\left(\frac{\widehat{m}_h(\chi)-\mathbb{E}\left(\widehat{m}_h(\chi)\right)}{\sqrt{Var\left(\widehat{m}_h(\chi)\right)}}\le \frac{y-\sqrt{nF_\chi(h)}(\mathbb{E}\left(\widehat{m}_h(\chi)\right)-m(\chi))}{\sqrt{nF_\chi(h)Var\left(\widehat{m}_h(\chi)\right)}}\right)= \\
&\quad \Phi\left(X_y\right)-P\left(\frac{\widehat{m}_h(\chi)-\mathbb{E}\left(\widehat{m}_h(\chi)\right)}{\sqrt{Var\left(\widehat{m}_h(\chi)\right)}}\le X_y\right)\longrightarrow 0,
\end{aligned}$$

where $X_y=\frac{y-\sqrt{nF_\chi(h)}(\mathbb{E}(\widehat{m}_h(\chi))-m(\chi))}{\sqrt{nF_\chi(h)Var(\widehat{m}_h(\chi))}}$ and the convergence to zero is attained applying the asymptotic normality in (4.27). In short, it is obtained that:

$$T_3(y)\longrightarrow 0 \text{ for any fixed value of } y. \tag{4.28}$$

Then, the term $T_1(y)$ is studied. This is the equivalent to $T_3(y)$, but in this case one deals with kernel estimator $\widehat{m}_b$ and the bootstrap estimator $\widehat{m}^*_{hb}$ and works conditionally on the sample $\mathcal{S}$. As in the previous case, it is sufficient to prove that

$$\frac{\widehat{m}^*_{hb}(\chi) - \mathbb{E}^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}{\sqrt{Var^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}} \xrightarrow{d} N(0,1), \text{ a.s., conditionally on } \mathcal{S}. \tag{4.29}$$

First, using the decomposition

$$\widehat{m}^*_{hb}(\chi) = \widehat{g}^*_{hb}(\chi)/\widehat{f}_h(\chi),$$

where

$$\widehat{g}^*_{hb}(\chi) = \frac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i,\chi)/h)Y_i^*}{nF_\chi(h)},$$

one obtains the following equivalence:

$$\frac{\widehat{m}^*_{hb}(\chi) - \mathbb{E}^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}{\sqrt{Var^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}} = \frac{\widehat{g}^*_{hb}(\chi)/\widehat{f}_h(\chi) - \mathbb{E}^{\mathcal{S}}\left(\widehat{g}^*_{hb}(\chi)/\widehat{f}_h(\chi)\right)}{\sqrt{Var^{\mathcal{S}}\left(\widehat{g}^*_{hb}(\chi)/\widehat{f}_h(\chi)\right)}} =$$

$$\frac{(1/\widehat{f}_h(\chi))\widehat{g}^*_{hb}(\chi) - 1/\widehat{f}_h(\chi)\mathbb{E}^{\mathcal{S}}\left(\widehat{g}^*_{hb}(\chi)\right)}{\sqrt{1/\widehat{f}_h(\chi)^2 Var^{\mathcal{S}}\left(\widehat{g}^*_{hb}(\chi)\right)}} = \frac{\widehat{g}^*_{hb}(\chi) - \mathbb{E}^{\mathcal{S}}\left(\widehat{g}^*_{hb}(\chi)\right)}{\sqrt{Var^{\mathcal{S}}\left(\widehat{g}^*_{hb}(\chi)\right)}}. \tag{4.30}$$

That is, it is enough to study the asymptotic normality of $\widehat{g}^*_{hb}(\chi)$, properly standardized. As this term, conditionally on $\mathcal{S}$ can be seen as a sum of independent terms, one can prove its asymptotic normality by means of the Liapunov's condition. That is, if one proves that:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathbb{E}^{\mathcal{S}}|(nF_\chi(h))^{-1}(Y_i^* - \mathbb{E}^{\mathcal{S}}(Y_i^*))K(d(\boldsymbol{\chi}_i,\chi)/h)|^3}{(\sqrt{Var^{\mathcal{S}}(\widehat{g}^*_{hb}(\chi))})^3} = 0 \text{ a.s.} \tag{4.31}$$

then, one gets:

$$\frac{\widehat{g}^*_{hb}(\chi) - \mathbb{E}^{\mathcal{S}}\left(\widehat{g}^*_{hb}(\chi)\right)}{\sqrt{Var^{\mathcal{S}}\left(\widehat{g}^*_{hb}(\chi)\right)}} \xrightarrow{d} N(0,1) \text{ a.s.}$$

and, as a consequence of (4.30), also (4.29) will be fulfilled.

First, the numerator of the condition (4.31) is studied using the definition of $\widehat{g}^*_{hb}(\chi)$:

$$\sum_{i=1}^{n} \mathbb{E}^{\mathcal{S}}|(nF_\chi(h))^{-1}(Y_i^* - \mathbb{E}^{\mathcal{S}}(Y_i^*))K(d(\boldsymbol{\chi}_i,\chi)/h)|^3 =$$

$$(nF_\chi(h))^{-3}\sum_{i=1}^{n} K^3(d(\boldsymbol{\chi}_i,\chi)/h)\mathbb{E}^{\mathcal{S}}|(Y_i^* - \mathbb{E}^{\mathcal{S}}(Y_i^*))|^3 =$$

$$(nF_\chi(h))^{-3}\mathbb{E}^{\mathcal{S}}|(Y_1^* - \mathbb{E}^{\mathcal{S}}(Y_1^*))|^3\sum_{i=1}^{n} K^3(d(\boldsymbol{\chi}_i,\chi)/h) = O((nF_\chi(h))^{-2}) \text{ a.s.}$$

Note that it has been used the fact that, conditioned to $\mathcal{S}$, $Y_i^*(i=1,\ldots,n)$ are identically distributed, as well as $\mathbb{E}^{\mathcal{S}}|Y_1^* - \mathbb{E}^{\mathcal{S}}(Y_1^*)|^3 = O(1)$ a.s. and $\sum_{i=1}^n K^3(d(\boldsymbol{\chi}_i,\chi)/h)) = O(nF_\chi(h))$ a.s.

Let's prove that $\mathbb{E}^{\mathcal{S}}|(Y_1^* - \mathbb{E}^{\mathcal{S}}(Y_1^*))|^3$ is bounded a.s. On the one hand, one has that

$$\mathbb{E}^{\mathcal{S}}|(Y_1^* - \mathbb{E}^{\mathcal{S}}(Y_1^*))|^3 = \mathbb{E}^{\mathcal{S}}|\varepsilon_1^*|^3 = n^{-1}\sum_{i=1}^{n}\left|\widehat{\varepsilon}_{i,b} - \overline{\widehat{\varepsilon}}_b\right|^3. \tag{4.32}$$

On the other hand, it verifies that

$$\widehat{\varepsilon}_{i,b} = \varepsilon_i + m(\boldsymbol{\chi}_i) - \widehat{m}_b(\boldsymbol{\chi}_i) = \varepsilon_i + o(1) \text{ a.s. uniformly in i.} \tag{4.33}$$

(Note that from the results in Ferraty and Vieu (2004) one has the uniform convergence of $\widehat{m}_b$).
In addition, the Strong Law of Large numbers gives

$$n^{-1}\sum_{i=1}^{n}\varepsilon_i \to 0 \text{ a.s. and } n^{-1}\sum_{i=1}^{n}|\varepsilon_i|^3 \to \mathbb{E}(|\varepsilon_i|^3)\text{a.s.}$$

Finally, (4.32), (4.33) and (4.3.3) give that $\mathbb{E}^{\mathcal{S}}|(Y_1^* - \mathbb{E}^{\mathcal{S}}(Y_1^*))|^3$ is bounded a.s.

Now we prove that

$$\sum_{i=1}^{n} K^3(d(\boldsymbol{\chi}_i,\chi)/h)) = O(nF_\chi(h)) \text{ a.s.} \tag{4.34}$$

For that, the following result plays a main role:

$$\#\{i=1,\ldots,n; \chi \in B(\boldsymbol{\chi}_i,h)\} = O(nF_\chi(h)) \text{ a.s.} \tag{4.35}$$

(For details on (4.35), consider (31) in Ferraty and Vieu (2004) with uniform kernel in $[0,1]$). Finally, (4.35) together with the fact that K is bounded with support $[0,1]$, give (4.34).

In order to study the denominator of the Condition (4.31), it is enough to consider Lemma 9 and (4.22) in Lemma 8 to get that

$$Var^{\mathcal{S}}(\widehat{m}^*_{hb}(\chi)) = O((nF_\chi(h))^{-1}) \text{ a.s.}$$

and, as a consequence:

$$(Var^{\mathcal{S}}(\widehat{g}^*_{hb}(\chi)))^{3/2} = O((nF_\chi(h))^{-3/2}) \text{ a.s.}$$

Using together the results for numerator and denominator one obtains that the expression in (4.31) is $O((nF_\chi(h))^{-1/2}) = o(1)$ a.s. by assumption (4.9).

As a consequence, it is obtained that

$$T_1(y) \longrightarrow 0 \text{ a.s. for any fixed value of } y. \tag{4.36}$$

Now, from (4.28), (4.36), noting that the uniform convergence for any $y \in \mathbb{R}$ follows from Polya's theorem (see Theorem 26 in Appendix A) together with the continuity of the function Φ, one has that:

$$\sup_{y\in\mathbb{R}} |T_1(y)| + \sup_{y\in\mathbb{R}} |T_3(y)| \to 0 \text{ a.s.} \tag{4.37}$$

Finally, it remains to study the term $T_2(y)$. Using the fact that, for any $a \in \mathbb{R}$ and $c > 0$,

$$\sup_{y\in\mathbb{R}} |\Phi(a + cy) - \Phi(y)| \leq |a| + \max\{c, c^{-1}\} - 1,$$

and considering

$$a = \frac{\sqrt{nF_\chi(h)}\left(\mathbb{E}\left(\widehat{m}_h(\chi)\right) - m(\chi) - \mathbb{E}^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right) + \widehat{m}_b(\chi)\right)}{\sqrt{nF_\chi(h) Var^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}}$$

and

$$c = \sqrt{\frac{Var\left(\widehat{m}_h(\chi)\right)}{Var^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}},$$

one has that

$$\sup_{y\in\mathbb{R}}|T_2(y)| \leq \left|\frac{\sqrt{nF_\chi(h)}\left(\mathbb{E}\left(\widehat{m}_h(\chi)\right)-m(\chi)-\mathbb{E}^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)+\widehat{m}_b(\chi)\right)}{\sqrt{nF_\chi(h)Var^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}}\right| + \max\left\{\sqrt{\frac{Var\left(\widehat{m}_h(\chi)\right)}{Var^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}}, \sqrt{\frac{Var^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)}{Var\left(\widehat{m}_h(\chi)\right)}}\right\} - 1. \quad (4.38)$$

From (4.38) one has that, in order to obtain

$$\sup_{y\in\mathbb{R}}|T_2(y)| \to 0 \text{ a.s.}, \quad (4.39)$$

it is sufficient to prove the next expression related to the numerator of a, as all the other terms are negligible by Lemma 9 (for the term c) together with the statement (4.22) in Lemma 8 (for the denominator in a),

$$\left|\sqrt{nF_\chi(h)}\left(\mathbb{E}\left(\widehat{m}_h(\chi)\right)-m(\chi)-\mathbb{E}^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)+\widehat{m}_b(\chi)\right)\right| \to 0 \text{ a.s.} \quad (4.40)$$

For that, let us consider the following decomposition:

$$\begin{aligned}
&\mathbb{E}^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right)-\widehat{m}_b(\chi) = \mathbb{E}^{\mathcal{S}}\left(\widehat{g}^*_{hb}(\chi)/\widehat{f}_h(\chi)-\widehat{m}_b(\chi)\right) = \\
&\frac{(nF_\chi(h))^{-1}}{\widehat{f}_h(\chi)}\sum_{i=1}^{n}\left(\widehat{m}_b(\boldsymbol{\chi}_i)-\widehat{m}_b(\chi)\right)K(d(\boldsymbol{\chi}_i,\chi)/h) = \\
&\frac{(nF_\chi(h))^{-1}}{\widehat{f}_h(\chi)}\sum_{i=1}^{n}\left(\widehat{m}_b(\boldsymbol{\chi}_i)-\widehat{m}_b(\chi)\pm\mathbb{E}(\widehat{m}_b(\boldsymbol{\chi}_i))\pm\right. \\
&\left.\mathbb{E}(\widehat{m}_b(\chi))\pm m(\boldsymbol{\chi}_i)\pm m(\chi)\right)K(d(\boldsymbol{\chi}_i,\chi)/h) = \\
&= U_1+U_2+U_3, \quad (4.41)
\end{aligned}$$

where

$$U_1 = \frac{(nF_\chi(h))^{-1}}{\widehat{f}_h(\chi)}\sum_{i=1}^{n}\left(\widehat{m}_b(\boldsymbol{\chi}_i)-\widehat{m}_b(\chi)-\mathbb{E}\left(\widehat{m}_b(\boldsymbol{\chi}_i)\right)+\mathbb{E}\left(\widehat{m}_b(\chi)\right)\right)K(d(\boldsymbol{\chi}_i,\chi)/h),$$

$$U_2 = \frac{(nF_\chi(h))^{-1}}{\widehat{f}_h(\chi)}\sum_{i=1}^{n}\left(\mathbb{E}\left(\widehat{m}_b(\boldsymbol{\chi}_i)\right)-\mathbb{E}\left(\widehat{m}_b(\chi)\right)-m(\boldsymbol{\chi}_i)+m(\chi)\right)K(d(\boldsymbol{\chi}_i,\chi)/h)$$

and

$$U_3 = \frac{(nF_\chi(h))^{-1}}{\widehat{f}_h(\chi)} \sum_{i=1}^{n} (m(\boldsymbol{\chi}_i) - m(\chi))\, K(d(\boldsymbol{\chi}_i, \chi)/h).$$

By means of similar techniques as those used in Ferraty and Vieu (2004) to obtain the rate of convergence of $\widehat{m}_h(\chi)$, it is easy to show that

$$U_3 = \mathbb{E}(\widehat{m}_h(\chi)) - m(\chi) + o\left((nF_\chi(h))\right)^{-1/2} \text{ a.s.} \tag{4.42}$$

Next step will be to prove that

$$U_2 = o\left((nF_\chi(h))\right)^{-1/2} \text{ a.s.} \tag{4.43}$$

Taking into account that $\boldsymbol{\chi}_i$ are identically distributed, together with the convergence of $\widehat{f}_h(\chi)$ (Lemma 2.6 in Delsol, 2009) and (4.35), one has that (4.43) is true if one checks:

$$\sup_{d(\chi_1,\chi)\leq h} |\mathbb{E}(\widehat{m}_b(\chi_1)) - \mathbb{E}(\widehat{m}_b(\chi)) - m(\chi_1) + m(\chi)| = o\left((nF_\chi(h))\right)^{-1/2} \text{ a.s.} \tag{4.44}$$

From a slight modification in the proof of Lemma 8 one obtains the next expression for (4.44):

$$\begin{aligned} &|\mathbb{E}(\widehat{m}_b(\chi_1)) - \mathbb{E}(\widehat{m}_b(\chi)) - m(\chi_1) + m(\chi)| = \\ &|\varphi'_{\chi_1}(0)\frac{M_{0\chi_1}}{M_{1\chi_1}} - \varphi'_\chi(0)\frac{M_{0\chi}}{M_{1\chi}}|b + O\left(\frac{1}{nF_{\chi_1}(b)}\right) + O\left(\frac{1}{nF_\chi(b)}\right) + \\ &O(b^{1+\alpha}). \end{aligned} \tag{4.45}$$

Note that last term $O(b^{1+\alpha})$ is used instead of $o(b)$ in Lemma 8. This is possible due to the next development considering the decomposition (4.23) and following the proof of Lemma 17 in Appendix A

$$\frac{\mathbb{E}\left(\widehat{g}_b(\chi)\right)}{\mathbb{E}\left(\widehat{f}_b(\chi)\right)} - m(\chi) = (\ldots) = \frac{\mathbb{E}\left(\varphi_\chi(d(\boldsymbol{\chi}_i,\chi))K(d(\boldsymbol{\chi}_i,\chi)/b)\right)}{\mathbb{E}\left(K(d(\boldsymbol{\chi}_i,\chi)/b)\right)}, \tag{4.46}$$

but, focusing now on the numerator of the last expression:

$$\int \varphi_\chi(t)K(t/b)dP^{d(\boldsymbol{\chi}_i,\chi)}(t) = \int \varphi_\chi(bt)K(t)dP^{d(\boldsymbol{\chi}_i,\chi)/b}(t),$$

where one can apply Taylor expansion of order zero in a neighbourhood of zero:

$$\begin{aligned} &\exists x' \in (0, bt); \varphi_\chi(bt) = \varphi_\chi(0) + \frac{\varphi'_\chi(x')}{1}(bt - 0) = \varphi'_\chi(x')bt = \\ &\varphi'_\chi(x')bt \pm \varphi'_\chi(0)bt = \varphi'_\chi(0)bt + (\varphi'_\chi(x') - \varphi'_\chi(0))bt. \end{aligned}$$

Applying Assumption (4.15) one has that

$$|\varphi'_\chi(x') - \varphi'_\chi(0)|bt \leq c|x' - 0|^\alpha bt \leq c|bt|^\alpha bt = c(bt)^{\alpha+1}$$

and thus:

$$\begin{aligned}
&\int \varphi_\chi(bt)K(t)dP^{d(\boldsymbol{\chi}_i,\chi)/b}(t) = \\
&\int \varphi'_\chi(0)btK(t)dP^{d(\boldsymbol{\chi}_i,\chi)/b}(t) + \int (\varphi'_\chi(x') - \varphi'_\chi(0))btK(t)dP^{d(\boldsymbol{\chi}_i,\chi)/b}(t) = \\
&\int \varphi'_\chi(0)btK(t)dP^{d(\boldsymbol{\chi}_i,\chi)/b}(t) + O(b^{\alpha+1}),
\end{aligned} \tag{4.47}$$

using in last inequality that $\int t^{\alpha+1}K(t)dP^{d(\boldsymbol{\chi}_i,\chi)/b}(t)$ is bounded.

Applying (4.47) into the numerator of (4.46), together with Lemma 8, one obtains the expression involved into (4.45).

Now, in order to get (4.44) it only remains to verify that the order of the remaining terms maintains uniformly $\forall \chi_1 / d(\chi_1, \chi) \leq h$ and this can be done following the proof of Lemma 5 in Ferraty, Van Keilegom and Vieu (2010) as dependence does not affect these terms.

Finally, it is necessary to study the term U_1. Again, one will obtain $U_1 = o\left((nF_\chi(h))\right)^{-1/2}$ a.s. via the following result:

$$\sup_{d(\chi_1,\chi)\leq h} |\widehat{m}_b(\chi_1) - \widehat{m}_b(\chi) - \mathbb{E}(\widehat{m}_b(\chi_1)) + \mathbb{E}(\widehat{m}_b(\chi))| = o\left((nF_\chi(h))\right)^{-1/2} \text{ a.s.} \tag{4.48}$$

The expression involved in (4.48) can be also decompose in several terms, focusing on the most complicated one, which is

$$\widehat{g}_b(\chi_1) - \widehat{g}_b(\chi) - \mathbb{E}(\widehat{g}_b(\chi_1)) + \mathbb{E}(\widehat{g}_b(\chi)).$$

Applying assumption (4.18), it is known that the ball $B(\chi, h)$ can be covered by r_n balls $B(\chi_{k_n}, l_n)$ and so:

$$\begin{aligned}
&\sup_{d(\chi_1,\chi)\leq h} |\widehat{g}_b(\chi_1) - \widehat{g}_b(\chi) - \mathbb{E}(\widehat{g}_b(\chi_1)) + \mathbb{E}(\widehat{g}_b(\chi))| \leq \\
&\max_{1\leq k\leq r_n} |\widehat{g}_b(\chi_{k_n}) - \widehat{g}_b(\chi) - \mathbb{E}(\widehat{g}_b(\chi_{k_n})) + \mathbb{E}(\widehat{g}_b(\chi))| + \\
&+ \max_{1\leq k\leq r_n} \sup_{\chi_1\in B(\chi_{k_n},l_n)} |\widehat{g}_b(\chi_{k_n}) - \widehat{g}_b(\chi_1) - \mathbb{E}(\widehat{g}_b(\chi_{k_n})) + \mathbb{E}(\widehat{g}_b(\chi_1))| = \\
&V_1 + V_2.
\end{aligned} \tag{4.49}$$

Note that $F_{\chi_{k_n}}(b)$ can be replaced by $F_\chi(b)$. Let us define, for $i = 1, \ldots, n$ and $k = 1, \ldots, r_n$:

$$Z_{ik} = (F_\chi(b))^{-1} Y_i \{K(d(\boldsymbol{\chi}_i, \chi)/b) - K(d(\boldsymbol{\chi}_i, \chi_{k_n})/b)\}. \tag{4.50}$$

Using assumptions (4.16), (4.17) and knowing that

$$\mathbb{E}(K(d(\boldsymbol{\chi}_i, \chi)/b)) = \int K(t) dP^{d(\boldsymbol{\chi}_i,\chi)/b}(t) =$$
$$K(1)F_\chi(b) - \int_0^1 K'(s)F_\chi(bs)ds = F_\chi(b)\left(K(1) - \int_0^1 K'(s)\frac{F_\chi(s)}{F_\chi(b)}ds\right) =$$
$$F_\chi(b)\left(K(1) - \int_0^1 K'(s)\tau_b(s)ds\right) \longrightarrow$$
$$F_\chi(b)\left(K(1) - \int_0^1 K'(s)\tau_0(s)ds\right) = F_\chi(b)M_{1\chi}$$

and

$$\mathbb{E}(K^2(d(\boldsymbol{\chi}_i, \chi)/b)) =$$
$$\int K^2(t) dP^{d(\boldsymbol{\chi}_i,\chi)/b}(t) K^2(1)F_\chi(b) - \int_0^1 (K^2)'(s)F_\chi(bs)ds =$$
$$F_\chi(b)\left(K^2(1) - \int_0^1 (K^2)'(s)\frac{F_\chi(s)}{F_\chi(b)}ds\right) =$$
$$F_\chi(b)\left(K^2(1) - \int_0^1 (K^2)'(s)\tau_b(s)ds\right) \longrightarrow$$
$$F_\chi(b)\left(K^2(1) - \int_0^1 (K^2)'(s)\tau_0(s)ds\right) = F_\chi(b)M_{2\chi},$$

one gets that

$$\mathbb{E}(|Z_{ik}|^r) = O([F_\chi(b)^{-1}h/b]^{r-1}) \forall r \geq 2.$$

Then, assuming that $F_\chi(b)^{-1}h/b < C$, one can apply Lemma 21 resulting:

$$V_1 = \max_{1\leq k\leq r_n} \left| \sum_{i=1}^n n^{-1}(Z_{ik} - \mathbb{E}(Z_{ik})) \right| = O_{\text{a.s.}}\left(n^{-1/2+1/p}\log n\right) =$$
$$o\left((nF_\chi(h))^{-1/2}\right) \text{ a.s.}$$

where the last equality comes from $n^{1/p}F_\chi(h)^{1/2}\log n = o(1)$.
Also

$$V_2 = o\left((nF_\chi(h))^{-1/2}\right) \text{ a.s.}$$

(as the steps of the proof of Lemma 6 in Ferraty, Van Keilegom and Vieu (2010) can be followed here) and so,

$$U_1 = o\left((nF_\chi(h))\right)^{-1/2} \text{ a.s.}$$

Gathering all the results involving expression (4.41) one can conclude that:

$$\mathbb{E}^{\mathcal{S}}\left(\widehat{m}^*_{hb}(\chi)\right) - \widehat{m}_b(\chi) = \mathbb{E}(\widehat{m}_h(\chi)) - m(\chi) + o\left((nF_\chi(h))\right)^{-1/2}$$

which implies that (4.40) is true and so it is (4.39), which concludes the proof of the Theorem. □

Proof of Theorem 5

The proof for Theorem 5 follows the same steps as the one for Theorem 4 for the naïve bootstrap. As can be seen in Ferraty, Van Keilegom and Vieu (2010), there exists a difference between the proof of Theorem 4 and Theorem 5 under each bootstrap procedure. That difference affects the proof of Lemma 9. However, as it can be seen in the proof of the mentioned Lemma 9, it holds under dependence conditions for both bootstrap procedures, reasoning as in Ferraty, Van Keilegom and Vieu (2010) making use of our Lemmas 19, 27 and 8 and following, step by step, the proof of (9) given in Ferraty et al. (2007) (for the independent case). Thus, once Lemma 9 holds under both bootstrap procedures, one can apply the same proof of Theorem 4 to our Theorem 5. □

4.4 Simulation study

This section is devoted to illustrate, when finite sample sizes are used, the accuracy of the confidence interval for $m(\chi)$ constructed from the proposed bootstrap methodology. For that, such interval will be compared with the true (and, in practice, unknown) confidence interval. In addition, to show the behaviour of our bootstrap interval against that of the interval obtained from the asymptotic distribution of $\widehat{m}_h(\chi)$, some results for the asymptotic interval will be given. Because of its generality, it will be focused on the wild bootstrap procedure.

In a first example smooth curves are considered, while in a second one the case of rough curves is dealt.

4.4.1 Building the confidence intervals

Given a curve χ and a model

$$Y_i = m(\boldsymbol{\chi}_i) + \varepsilon_i \ (i = 1, \ldots, n), \tag{4.51}$$

where the process $\{(\boldsymbol{\chi}_i, Y_i)\}$ is α-mixing and identically distributed as $(\boldsymbol{\chi}, Y)$, and χ is observed from $\boldsymbol{\chi}$, the true, bootstrap and asymptotic $(1-\alpha)$-confidence intervals for $m(\chi)$ were constructed as

$$I^{true}_{\chi,1-\alpha} = (\widehat{m}_h(\chi) + q^{true}_{\alpha/2}(\chi), \widehat{m}_h(\chi) + q^{true}_{1-\alpha/2}(\chi)),$$

$$I^{*}_{\chi,1-\alpha} = (\widehat{m}_h(\chi) + q^{*}_{\alpha/2}(\chi), \widehat{m}_h(\chi) + q^{*}_{1-\alpha/2}(\chi))$$

and

$$I^{asymp}_{\chi,1-\alpha} = (\widehat{m}_h(\chi) + q^{asymp}_{\alpha/2}(\chi), \widehat{m}_h(\chi) + q^{asymp}_{1-\alpha/2}(\chi)),$$

respectively, where the quantiles $q^{true}_p(\chi)$, $q^{*}_p(\chi)$ and $q^{asymp}_p(\chi)$ were computed in the following way:

- Theoretical quantiles ($q^{true}_p(\chi)$).
    1. Generate n_{MC} samples $\{(\boldsymbol{\chi}^s_i, Y^s_i), i = 1, \ldots, n\}^{n_{MC}}_{s=1}$ from Model (4.51).
    2. Carry out n_{MC} estimates $\{\widehat{m}^s_h(\chi)\}^{n_{MC}}_{s=1}$, where $\widehat{m}^s_h(\cdot)$ is the functional kernel estimator (4.3) derived from the s^{th} sample $\{(\boldsymbol{\chi}^s_i, Y^s_i)\}^{n}_{i=1}$.
    3. Compute the set of approximation errors $ERRORS.MC = \{m(\chi) - \widehat{m}^s_h(\chi)\}^{n_{MC}}_{s=1}$.

4. Compute the theoretical quantile, $q_p^{true}(\chi)$, from the quantile of order p of $ERRORS.MC$.

- Bootstrap quantiles ($q_p^*(\chi)$).

1. Generate the sample $\mathcal{S} = \{(\boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{\chi}_n, Y_n)\}$ from Model (4.51).
2. Compute $\widehat{m}_b(\chi)$ over the dataset $\mathcal{S}$.
3. Repeat B times the bootstrap algorithm over $\mathcal{S}$ by using i.i.d. random variables, V_i, drawn from mixture of the two Dirac distributions $0.1(5+\sqrt{5})\delta_{(1-\sqrt{5})/2} + 0.1(5-\sqrt{5})\delta_{(1+\sqrt{5})/2}$, giving the B bootstrap estimates $\{\widehat{m}_{hb}^{*,r}(\chi)\}_{r=1}^{B}$.
4. Compute the set of bootstrap errors $ERRORS.BOOT = \{\widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi)\}_{r=1}^{B}$.
5. Compute the bootstrap quantile, $q_p^*(\chi)$, from the quantile of order p of $ERRORS.BOOT$.

- Asymptotic quantiles ($q_p^{asymp}(\chi)$).

1. Generate the sample $\mathcal{S} = \{(\boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{\chi}_n, Y_n)\}$ from Model (4.51).
2. Use the sample $\mathcal{S}$ to estimate the constants $F_\chi(h)$, $M_{1\chi}$, $M_{2\chi}$ and σ_ε as suggested in Delsol (2009), pages 18 and 20.
3. Compute the asymptotic quantile, $q_p^{asymp}(\chi)$, from the quantile of order p of the corresponding normal distribution.

Finally, the estimate $\widehat{m}_h(\chi)$ needed for each of the three intervals was obtained from $\mathcal{S}$.

The quadratic kernel, $K(u) = 1.5(1-u^2)1_{[0,1]}(u)$, was considered in the estimates $\widehat{m}_h$ and $\widehat{m}_{hb}^*$, while the bandwidth $b = b_{CV}$ was selected by means of the cross-validation methodology. Then, $h = b_{CV}$ was set.

4.4.2 Model 1: smooth curves

The first simulated model, Model 1, is based on the one used in Delsol (2009), where smooth curves $\boldsymbol{\chi}$ were considered. Some modifications were included to adapt his model to our context. Specifically, the discretized functional covariate in Model 1 was

$$\boldsymbol{\chi}_i(t_j) = \cos(a_i + \pi(2t_j - 1)), \tag{4.52}$$

where $\{a_i\}$ comes from AR(1) gaussian process with correlation coefficient $\rho_a = 0.7$ and variance $\sigma_a^2 = 0.05$. Values $0 = t_1 < t_2 < \cdots < t_{99} < t_{100} = 1$ equally spaced were considered. The regression operator was

$$m(\boldsymbol{\chi}) = \frac{1}{2\pi} \int_{1/2}^{3/4} (\boldsymbol{\chi}'(t))^2 dt$$

while the errors $\{\varepsilon_i\}$ were independent centered gaussians of variance equal to 0.1 times the empirical variance of $\{m(\boldsymbol{\chi}_1), \ldots, m(\boldsymbol{\chi}_n)\}$.

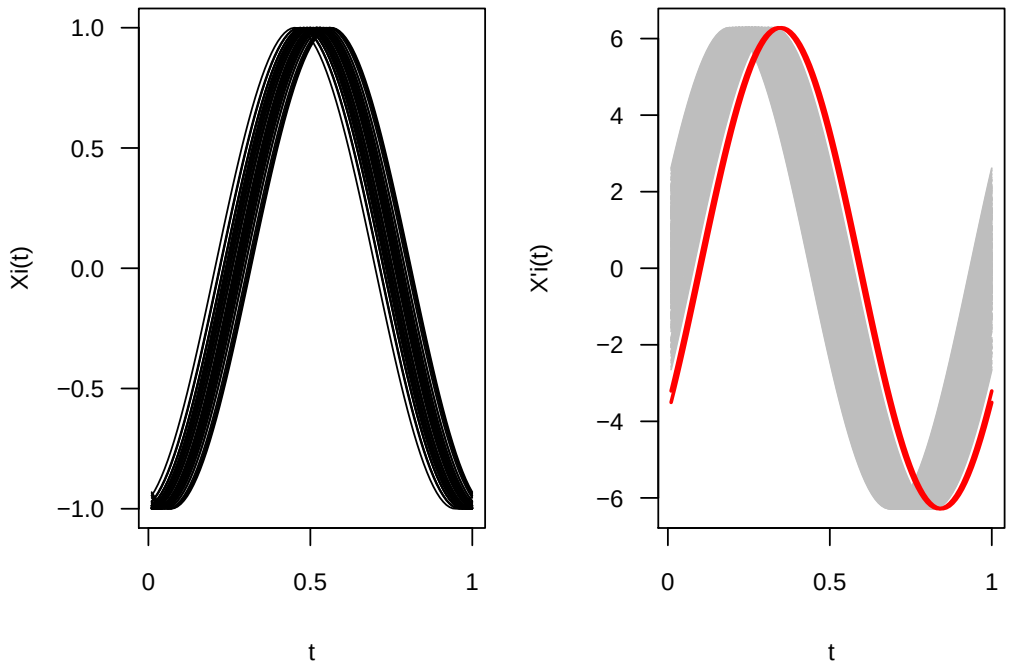


Figure 4.1: Left panel: first 50 curves in a training sample $\mathcal{S}$ ($n = 250$ was considered) generated from Model 1. Right panel: first derivative of the curves in $\mathcal{S}$, together with the first derivative of the curves χ_{41} and χ_{94} (in red) in the test sample $\mathcal{C}$.

Note that Model 1 deals with smooth curves (see left panel in Figure 4.1), this fact suggesting the use of a semi-metric based on some derivative of the curve (for details, see Section 13.6 in Ferraty and Vieu, 2006). Specifically, as recommended in Delsol (2009), the semi-metric $(d_1^{deriv}(\cdot,\cdot))$ considered in

Model 1 was based on the first derivative of the curve:

$$d_1^{deriv}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = \sqrt{\int_0^1 (\boldsymbol{\chi}_i'(t) - \boldsymbol{\chi}_j'(t))^2 dt}.$$

True, bootstrap and asymptotic $(1-\alpha)$-confidence intervals for $m(\chi)$ with $\chi \in \mathcal{C}$ were computed and compared. The test sample $\mathcal{C} = \{\chi_1, \ldots, \chi_{n_\mathcal{C}}\}$, consisting in $n_\mathcal{C}$ independent curves, was generated in the following way: first, $n_\mathcal{C}$ independent functional time series were obtained from the process $\{\boldsymbol{\chi}_i\}$ defined in (4.52); then, a curve χ was selected at random in each of such $n_\mathcal{C}$ functional time series. Note that from the procedure explained in the previous Section 4.4.1 one obtains one $(1-\alpha)$-confidence interval of each type for $m(\chi)$: true ($I^{true}_{\chi,1-\alpha}$), bootstrap ($I^{*}_{\chi,1-\alpha}$) and asymptotic ($I^{asymp}_{\chi,1-\alpha}$) confidence intervals. To compare the accuracy of each type of interval, the empirical coverages are obtained by repeating the procedure M times and computing the proportion of times that each interval contains the value $m(\chi)$.

Values $n_{MC} = 2000$, $B = 500$, $n_\mathcal{C} = 100$, $M = 500$, $1-\alpha = 0.95, 0.90$ and $n = 100, 250$ were considered.

Table 4.1 reports the average over $\mathcal{C}$ of the empirical coverage of the three computed confidence intervals. As expected, the accuracy of the coverages improves as the sample size, n, increases. In addition, coverages of the bootstrap intervals are closer to the theoretical coverages than the corresponding to the asymptotic intervals.

Table 4.1: Average over $\mathcal{C}$ of the empirical coverage of the true, bootstrap and asymptotic confidence intervals for Model 1. Standard deviation appears in brackets.

$1-\alpha$	0.95		0.90	
n	100	250	100	250
Coverage (I^{true})	0.946 (0.01)	0.951 (0.01)	0.896 (0.02)	0.903 (0.02)
Coverage (I^{*})	0.890 (0.12)	0.921 (0.08)	0.849 (0.12)	0.877 (0.08)
Coverage (I^{asymp})	0.852 (0.14)	0.898 (0.11)	0.794 (0.14)	0.842 (0.12)

Figure 4.2 shows a comparison of the empirical coverages of $I^{true}_{\chi,1-\alpha}$, $I^{*}_{\chi,1-\alpha}$ and $I^{asymp}_{\chi,1-\alpha}$ for each $\chi \in \mathcal{C}$. On the one hand, this figure clearly reflects the underestimation of the coverage by the asymptotic intervals, this fact being attenuated by the bootstrap ones. Therefore, at least in this example, bootstrap methodology is a nice alternative to the asymptotic one. On the

other hand, focusing on the empirical coverages by the bootstrap intervals, it is remarkable the presence of two confidence intervals with poor empirical coverages. Specifically, they correspond to $m(\chi_{41})$ and $m(\chi_{94})$ ($\chi_i \in \mathcal{C}$, $i = 41, 94$).

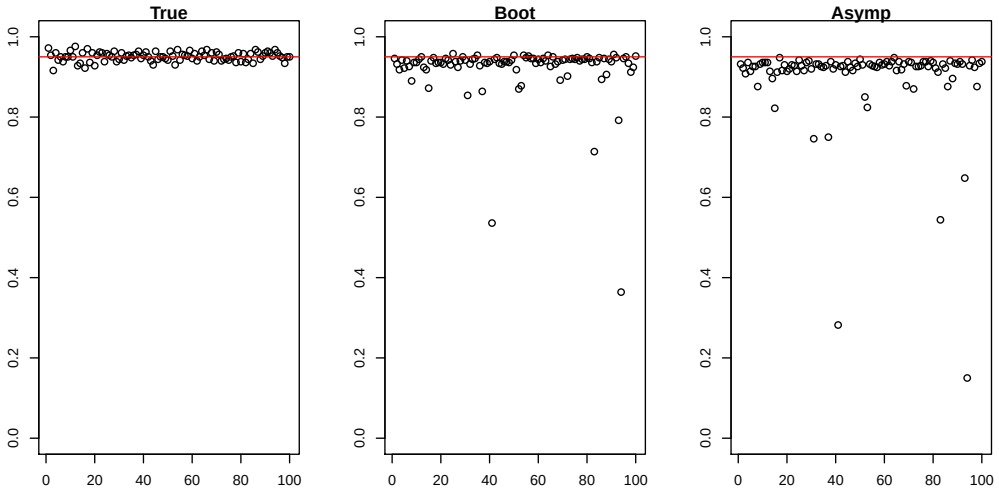


Figure 4.2: Empirical coverage of the true, bootstrap and asymptotic confidence intervals for Model 1 for each $\chi \in \mathcal{C}$ (values $1 - \alpha = 0.95$ and $n = 250$ are considered). Solid line is located at a height $1 - \alpha$.

In an attempt to find the reasons of those poor behaviours, Figure 4.1 (right panel) shows the first derivative of the curves in a training sample $\mathcal{S}$, together with the first derivative of the curves χ_{41} and χ_{94} in the test sample $\mathcal{C}$. It seems that χ_{41} and χ_{94} are atypical curves respect to $\mathcal{S}$. As attested from Figure 4.3 (left panel), this fact causes poor predictions for $m(\chi_{41})$ and $m(\chi_{94})$ and, therefore, poor confidence intervals.

Figure 4.3 (right panel) reports, for each $\chi \in \mathcal{C}$, the confidence intervals obtained by means of the bootstrap methodology (using the training sample $\mathcal{S}$ referred in the previous paragraph). True confidence intervals are also shown. Except for the cases of the atypical curves χ_{41} and χ_{94}, bootstrap intervals are close to the true ones.

In addition to the coverage, also the estimated density for the approximation errors is compared. For that, we denote by $\widehat{f}_{true,\chi}(\cdot)$, $\widehat{f}_{*,\chi}(\cdot)$ and $\widehat{f}_{asymp,\chi}(\cdot)$ the corresponding theoretical, bootstrap and asymptotic densities.

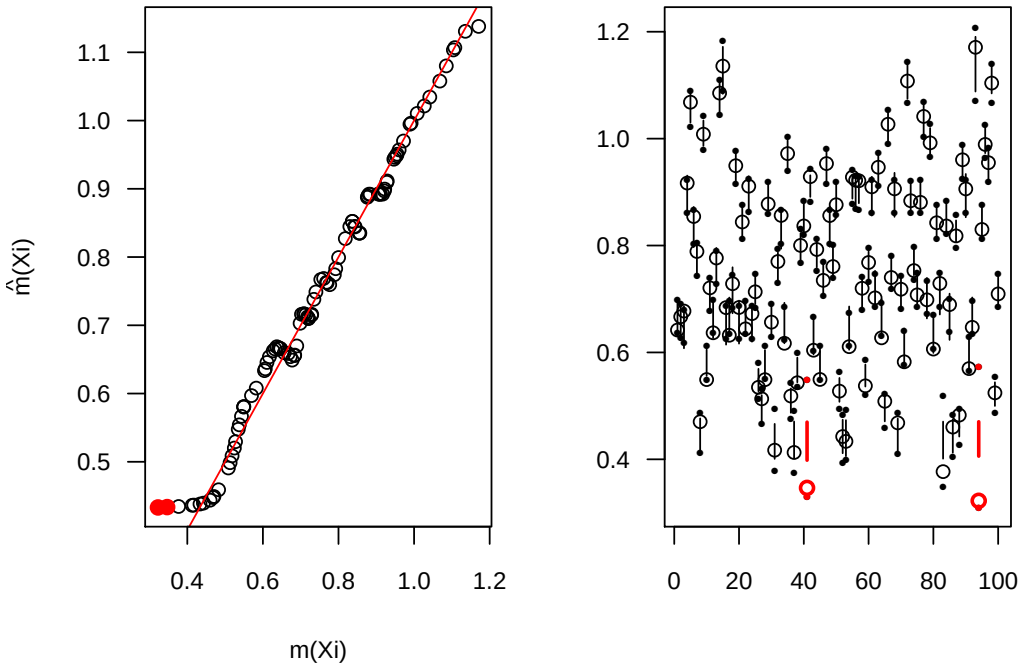


Figure 4.3: Left panel: predicted values ($\widehat{m}_h(\chi)$) (from a training sample $\mathcal{S}$ from Model 1) for each $\chi \in \mathcal{C}$ vs true values ($m(\chi)$). Full circles (in red) correspond to the atypical curves χ_{41} and $\chi_{94} \in \mathcal{C}$ ($n = 250$ was considered). Right panel: for each curve χ in $\mathcal{C}$, the vertical line represents the bootstrap confidence interval for $m(\chi)$ obtained from $\mathcal{S}$, while the dots delimit the true confidence interval ($1 - \alpha = 0.95$ was considered). In addition, the hollow circle locates the regression value $m(\chi)$. Outputs for the atypical curves are coloured in red.

Figure 4.4 displays the estimated densities for the three different approximation errors over four curves in $\mathcal{C}$. The first thing one can see in this figure is that both bootstrap and asymptotic curves well approach the true density. In the first plot, the asymptotic density is almost the same as the true one, and also in other two plots the asymptotic density seems to approximate better the true density than the bootstrapped density. It is now important to remember that, even if the whole density seems to be a good approximation, the real importance when dealing with confidence intervals is to obtain good approximations for the quantiles. Then, the key point is to better approximate the tails and not only the central part of the density.

Taking into account jointly the density estimation and the empirical coverage, we can see that the bootstrap procedure performs reasonable, due to

the trade-off between quantile and density estimation. The analysis will focus now in the true and the bootstrap errors densities in Figure 4.5, in which the comparison is given for eight curves in $\mathcal{C}$. It includes also the variational distances between $\widehat{f}_{true,\chi}(\cdot)$ and $\widehat{f}_{*,\chi}(\cdot)$ for those selected $\chi \in \mathcal{C}$. Note that such distance is defined as

$$dist_{\chi} = 0.5 \int \left| \widehat{f}_{true,\chi}(t) - \widehat{f}_{*,\chi}(t) \right| dt. \tag{4.53}$$

Finally, Figure 4.6 displays a boxplot of the variational distances for all $\chi \in \mathcal{C}$, indicating that the true errors can be well approximated by the bootstrapped errors.

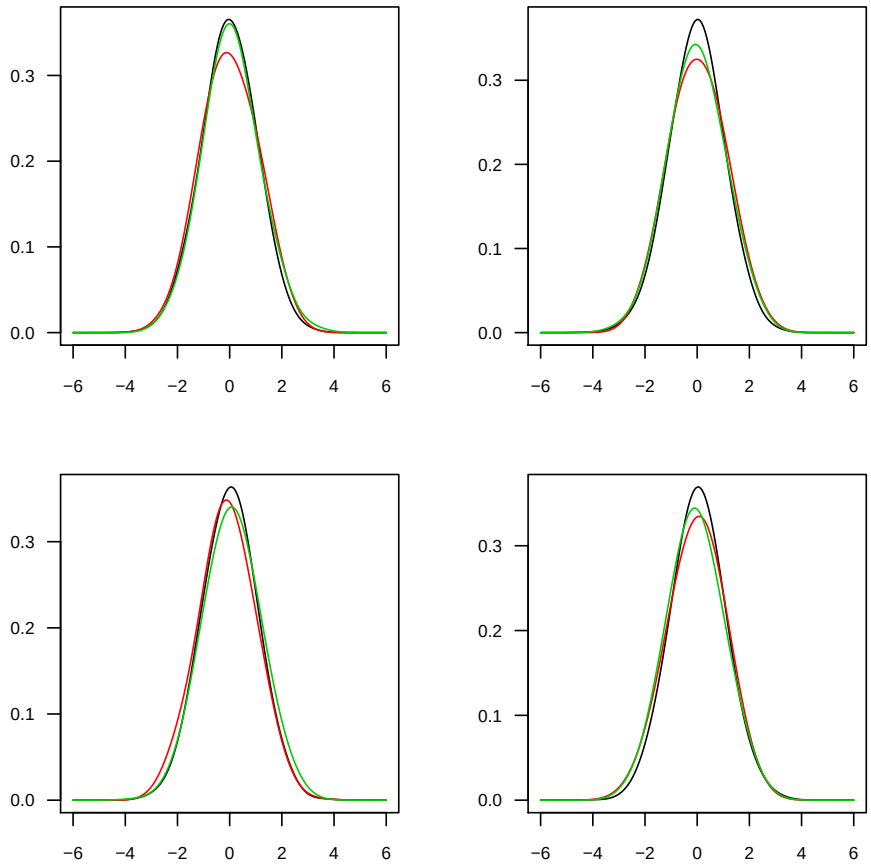


Figure 4.4: Estimated densities of the true error (black line), the bootstrapped error (red line) and the asymptotic error (green line), for four curves in $\mathcal{C}$.

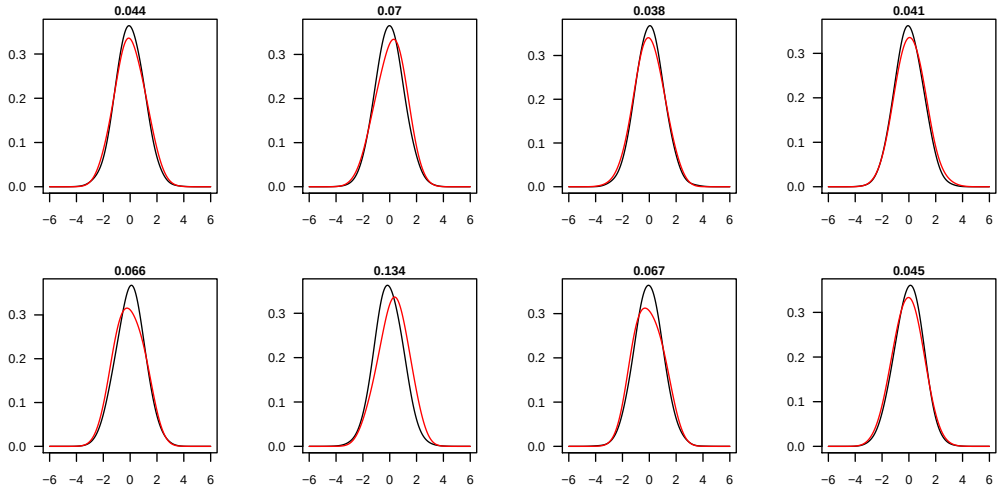


Figure 4.5: Estimated densities of the true error (black line) and the bootstrapped error (red line), for eight curves in $\mathcal{C}$.

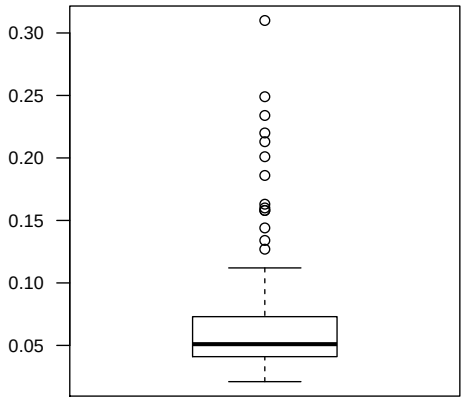


Figure 4.6: Boxplot of the variational distances between $\widehat{f}_{true,\chi}(\cdot)$ and $\widehat{f}_{*,\chi}(\cdot)$.

4.4.3 Model 2: rough curves

To provide further evidence of the interest of our methodology, a second example, dealing with rough curves, is given. Specifically, in Model 2 the discretized functional covariate was

$$\boldsymbol{\chi}_i(t_j) = b_{1i}\cos(b_{2i}t_j) + \sum_{k=1}^{j} B_{ik}/b,$$

where $b = 5$, $\{b_{1i}\}$ and $\{b_{2i}\}$ came from AR(1) and MA(1) gaussian processes with parameters $\rho_{b_1} = 0.9$ and $\theta_{b_2} = -0.5$, respectively, and variances $\sigma^2_{b_1} = \sigma^2_{b_2} = 0.1$. B_{ik} were i.i.d. realizations of $N(0, \sigma)$ with $\sigma = 0.1$ and $0 = t_1 < t_2 < \cdots < t_{99} < t_{100} = \pi$ were 100 equally spaced measurements. The regression operator was

$$m(\boldsymbol{\chi}) = \int_0^{\pi} (\boldsymbol{\chi}(t))^2 dt$$

and the errors were generated as in Model 1.

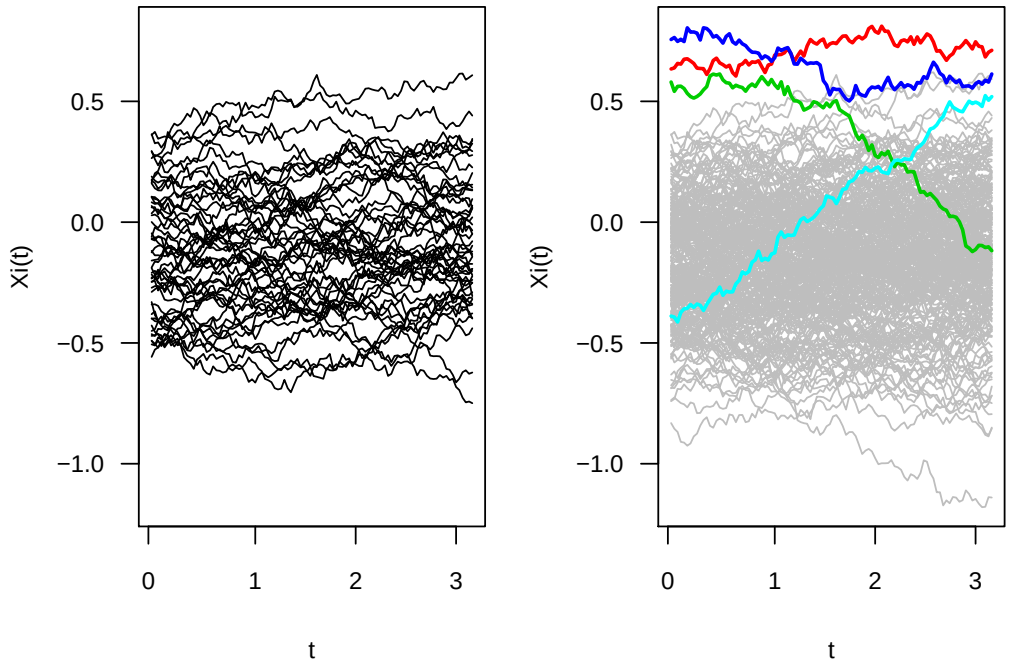


Figure 4.7: Left panel: first 50 curves in a training sample $\mathcal{S}$ ($n = 250$ was considered) generated from Model 2. Right panel: curves in $\mathcal{S}$ together with the curves χ_{20}, χ_{28}, χ_{39} and χ_{75} in the test sample $\mathcal{C}$.

Figure 4.7 (left panel) shows some sequential curves corresponding to a functional time series generated from Model 2.

Note that Model 2 adapts the model considered in Ferraty, Van Keilegom and Vieu (2012) to a setting of both scalar response and dependent curves. As recommended in that paper, the semi-metric $(d_4^{proj}(\cdot,\cdot))$ was based on the projection on the four eigenvectors, $v_1(\cdot),\ldots,v_4(\cdot)$, associated with the four largest eigenvalues of the empirical covariance operator of the functional predictor $\boldsymbol{\chi}$:

$$d_4^{proj}(\boldsymbol{\chi}_i,\boldsymbol{\chi}_j)=\sqrt{\sum_{k=1}^{4}\left(\int_0^{\pi}(\boldsymbol{\chi}_i(t)-\boldsymbol{\chi}_j(t))v_k(t)dt\right)^2}. \tag{4.54}$$

Table 4.2 reports the average over $\mathcal{C}$ of the empirical coverage of the three computed confidence intervals. The accuracy of the coverages improves as the sample size, n, increases. Coverages of both the bootstrap and the asymptotic intervals are worse than the ones obtained in the previous example of smooth curves, this fact showing the difficulties of the inference when dealing with curves with higher variability. In any case, bootstrap intervals continue to be better than the asymptotic ones (at least in this example).

Table 4.2: Average over $\mathcal{C}$ of the empirical coverage of the true, bootstrap and asymptotic confidence intervals for Model 2. Standard deviation appears in brackets.

$1-\alpha$	0.95		0.90	
n	100	250	100	250
Coverage (I^{true})	0.950 (0.01)	0.949 (0.01)	0.903 (0.02)	0.897 (0.02)
Coverage (I^{*})	0.804 (0.18)	0.889 (0.07)	0.774 (0.18)	0.861 (0.07)
Coverage (I^{asymp})	0.755 (0.17)	0.818 (0.06)	0.693 (0.16)	0.750 (0.06)

Figure 4.8 compares the empirical coverages of $I^{true}_{\chi,1-\alpha}$, $I^{*}_{\chi,1-\alpha}$ and $I^{asymp}_{\chi,1-\alpha}$ for each $\chi\in\mathcal{C}$. The underestimation of the coverage by the asymptotic intervals is clearly shown in this figure, this fact being attenuated by the bootstrap ones (as in the case of Model 1). Therefore, at least in this example, bootstrap methodology is a nice alternative to asymptotic one. Focusing on the empirical coverages by the bootstrap intervals, it is noted, again, the presence of some (four) confidence intervals with poor empirical coverages. They correspond to $m(\chi_{20})$, $m(\chi_{28})$, $m(\chi_{39})$ and $m(\chi_{75})$ $(\chi_i\in\mathcal{C},\, i=20,28,39,75)$.

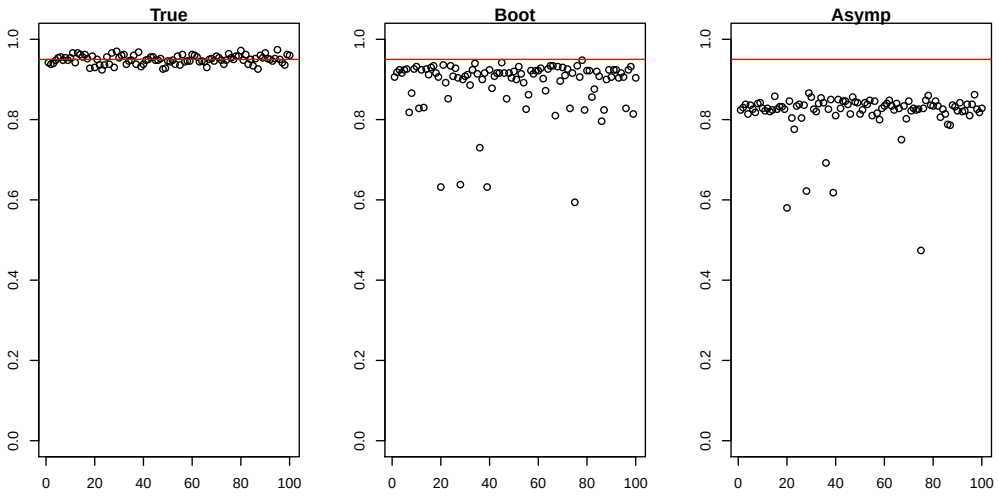


Figure 4.8: Empirical coverage of the true, bootstrap and asymptotic confidence intervals for Model 2 when values $1-\alpha = 0.95$ and $n = 250$ are considered. Solid line is located at a height $1-\alpha$.

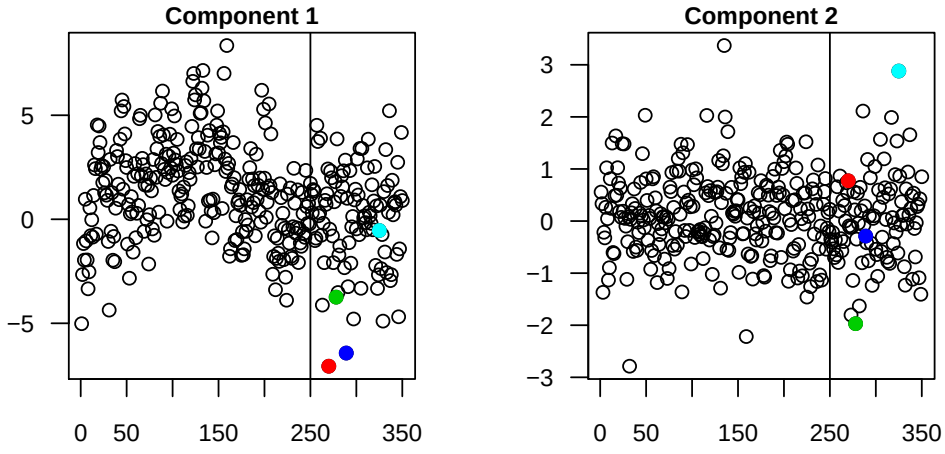


Figure 4.9: Left of the vertical line: scores of the first (left panel) and second (right panel) principal component of the curves in a training sample $\mathcal{S}$ (sample size $n = 250$). Right of the vertical line: scores of the curves in the test sample $\mathcal{S}$. Full circles correspond to the curves χ_{20}, χ_{28}, χ_{39} and $\chi_{75} \in \mathcal{C}$.

Figure 4.9 shows the scores of the first (left panel) and second (right panel) principal components of the curves in a training sample $\mathcal{S}$. The scores corresponding to the curves in the test sample $\mathcal{C}$ are also included. This figure shows that the scores of the first principal component of χ_{20} and χ_{39} are atypical with respect to the scores of the curves in the training sample. The same occurs for the scores of the second principal component of χ_{28} and χ_{75}. Note that the atypical behaviour of these four curves is supported by Figure 4.7 (right panel), which shows the curves in $\mathcal{S}$ together with χ_{20}, χ_{28}, χ_{39} and χ_{75}.

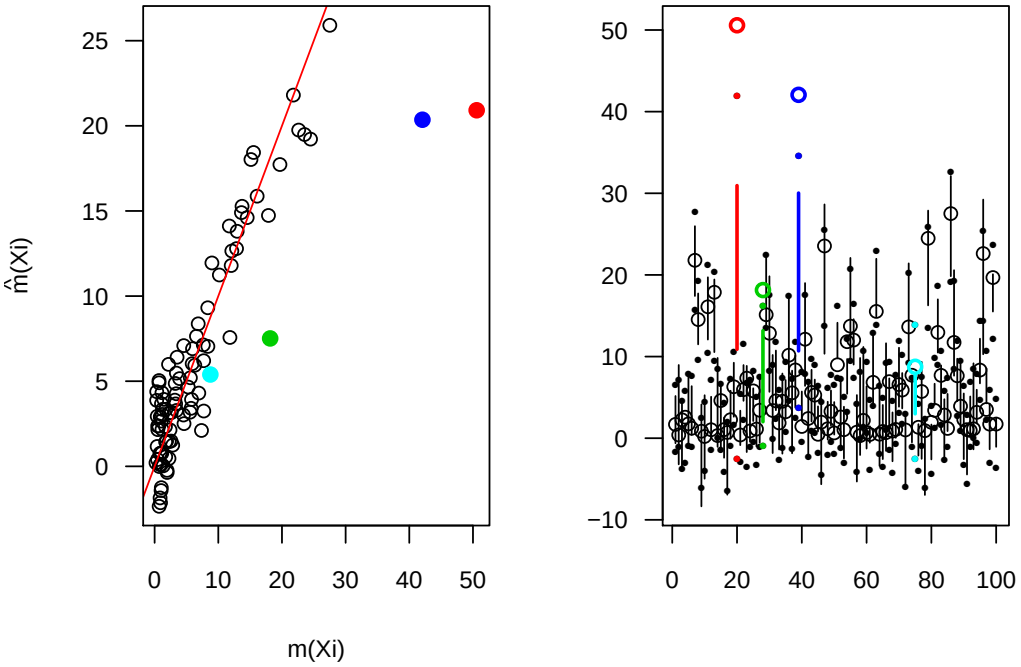


Figure 4.10: Left panel: predicted values ($\widehat{m}_h(\chi)$) (from a training sample $\mathcal{S}$ from Model 2) for each $\chi \in \mathcal{C}$ vs observed values ($m(\chi)$). Full circles correspond to the atypical curves χ_{20}, χ_{28}, χ_{39} and $\chi_{75} \in \mathcal{C}$ ($n = 250$ was considered). Right panel: for each curve χ in $\mathcal{C}$, the vertical line represents the bootstrap confidence interval obtained from $\mathcal{S}$, while the dots delimit the true confidence interval ($1 - \alpha = 0.95$ was considered). In addition, the hollow circle locates the regression value $m(\chi)$. Outputs in colour other than black correspond to the atypical curves.

Figure 4.10 (left panel) displays the points $(m(\chi), \widehat{m}_h(\chi))$ for $\chi \in \mathcal{C}$. The expected poor estimation of $m(\chi)$ in (three of the four) atypical curves χ_i, $i = 20, 28, 39, 75$ is attested from such figure, this fact causing the poor behaviour of the confidence intervals associated to those curves.

Figure 4.10 (right panel) reports, for each $\chi \in \mathcal{C}$, the confidence intervals obtained by means of the bootstrap methodology (using the training sample $\mathcal{S}$ referred in a previous paragraph). True confidence intervals are also shown. Excepting the cases of the atypical curves, bootstrap intervals are close to the true ones.

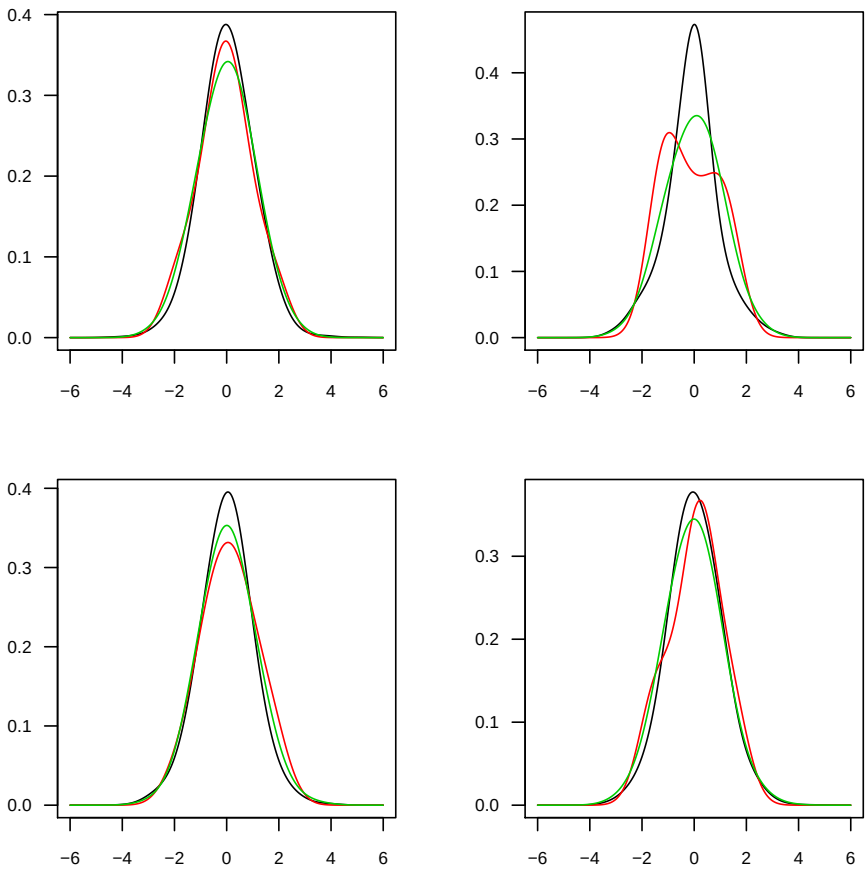


Figure 4.11: Estimated densities of the true error (black line), the bootstrapped error (red line) and the asymptotic error (green line), for four curves in $\mathcal{C}$.

This analysis finishes comparing again the estimated densities for the approximation errors. Figure 4.11 displays the estimated densities for the three different approximation errors over four curves in $\mathcal{C}$. Comparing with Figure 4.4 for the smooth curves, the density approximation seems to be worse when dealing with the rough curves due to the higher variability. However, we can see again that both bootstrap and asymptotic errors are a good approach for the true density, except the second plot in which both behaves badly.

The comparison focuses among the true and the bootstrap errors in Figure 4.12, in which the comparison is given for eight curves in $\mathcal{C}$. It includes also the variational distances defined in (4.53) between $\widehat{f}_{true,\chi}(\cdot)$ and $\widehat{f}_{*,\chi}(\cdot)$ for those selected $\chi \in \mathcal{C}$. As expected, the individual distances are now higher than the ones obtained with smooth curves, but the bootstrap density still behaving correctly.

Finally, Figure 4.13 displays the boxplot of the variational distances for all $\chi \in \mathcal{C}$, indicating that the true errors can be well approximated by the bootstrapped errors, even if this distance is now higher than in Figure 4.6.

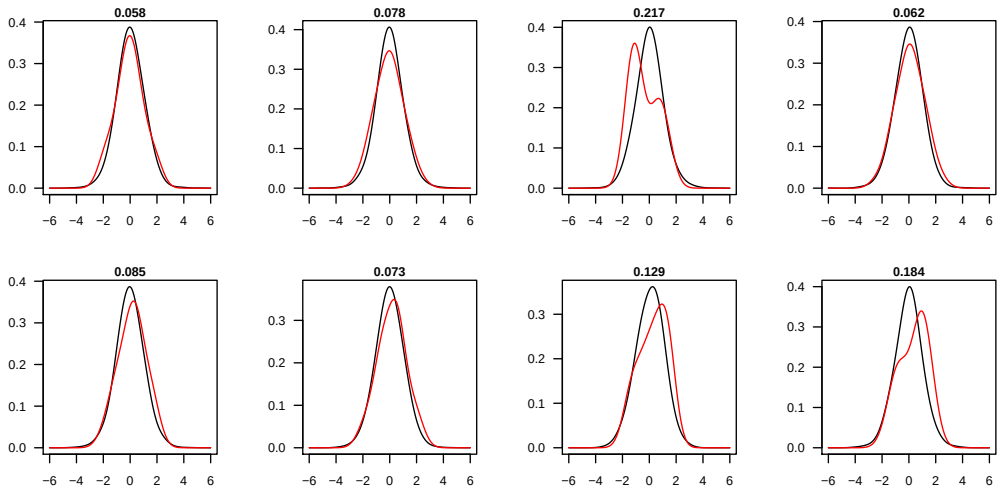


Figure 4.12: Estimated densities of the true error (black line) and the bootstrapped error (red line), for eight curves in $\mathcal{C}$.

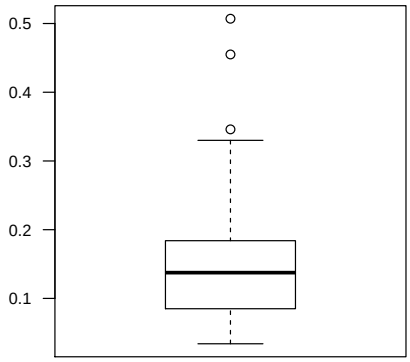


Figure 4.13: Boxplot of the variational distances between $\widehat{f}_{true,\chi}(\cdot)$ and $\widehat{f}_{*,\chi}(\cdot)$.

4.5 Application to electricity data

This section applies the methodology proposed in this chapter to the construction of confidence intervals for the mean hourly electricity demand/price in Spain given the daily curve of electricity demand/price in the previous day. As in the simulation study presented in Section 4.4, the wild bootstrap procedure will be considered. 1000 bootstrap replicates were drawn, the quadratic kernel was used and equal smoothing parameters $h = b = b_{CV}$ were considered, where b_{CV} was selected from a cross-validation method. In addition, the class of projection-based semi-metrics $\{d_v^{proj}(\cdot,\cdot)\}_v$ (see (4.54) for the case of $v = 4$) was considered, the quantity of eigenvectors v being also chosen from cross-validation. The confidence level considered was $1 - \alpha = 0.95$.

4.5.1 Case study: electricity demand

It is known that electricity demand shows vastly different patterns on workdays, public holidays and weekend (for details see Section 1.4). Thus, in order to accommodate this fact to model (4.2), only workdays are considered; in addition, to avoid (or attenuate) the effect that abrupt changes in temperature exert on electricity demand, our database was reduced to the second quarter of the year 2012. In summary, our database, $\mathcal{B}$, consists in the workdays of the second quarter of the year 2012. Each daily functional datum, χ_i, comes from the 24 hourly observations of electricity demand in

Spain for each day in our database.

In the following, two applications are presented. In the first one, the mean hourly electricity demand is estimated for each hour in a fixed day, while in the second application it is estimated for a fixed hour in different days.

In the first application, bootstrap confidence intervals were built for each mean hourly electricity demand corresponding to the last day in the database (Friday, June 29, 2012) given the daily curve of electricity demand in the previous day. Therefore, 24 confidence intervals need to be computed. The interval corresponding to the hour $t = 1, \ldots, 24$ was based on the regression model

$$\boldsymbol{\chi}_{i+1}(t) = m_t^{(1)}(\boldsymbol{\chi}_i) + \varepsilon_{i,t}^{(1)} \ (i = 1, \ldots, n); \tag{4.55}$$

that is, to do inference on the mean hourly electricity demand at hour t, the functional $G(\cdot)$ in model (4.1) is defined as $G_t(\boldsymbol{\chi}_{i+1}) = \boldsymbol{\chi}_{i+1}(t)$. Historical curves consisted in the days in our database, $\mathcal{B}$, previous to Friday, June 29, 2012 (fixed historical curves, not dependent on the prediction horizon t; equivalently, not dependent on the model). Figure 4.14 (left panel) displays the corresponding bootstrap confidence intervals. Note that the small sizes of such intervals respect to the big magnitude of the observed demand suggest a good accuracy of the bootstrap confidence intervals.

In order to maintain the prediction horizon (note that in the previous application 24 prediction horizons were considered), a second application was implemented. Specifically, a confidence interval was constructed for each mean electricity demand at fixed hour 20:00 corresponding to each of the $d = 1, \ldots, 21$ workdays in June 2012, given the daily curve of electricity demand in the previous day in our database. Therefore, 21 confidence intervals need to be computed. In this second application, historical data consisted in the workdays included in the 61 previous days (two previous months) to the day to predict, while the modelling to obtain the confidence interval corresponding to the day d was done by means of the regression model

$$\boldsymbol{\chi}_{i+1,d}(20) = m_d^{(2)}(\boldsymbol{\chi}_{i,d}) + \varepsilon_{i,d}^{(2)} \ (i = 1, \ldots, n_d); \tag{4.56}$$

that is, $G(\cdot)$ in model (4.1) was defined as $G(\boldsymbol{\chi}_{i+1}) = \boldsymbol{\chi}_{i+1}(20)$, and the historical curves were changing as d (equivalently, the model) does (in opposite at what occurred in the previous application). Figure 4.14 (right panel) shows the corresponding bootstrap confidence intervals. In this case, it can be observed that the ratio 'length of the interval/magnitude of the observed demand' is slightly greater than in the first application. This fact could be a

consequence of decreased sample size in (roughly) a 33 percent. In any case, it seems that the accuracy of the bootstrap confidence intervals continue to be sufficiently good.

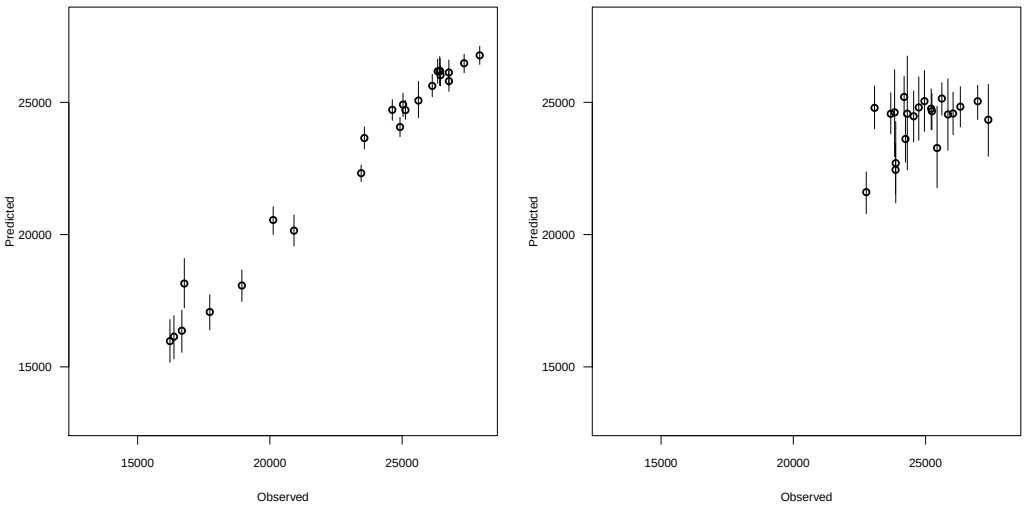


Figure 4.14: Left panel: Bootstrap confidence intervals computed for the electricity demand, for the 24 hours of Friday, June 29, 2012. Right panel: Bootstrap confidence intervals computed for the electricity demand, for the workdays in June, 2012 (fixed hour: 20:00).

4.5.2 Case study: electricity price

Electricity price shows also vastly different patterns on workdays, public holidays and weekend as the case of electricity demand, but maybe at a lower level (for details see Section 1.4). Thus, in order to accommodate this fact to our model (4.2), the study is focused again on workdays from the second quarter of the year 2012. In summary, analogous to the demand case study, our database $\mathcal{B}$, consists in the workdays of the second quarter of the year 2012. Each daily functional datum, χ_i, comes from the 24 hourly observations of electricity price in Spain for each day in the database.

Same structure as in the previous case study can be followed, presenting two applications: one for the estimation of the mean hourly electricity price for each hour in a fixed day and other for a fixed hour in different days.

First application follows the same procedure as in the demand case, focus on bootstrap confidence intervals for each mean hourly electricity price

corresponding to the last day in our database (Friday, June 29, 2012) given the daily curve of electricity price in the previous day in our database, using model (4.55). Historical curves consisted in the days in our database, $\mathcal{B}$, previous to Friday, June 29, 2012. Figure 4.15 (left panel) displays the corresponding bootstrap confidence intervals. In this situation the magnitude of the observed price is smaller than in demand but the sizes of such intervals are also small in proportion, which suggest a good accuracy of the bootstrap confidence intervals.

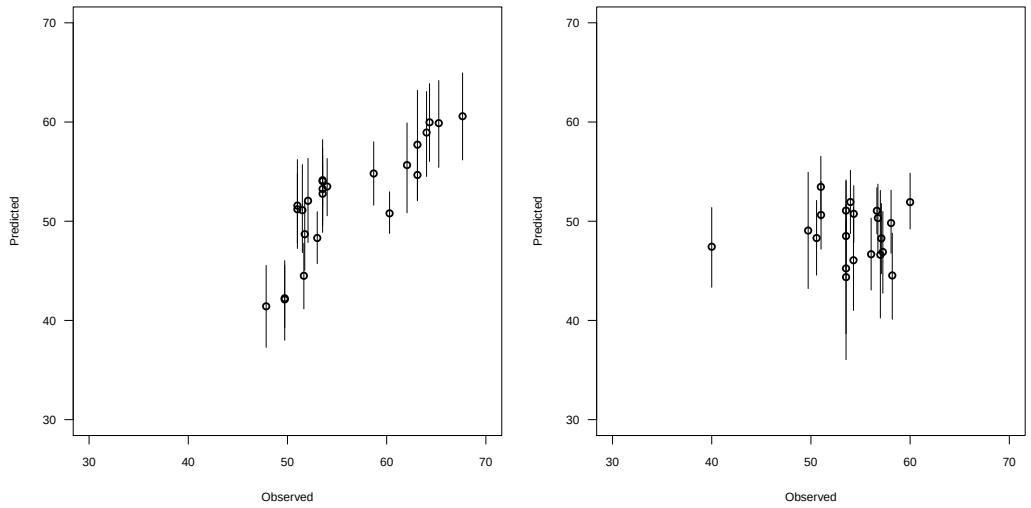


Figure 4.15: Left panel: Bootstrap confidence intervals computed for the electricity price, for the 24 hours of Friday, June 29, 2012. Right panel: Bootstrap confidence intervals computed for the electricity price, for the workdays in June, 2012 (fixed hour: 20:00).

A second application was implemented for each mean electricity price at fixed hour 20:00 corresponding, again, to each of the $d = 1, \ldots, 21$ workdays in June 2012, given the daily curve of electricity price in the previous day in our database. Therefore, 21 confidence intervals need to be computed, following model 4.56. Figure 4.15 (right panel) shows the corresponding bootstrap confidence intervals. In this case, it can be observed that the ratio 'length of the interval/magnitude of the observed price' is similar to the first application and thus, the accuracy of the bootstrap confidence intervals continue to be sufficiently good for this example.

4.6 Conclusions

This chapter has revisited the work by Ferraty, Van Keilegom and Vieu (2010) to extend its asymptotic results, established for independent samples, to the case of dependent ones. Based on both naïve and wild bootstrap procedures, pointwise confidence intervals for the regression function in a nonparametric model with functional predictor have been built, and their asymptotic validity has been established. Examples on finite sample sizes (via simulations and applications to real data) have shown that such results are useful in practice.

Both theoretical developments and applications presented in this chapter can be found in Raña, P., Aneiros, G., Vilar, J. and Vieu, P. (2016) in *Electronic Journal of Statistics.*

Interesting challenges remain as open problems to be dealt in a future. They include the topic of bandwidths selection: based on the simulation study presented in Ferraty, Van Keilegom and Vieu (2010), equal bandwidths chosen by cross-validation were considered in our applications; despite it is out of the scope of this paper, theoretical results on bandwidth selection for the proposed methodology would contribute greatly to the statistical literature (although we are aware of the difficulty of obtaining them). In addition, researches based on bootstraping pairs (instead of on bootstrapping residuals, as was done here) would give rise to new tools in the setting of nonparametric functional data analysis. To expand the range of possible applications, it would be very interesting to obtain results under dependence conditions in the random errors of the regression model.

Extension to functional response is certainly another interesting problem to be dealt. In fact, Ferraty, Van Keilegom and Vieu (2010) already extended it when dealing with independent data, giving some kind of functional pseudo-confidence area.

Note that, along the application given in this chapter, one deals with confidence intervals for the conditional expectation of the demand/price in a fixed hour, given the daily curve of the demand/price in the previous day. Thus, as in the practice this value is unknown, one can use the observed demand/price instead but, it must be pointed out that the confidence intervals are not design to capture this value. In order to extend this application, Chapter 6 will consider prediction intervals within this context.

Chapter 5

Confidence Intervals in Semi-Functional Partial Linear Regression

5.1 Introduction

This chapter is devoted to study the SFPL regression model and to propose naïve and wild bootstrap procedures to construct pointwise confidence intervals for each part of the cited model. Specifically, it deals with a predictor of functional nature and scalar covariates with nonparametric and linear effect, respectively. Assuming α-mixing conditions on the sample, the asymptotic validity of both procedures is obtained. A simulation study shows the performance of the procedures when finite sample sizes are used. In addition, an application to electrical data from the Spanish Electricity Market illustrates the helpful of the proposed methodology.

This chapter replicates the study developed in Chapter 4 for FNP regression, but considering now the SFPL model. Unlike the previous chapter, as far as we know, there is no preceding study in the literature regarding validity of the bootstrap in this model (even for the independent case). Moreover, it is even difficult to find applications of this kind of bootstrap procedures applied to scalar partial linear regression. One can find in Liang et al. (2000) and You and Chen (2006) proposals for bootstrap approximation in partial linear regression but in the case of fixed design, independent data and regarding the linear component of the model. For these reasons, this chapter shows the first approach to the validity of the bootstrap procedures developed in the context of SFPL model with dependent data (and, as a particular case,

to independent data), considering both linear and nonparametric parts of the model.

The rest of the chapter is organized as follows: Section 5.2 presents the SFPL model, its estimators and the bootstrap procedures. All the asymptotic theory is joined in Section 5.3, including in Subsection 5.3.1 the necessary assumptions for the good performance of the asymptotic results. The main contributions of this chapter are the four theorems stated in Subsection 5.3.2, which give the validity of the two bootstrap procedures for the two parts of the SFPL model (linear and nonparametric part). Section 5.4 and 5.5 include the application of the proposed bootstrap procedures to the construction of confidence intervals for both simulated and real data, respectively. Finally, Section 5.6 concludes the chapter.

5.2 The model and the bootstrap procedures

Throughout this memory, prediction on functional time series is of main importance and so, when dealing with SFPL, the interest lies in the model

$$G(\boldsymbol{\chi}_{i+1}) = \boldsymbol{X}_i^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

where G is a known real-valued operator, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown real parameters, m is an unknown smooth real-valued operator and ε_i are i.i.d. mean zero random errors. The explanatory random variables $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ and $\boldsymbol{\chi}_i$ are valued in $\mathbb{R}^p$ and some infinite-dimensional space $\mathcal{H}$, respectively. The functional space $\mathcal{H}$ is endowed with a semi-metric $d(\cdot, \cdot)$.

Asymptotic theory will be obtained for the more general SFPL model:

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{5.1}$$

where the sequence $\{(\boldsymbol{X}_i, \boldsymbol{\chi}_i, Y_i)\}$ is α-mixing. It is assumed that, for all $i = 1, \ldots, n$, $(\boldsymbol{X}_i, \boldsymbol{\chi}_i, Y_i)$ is identically distributed as $(\boldsymbol{X}, \boldsymbol{\chi}, Y)$, while the corresponding random errors $\{\varepsilon_i\}$ are i.i.d. as ε, verifying $\mathbb{E}(\varepsilon|\boldsymbol{X}, \boldsymbol{\chi}) = 0$ and $\mathbb{E}(\varepsilon^2|\boldsymbol{X}, \boldsymbol{\chi}) = \sigma_\varepsilon^2(\boldsymbol{X}, \boldsymbol{\chi}) < \infty$.

5.2.1 Estimators

Let us denote

$$\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T,\ \mathbf{Y} = (Y_1, \ldots, Y_n)^T,\ \mathbf{W}_h = (w_h(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j))$$

and, for any $(n \times q)$ matrix $\mathbf{A}$ $(q \geq 1)$,

$$\widetilde{\mathbf{A}}_h = (\mathbf{I} - \mathbf{W}_h)\mathbf{A}.$$

In addition, it is denoted

$$\mathcal{S} = \{(\boldsymbol{X}_1, \boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{\chi}_n, Y_n)\}.$$

The following estimators $\widehat{\boldsymbol{\beta}}_h$ and $\widehat{m}_h(\cdot)$ of the vector parameter $\boldsymbol{\beta}$ and the function $m(\cdot)$ in (5.1) will be considered:

$$\widehat{\boldsymbol{\beta}}_h = (\widetilde{\mathbf{X}}_h^T \widetilde{\mathbf{X}}_h)^{-1} \widetilde{\mathbf{X}}_h^T \widetilde{\mathbf{Y}}_h$$

and

$$\widehat{m}_h(\chi) = \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)(Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}_h),$$

respectively.

Estimator $\widehat{\boldsymbol{\beta}}_h$ can be motivated in the following way: if the true vector of parameters $\boldsymbol{\beta}$ was known, one could estimate $m(\chi)$ by means of $\widehat{m}_{h,\boldsymbol{\beta}}(\chi) = \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)(Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta})$. Meanwhile, as $\boldsymbol{\beta}$ is unknown, it is estimated by $\widehat{\boldsymbol{\beta}}_h$

$$\widehat{\boldsymbol{\beta}}_h = \arg \min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j - \widehat{m}_{h,\boldsymbol{\beta}}(\boldsymbol{\chi}_i) \right)^2.$$

It is then natural to estimate $m(\chi)$ by means of the kernel estimator $\widehat{m}_h(\chi)$.

In this chapter, the Nadaraya-Watson type weights are used:

$$w_h(\boldsymbol{\chi}_i, \chi) = \frac{K(d(\boldsymbol{\chi}_i, \chi)/h)}{\sum_{i=1}^{n} K(d(\boldsymbol{\chi}_i, \chi)/h)},$$

where $K(\cdot)$ is a real function (the kernel) and $h > 0$ is a smoothing parameter.

5.2.2 Bootstrap in SFPL models

Two bootstrap procedures are developed, which generalized to the SFPL model the procedures already studied in Chapter 4 for the FNP model. The first one, called "Naïve bootstrap", is designed for homoscedastic models (that is, when $\sigma_\varepsilon^2(\boldsymbol{X}, \boldsymbol{\chi}) = \mathbb{E}(\varepsilon^2 | (\boldsymbol{X}, \boldsymbol{\chi})) = \sigma_\varepsilon^2$) and the second one, called "Wild bootstrap", when it is heteroscedastic.

The bootstrap procedures follow the next algorithms:

Naïve bootstrap.

Step 1: Construct the residuals $\widehat{\varepsilon}_{i,b} = Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b - \widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$.

Step 2: Draw n i.i.d. random variables $\varepsilon_1^*, \ldots, \varepsilon_n^*$ from the empirical distribution function of $(\widehat{\varepsilon}_{1,b} - \overline{\widehat{\varepsilon}}_b, \ldots, \widehat{\varepsilon}_{n,b} - \overline{\widehat{\varepsilon}}_b)$, where $\overline{\widehat{\varepsilon}}_b = n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_{i,b}$.

Step 3: Obtain $Y_i^* = \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b + \widehat{m}_b(\boldsymbol{\chi}_i) + \varepsilon_i^*$, $i = 1, \ldots, n$.

Step 4: Define

$$\widehat{\boldsymbol{\beta}}_b^* = (\widetilde{\mathbf{X}}_b^T\widetilde{\mathbf{X}}_b)^{-1}\widetilde{\mathbf{X}}_b^T\widetilde{\mathbf{Y}^*}_b$$

and

$$\widehat{m}_{hb}^*(\chi) = \sum_{i=1}^n w_h(\boldsymbol{\chi}_i, \chi)(Y_i^* - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b^*),$$

Wild bootstrap.

Change Step 2 in the naïve bootstrap: define $\varepsilon_i^* = \widehat{\varepsilon}_{i,b}V_i$, $i = 1, \ldots, n$, where $V_1, \ldots, V_n$ are i.i.d. random variables that are independent of the data $\mathcal{S}$ and that satisfy $\mathbb{E}(V_1) = 0$, $\mathbb{E}(V_1^2) = 1$ and $\mathbb{E}(V_1^r) \leq \mathrm{C} < \infty$ for some $r > 6$. Maintain the other three steps.

5.3 Asymptotic theory

5.3.1 Assumptions

In the following, χ denotes a fixed element of the space $\mathcal{H}$.

Next set of assumptions was already established in Theorem 1 in Aneiros and Vieu (2008), to prove the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ and also the iterated logarithm law for the same SFPL model under dependence. Those results will play a main role in the proofs of our asymptotic results.

Semi-metric space

$$\boldsymbol{\chi} \text{ is valued in some given comptact subset } \mathcal{C} \text{ of } \mathcal{H} \text{ such that } \mathcal{C} \subset \bigcup_{k=1}^{\tau_n} \mathcal{B}(z_k, l_n), \tag{5.2}$$

where $\tau_n l_n^\gamma = \mathrm{C}$ (γ and C denote real positive constants), $\tau_n \to \infty$ and $l_n \to 0$ as $n \to \infty$.

Kernel

$$\begin{array}{c} K \text{ has support } [0,1] \text{ and is Lipschitz continuous on } [0,\infty). \\ \text{In addition, } \exists k/\forall u \in [0,1], -K'(u) > k > 0. \end{array} \tag{5.3}$$

Smoothness Denote $g_j(\chi) = \mathbb{E}(X_{ij}|\boldsymbol{\chi}_i = \chi),\ 1 \leq i \leq n, 1 \leq j \leq p$.
It is assumed that all the operators to be estimated are smooth, ie, for some $c < \infty$ and $\alpha > 0$,

$$\forall (u,v) \in \mathcal{C} \times \mathcal{C}, \forall f \in \{m, g_1, \ldots, g_p\}, \text{ we have: } |f(u) - f(v)| \leq c d(u,v)^{\alpha}. \tag{5.4}$$

Distributions For the probability distribution of the infinite-dimensional process $\boldsymbol{\chi}$, it is assumed that exists $F(\cdot)$, a positive valued function on $(0,\infty)$ and positive constants $\alpha_0, \alpha_1, \alpha_2$ such that

$$\begin{array}{l} \displaystyle\int_0^1 F(hs)ds > \alpha_0 F(h) \text{ and} \\ \alpha_1 F(h) \leq P(\boldsymbol{\chi} \in \mathcal{B}(t,h)) \leq \alpha_2 F(h)\ \forall t \in \mathcal{C}, h > 0. \end{array} \tag{5.5}$$

The joint probability distribution of $(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)$ is assumed that exists a function $\psi(h) = cF(h)^{1+\epsilon}$ $(c > 0, 0 \leq \epsilon \leq 1)$ and positive constants α_3, α_4 such that

$$0 < \alpha_3 \psi(h) \leq \sup_{i \neq j} P[(\boldsymbol{\chi}_j, \boldsymbol{\chi}_j) \in \mathcal{B}(t,h) \times \mathcal{B}(t,h)] \leq \alpha_4 \psi(h), \quad \forall t \in \mathcal{C}, h > 0. \tag{5.6}$$

Dependence structure $\{(\boldsymbol{X}_i, \boldsymbol{\chi}_i, Y_i)\}_{i=1}^n$ come from some stationary strong mixing process, with mixing coefficients $\{\alpha(n)\}$ that verify

$$\alpha(n) \leq cn^{-a}, a > 4.5, \tag{5.7}$$

while

$$\boldsymbol{\eta}_i \text{ is independent of } \varepsilon_i\ (i = 1, \ldots, n), \tag{5.8}$$

where $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ip})^T$, $\eta_{ij} = X_{ij} - \mathbb{E}(X_{ij}|\boldsymbol{\chi}_i) = X_{ij} - g_j(\boldsymbol{\chi}_i), j = 1, \ldots, p$.

Moments Denote $\boldsymbol{V}_{\boldsymbol{\varepsilon}} = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T), \boldsymbol{\varepsilon}^T = (\varepsilon_1, \ldots, \varepsilon_n), \boldsymbol{\eta}^T = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$.
It is supposed that:

$$\mathbb{E}|Y_1|^r + \mathbb{E}|X_{11}|^r + \ldots + \mathbb{E}|X_{1p}|^r < \infty \text{ for some } r > 6. \tag{5.9}$$

$$sup_{i,j} \mathbb{E}(|Y_i Y_j||(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)) < \infty \tag{5.10}$$

$$\max_{1\leq j\leq p}\sup_{i_1,i_2}\mathbb{E}(|X_{i_1,j}X_{i_2,j}||(\boldsymbol{\chi}_{i_1,j}\boldsymbol{\chi}_{i_2,j}))<\infty \tag{5.11}$$

$$\boldsymbol{B}=\mathbb{E}(\boldsymbol{\eta}_1\boldsymbol{\eta}_1^T),\ \boldsymbol{C}=\lim_{n\to\infty}n^{-1}\mathbb{E}(\boldsymbol{\eta}^T\boldsymbol{V}_{\boldsymbol{\varepsilon}}\boldsymbol{\eta}).$$
$$\boldsymbol{B}\text{ and }\boldsymbol{C}\text{ are positive definite matrix.} \tag{5.12}$$

$$s_n^{\frac{r(a+1)}{2(a+r)}}=o(n^{\theta})\text{ for some }\theta>2, \tag{5.13}$$

where $s_n=\sup_{\chi\in\mathcal{C}}(s_{n,1}(\chi)+s_{n,2}(\chi)+s_{n,3}(\chi))$, with

$$s_{n,1}(\chi)=\sum_{i=1}^{n}\sum_{j=1}^{n}|Cov(\Delta_i(\chi),\Delta_j(\chi)|\text{ with }\Delta_i(\chi)=K(\frac{d(\boldsymbol{\chi}_i,\chi)}{h})$$

$$s_{n,2}(\chi)=\sum_{i=1}^{n}\sum_{j=1}^{n}|Cov(\Gamma_i(\chi),\Gamma_j(\chi)|\text{ with }\Gamma_i(\chi)=Y_iK(\frac{d(\boldsymbol{\chi}_i,\chi)}{h})$$

$$s_{n,3}(\chi)=\max_{1\leq k\leq p}\sum_{i=1}^{n}\sum_{j=1}^{n}|Cov(\Gamma_{ik}(\chi),\Gamma_{jk}(\chi)|\text{ with }\Gamma_{ik}(\chi)=X_{ik}K(\frac{d(\boldsymbol{\chi}_i,\chi)}{h})$$

Small ball probabilities In order to manage the convergence rates found in the development of Theorems 10 and 11, it is necessary to consider the following assumptions:

$$\begin{gathered}nh^{4\alpha}\to 0,\ F(h)^{-1}n^{-1/4+1/r}\log n\to 0,\ nF(h)^{\frac{\varepsilon a(r-2)}{r}-1}=O(1)\text{ and}\\ F(h)^{-2}\big(n^{1-\frac{\theta(a+r)}{r(a+1)}}\big)^{-2}\log n=O(1)\text{ as }n\to\infty\\ \text{where }\alpha>0,0\leq\varepsilon\leq 1,a>4.5,r>6\text{ and }\theta>2\\ \text{were defined in the assumptions above.}\end{gathered} \tag{5.14}$$

Finally, assumptions formulated in Chapter 4, Section 4.3.1 for the FNP model will be also considered. To work with SFPL model, the same assumptions (4.4-4.20) are adopted but considering that now they will be applied over the model $\mathcal{Y}=m(\chi)+\varepsilon$ (where $\mathcal{Y}=Y-X\beta$ and m and ε were defined in (5.1)).

5.3.2 Asymptotic results

Let $P^{\mathcal{S}}$ denote probability, conditionally on the sample $\mathcal{S}$, and $\mathbf{a}$ a constant vector in $\mathbb{R}^p$.

Validity of the bootstrap procedures for the linear part

Theorem 10 (Naïve) *Under Assumptions (5.2)-(5.14), if the model is homoscedastic, for the naïve bootstrap one has:*

$$\sup_{y\in\mathbb{R}}\left|P^{\mathcal{S}}\left(\sqrt{n}\mathbf{a}^T(\widehat{\boldsymbol{\beta}}_b^*-\widehat{\boldsymbol{\beta}}_b)\leq y\right)-P\left(\sqrt{n}\mathbf{a}^T(\widehat{\boldsymbol{\beta}}_b-\boldsymbol{\beta})\leq y\right)\right|\to_P 0.$$

Theorem 11 (Wild) *Under Assumptions (5.2)-(5.14) if, in addition* $|\varepsilon|\leq c<\infty$, $F(h)^{-1}n^{-1/4+1/r}logn(loglogn)^{1/4}\to 0$, *for the wild bootstrap procedure one has that*

$$\sup_{y\in\mathbb{R}}\left|P^{\mathcal{S}}\left(\sqrt{n}\mathbf{a}^T(\widehat{\boldsymbol{\beta}}_b^*-\widehat{\boldsymbol{\beta}}_b)\leq y\right)-P\left(\sqrt{n}\mathbf{a}^T(\widehat{\boldsymbol{\beta}}_b-\boldsymbol{\beta})\leq y\right)\right|\to_P 0.$$

Remark 12 *Main assumption related to the use of wild bootstrap in Theorem 11, which is not necessary when the bootstrap procedure is naïve, is that random errors* ε_i *are bounded. This assumption is necessary in (5.67) in order to manage the convergence of the bootstrapped errors term. Also, one needs to assume* $F(h)^{-1}n^{-1/4+1/r}\log n(\log\log n)^{1/4}\to 0$, *which is a small variation of the assumptions given in (5.14).*

Remark 13 *Both theorems 10 and 11 establish the validity of the bootstrap procedures (naïve and wild bootstrap, respectively) for the linear part of the SFPL model. They represent a first extension of the theorems 4 and 5 from the FNP to the SFPL model, which will be completed with the results given in the next paragraph taking into account also the nonparametric part. Theorems 10 and 11 allows to approximate, through its bootstrapped estimator, the asymptotic distribution of the estimator of* $\mathbf{a}^T\boldsymbol{\beta}$ *within the SFPL model, which has been studied in Aneiros and Vieu (2008). Its main practical usefulness is related to the building of confidence intervals for* $\mathbf{a}^T\boldsymbol{\beta}$ *in a context of dependent data. Following a similar procedure as in Chapter 4, the quantiles of the distribution of* $\mathbf{a}^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)$ *can be approximated by means of the quantiles from the distribution of the bootstrapped error* $\mathbf{a}^T(\widehat{\boldsymbol{\beta}}_b-\widehat{\boldsymbol{\beta}}_b^*)$. *As one can generate as many replicates of the bootstrapped error as you want, the results of the Theorems 10 and 11, allow to obtain good approximations of the theoretical quantiles to build the confidence intervals for* $\mathbf{a}^T\boldsymbol{\beta}$.

As mentioned in the introduction of this chapter, there is no preceding in the literature regarding the validity of these bootstrap procedures within our context. In fact, one can find in Liang et al. (2000) a proposal for bootstrap approximation in partial linear regression, in the case of fixed design. They propose a naïve bootstrap that has been generalized by You and Chen (2006) to the wild bootstrap, allowing to consider heteroscedastic models. Both results take care of the linear component of the model with independent data. Thus, Theorems 10 and 11 show a first approach to this bootstrap procedures in the context of SFPL model under dependent data (and, as a particular case, under independent data too).

Validity of the bootstrap procedures for the nonparametric part

In order to manage the term $\boldsymbol{X}$ in the proofs of the following theorems, the next assumption is added:

$$||\boldsymbol{X}||_\infty \leq c < \infty. \tag{5.15}$$

Also, as Theorems 4 and 5 will be applied within the proofs of Theorems 14 and 15, respectively, it will be necessary to impose the following assumptions:

$$\begin{aligned} \max\{&\mathbb{E}(|(m(\boldsymbol{\chi}_i)+\varepsilon_i)(m(\boldsymbol{\chi}_j)+\varepsilon_j)|^p|\boldsymbol{\chi}_i,\boldsymbol{\chi}_j) \leq M \ a.s., \\ &\mathbb{E}(|m(\boldsymbol{\chi}_i)+\varepsilon_i|^p|\boldsymbol{\chi}_i,\boldsymbol{\chi}_j)\} \leq M \ a.s. \ \forall i,j \in \mathbb{Z}. \end{aligned} \tag{5.16}$$

$$\begin{aligned} &\text{Function } \mathbb{E}(|m(\boldsymbol{\chi})+\varepsilon||\boldsymbol{\chi}=\cdot) \text{ is continuous on a neighbourhood of } \chi, \\ &\text{and } \sup_{d(\chi_1,\chi)<\delta}\mathbb{E}(|m(\boldsymbol{\chi})+\varepsilon|^q|\boldsymbol{\chi}=\chi_1) < \infty \text{ for some } \delta>0; \forall q \geq 1. \end{aligned} \tag{5.17}$$

Theorem 14 (Naïve) *Under Assumptions (4.4)-(4.20) and (5.15)-(5.17), considering also assumptions for Theorem 10, if the model is homoscedastic, for the naïve bootstrap procedure, one has that*

$$\begin{aligned} \sup_{y\in\mathbb{R}} \Bigg| &P^{\mathcal{S}}\left(\sqrt{nF_\chi(h)}(\widehat{m}^*_{hb}(\chi)-\widehat{m}_b(\chi)) \leq y\right) - \\ &P\left(\sqrt{nF_\chi(h)}(\widehat{m}_h(\chi)-m(\chi)) \leq y\right)\Bigg| \to_P 0. \end{aligned}$$

Theorem 15 (Wild) *Under Assumptions (4.4)-(4.20) and (5.15)-(5.17), considering also assumptions for Theorem 11, for the wild bootstrap procedure, one has that*

$$\sup_{y\in\mathbb{R}}\left|P^{\mathcal{S}}\left(\sqrt{nF_\chi(h)}(\widehat{m}^*_{hb}(\chi)-\widehat{m}_b(\chi))\leq y\right)-P\left(\sqrt{nF_\chi(h)}(\widehat{m}_h(\chi)-m(\chi))\leq y\right)\right|\to_P 0.$$

Remark 16 *Theorems 14 and 15 can be seen as an extension of Theorems 4 and 5 in Section 4.3.2, but considering now the nonparametric part of the SFPL model. They complement Theorems 10 and 11 for the linear part, allowing to consider all the estimators within this model. Its main practical usefulness is, again, related to the building of confidence intervals for $m(\chi)$ in a context of dependent data (and, as a particular case, also for independent data). Following an analogous procedure as in the case of Theorems 4 and 5 from the FNP model (see Chapter 4), the α-quantile, $q_\alpha(\chi)$, of $m(\chi)-\widehat{m}_h(\chi)$ can be approximated by the α-quantile, $q^*_\alpha(\chi)$, obtained from the distribution of the bootstrapped errors $\widehat{m}_b(\chi)-\widehat{m}^*_{hb}(\chi)$. This fact, allows to use this bootstrapped errors for the nonparametric part, within the SFPL model, to approximate the desired confidence intervals for $m(\chi)$.*

5.3.3 Proofs

Proof Theorem 10

First, let $\mathbb{E}^{\mathcal{S}}$ and $Var^{\mathcal{S}}$ denote expectation and variance, respectively, conditionally on the sample $\mathcal{S}$, while Φ is the standard normal distribution function.

Let us write

$$P^{\mathcal{S}}\left(\sqrt{n}\mathbf{a}^T(\widehat{\boldsymbol{\beta}}^*_b-\widehat{\boldsymbol{\beta}}_b)\leq y\right)-P\left(\sqrt{n}\mathbf{a}^T(\widehat{\boldsymbol{\beta}}_b-\boldsymbol{\beta})\leq y\right)=T_1(y)+T_2(y), \quad (5.18)$$

where,

$$T_1(y)=P^{\mathcal{S}}\left(\sqrt{n}\mathbf{a}^T(\widehat{\boldsymbol{\beta}}^*_b-\widehat{\boldsymbol{\beta}}_b)\leq y\right)-\Phi\left(\frac{y}{\sqrt{\mathbf{a}^T\mathbf{A}\mathbf{a}}}\right) \quad (5.19)$$

and

$$T_2(y)=\Phi\left(\frac{y}{\sqrt{\mathbf{a}^T\mathbf{A}\mathbf{a}}}\right)-P\left(\sqrt{n}\mathbf{a}^T(\widehat{\boldsymbol{\beta}}_b-\boldsymbol{\beta})\leq y\right). \quad (5.20)$$

We have denoted $\mathbf{A}=\mathbf{B}^{-1}\mathbf{C}\mathbf{B}^{-1}$, where $\mathbf{B}$ and $\mathbf{C}$ were defined in assumption (5.12).

Note that, if one proves that

$$T_1(y) \longrightarrow_P 0 \text{ for any fixed value of } y \tag{5.21}$$

and

$$T_2(y) \longrightarrow 0 \text{ for any fixed value of } y, \tag{5.22}$$

then, from Polya's theorem (see Theorem 26 in Appendix A) together with the continuity of the function Φ, one will have that

$$\sup_{y\in\mathbb{R}} |T_1(y)| + \sup_{y\in\mathbb{R}} |T_2(y)| \to_P 0 \ , \tag{5.23}$$

which concludes the proof.

On the one hand, from Theorem 25 in Appendix A:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}) \xrightarrow{d} N(0, \mathbf{A});$$

this implies that:

$$\sqrt{n}\mathbf{a}^T(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}) \xrightarrow{d} N(0, \mathbf{a}^T\mathbf{A}\mathbf{a})$$

and, thus, (5.22) holds.

On the other hand, in order to obtain (5.21) it is sufficient to prove that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}}_b) \xrightarrow{d} N(0, \mathbf{A}), \text{ in probability conditionally on } \mathcal{S}. \tag{5.24}$$

First, for a given function $g(\cdot) = m(\cdot)$ or $g(\cdot) = \widehat{m}_b(\cdot)$, denote

$$\widetilde{g}_b(\chi) = g(\chi) - \sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i, \chi) g(\boldsymbol{\chi}_i).$$

Then, one can write

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}}_b) = \quad & (n^{-1}\widetilde{\mathbf{X}}_b^T\widetilde{\mathbf{X}}_b)^{-1}n^{-1/2}\left(\sum_{i=1}^{n}\widetilde{\boldsymbol{X}}_i\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i) - \right. \\ & \left. \sum_{i=1}^{n}\widetilde{\boldsymbol{X}}_i\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon_k^* + \sum_{i=1}^{n}\widetilde{\boldsymbol{X}}_i\varepsilon_i^*\right) = \\ & (n^{-1}\widetilde{\mathbf{X}}_b^T\widetilde{\mathbf{X}}_b)^{-1}n^{-1/2}(S_{n1}^* - S_{n2}^* + S_{n3}^*). \end{aligned} \tag{5.25}$$

Using this decomposition, if one proves:

$$S_{n1}^* - S_{n2}^* + S_{n3}^* = \sum_{i=1}^{n}\boldsymbol{\eta}_i\varepsilon_i^* + o_{P\mathcal{S}}(n^{1/2}), \tag{5.26}$$

and

$$n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\eta}_i\varepsilon_i^* \xrightarrow{d} N(0,\mathbf{C}), \text{ in probability conditionally on the sample } \mathcal{S}, \tag{5.27}$$

on gets, together with Lemma 24 in Appendix A, the asymptotic normality of (5.24).

To obtain (5.26), the order of $\max_i|\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i)|$ and

$$\max_i\left(\sum_{k=1}^{n}w_b(\boldsymbol{\chi}_i,\boldsymbol{\chi}_k)\varepsilon_k^*\right) \tag{5.28}$$

play a main role. So, the first thing to do is to obtain such values. On the one hand, regarding $\max_i|\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i)|$, one has the following decomposition:

$$\begin{aligned}
&\widetilde{m}_b(\chi)-\widetilde{\widehat{m}}_b(\chi)=\\
&m(\chi)-\sum_{i=1}^{n}w_b(\boldsymbol{\chi}_i,\chi)m(\boldsymbol{\chi}_i)-\widehat{m}_b(\chi)+\sum_{i=1}^{n}w_b(\boldsymbol{\chi}_i,\chi)\widehat{m}_b(\boldsymbol{\chi}_i)\leq\\
&\leq\sup_{\chi}|m(\chi)-\widehat{m}_b(\chi)|+\sup_{\chi}\sum_{i=1}^{n}w_b(\boldsymbol{\chi}_i,\chi)|m(\boldsymbol{\chi}_i)-\widehat{m}_b(\boldsymbol{\chi}_i)|.
\end{aligned}$$

Theorem 2 in Aneiros and Vieu (2008) gives that

$$\sup_{\chi}|m(\chi)-\widehat{m}_b(\chi)|=O\left(b^{\alpha}+\sqrt{\frac{\log n}{nF(b)}}\right)\ \text{a.s.} \tag{5.29}$$

while Lemma 23 in the Appendix A gives

$$\max_{i,j}|w_b(\boldsymbol{\chi}_i,\boldsymbol{\chi}_j)|=O\left(\frac{1}{nF(b)}\right)\ \text{a.s.} \tag{5.30}$$

In addition, using Assumption (5.5), it is easy to show, in a similar way as for obtaining (4.35), that

$$\max_j\#\{i=1,\ldots,n;\boldsymbol{\chi}_j\in B(\boldsymbol{\chi}_i,b)\}=O(nF(b))\ \text{a.s.} \tag{5.31}$$

Now, from (5.29), (5.30) and (5.31) one gets

$$\max_i|\widetilde{m}_b(\boldsymbol{\chi}_i)-\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i)|=O\left(b^{\alpha}+\sqrt{\frac{\log n}{nF(b)}}\right)\ a.s. \tag{5.32}$$

Now, (5.32) together with (34) in Aneiros-Pérez and Vieu (2008), that is:

$$\max_i |\widetilde{m}_b(\boldsymbol{\chi}_i)| = O\left(b^\alpha + F(b)^{-1} n^{-1/2+1/r} \log n\right), \tag{5.33}$$

gives that $\max_i |\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i)|$ and $\max_i |\widetilde{m}_b(\boldsymbol{\chi}_i)|$ have the same order a.s.:

$$\max_i |\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i)| = O\left(b^\alpha + F(b)^{-1} n^{-1/2+1/r} \log n\right) \ a.s. \tag{5.34}$$

On the other hand, regarding the order of (5.28), one can obtain it using remark 22 in Appendix A (which is related to Lemma 21), with $V_k = \varepsilon_k^*$. Remember that the hypothesis of this remark related the term V_k are given in Lemma 21: Let V_k be a zero-mean, stationary, α-mixing and real process, such that for some $r > 4$,

$$\max_{1\le k\le n} \mathbb{E}|V_k|^r \le \mathcal{C} < \infty.$$

While working with bootstrap estimators, one deals with expectation conditionally on the sample $\mathcal{S}$ and so, it is necessary to verify that, for some $r > 4$,

$$\max_{1\le k\le n} \mathbb{E}^{\mathcal{S}}|\varepsilon_k^*|^r \le \mathcal{C} < \infty. \tag{5.35}$$

In order to study (5.35), the following decomposition will be taken into account:

$$\widehat{\varepsilon}_{k,b} = \widetilde{\boldsymbol{X}}_k^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b) + \widetilde{m}_b(\boldsymbol{\chi}_k) - \sum_{j=1}^n w_b(\boldsymbol{\chi}_j, \boldsymbol{\chi}_k)\varepsilon_j + \varepsilon_k. \tag{5.36}$$

Denote:

$$D_b = \max_k |\widetilde{\boldsymbol{X}}_k^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b)| + \max_k |\widetilde{m}_b(\boldsymbol{\chi}_k)| + \max_k |\sum_{j=1}^n w_b(\boldsymbol{\chi}_j, \boldsymbol{\chi}_k)\varepsilon_j|. \tag{5.37}$$

On the one hand, (A.1) in Theorem 25 in Appendix A gives

$$|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b| = O_P(n^{-1/2}). \tag{5.38}$$

On the other hand, as

$$\frac{\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}}{n} = \sum_{i=1}^n \frac{\widetilde{\mathbf{X}}_i^T\widetilde{\mathbf{X}}_i}{n} \to \boldsymbol{B} \text{ a.s.}$$

(see Lemma 24 in Appendix A) one has, $\forall j \in \{1, \ldots, p\}$

$$\sum_{i=1}^{n} \frac{\widetilde{\mathbf{X}}_{i,j}^2}{n} \to b_{jj} \text{ a.s.} \Rightarrow \max_{1 \leq i \leq n} \frac{\widetilde{\mathbf{X}}_{i,j}^2}{n} = o(1) \text{ a.s.}$$

and so,

$$\max_{i,j} \frac{\widetilde{\mathbf{X}}_{i,j}^2}{n} = o(1) \text{ a.s.}$$

which implies that

$$\max_i |\widetilde{\boldsymbol{X}}_i|^2 = \max_i \sum_{j=1}^{p} \widetilde{\mathbf{X}}_{i,j}^2 \leq \max_i \max_j \widetilde{\mathbf{X}}_{i,j}^2 p = o(n)p \text{ a.s.} = o(n) \text{ a.s.}$$

and finally,

$$\max_k |\widetilde{\boldsymbol{X}}_k|^2 = o(n) \text{ a.s.} \tag{5.39}$$

(5.38) and (5.39) imply that

$$\max_k |\widetilde{\boldsymbol{X}}_k^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b)| = o_P(1) \tag{5.40}$$

In addition, applying Lemma 21 in Appendix A and taking into account (5.30), one has:

$$\max_k |\sum_{j=1}^{n} w_b(\boldsymbol{\chi}_j, \boldsymbol{\chi}_k)\varepsilon_j| = O(F(b)^{-1} n^{-1/2+1/r} \log n) \text{ a.s.} \tag{5.41}$$

(5.40), (5.41) and (5.33) allow to conclude that

$$D_b = o_P(1) \tag{5.42}$$

and as a consequence,

$$\widehat{\varepsilon}_{k,b} = \varepsilon_k + o_P(1) \text{ uniformly in } k. \tag{5.43}$$

Then, when studying the following term:

$$\mathbb{E}^{\mathcal{S}}(|\varepsilon_k^*|^r) = n^{-1} \sum_{i=1}^{n} |\widehat{\varepsilon}_{i,b} - \overline{\widehat{\varepsilon}}_b|^r \leq$$

$$cn^{-1} \sum_{i=1}^{n} \left(|\widehat{\varepsilon}_{i,b}|^r + |\overline{\widehat{\varepsilon}}_b|^r \right) = cn^{-1} \sum_{i=1}^{n} |\widehat{\varepsilon}_{i,b}|^r + c|\overline{\widehat{\varepsilon}}_b|^r \tag{5.44}$$

if one applies in (5.44) that $\widehat{\varepsilon}_{k,b} = \varepsilon_k + o_P(1)$, as

$$n^{-1}\sum_{i=1}^{n}\varepsilon_i \xrightarrow{p} 0 \text{ and } n^{-1}\sum_{i=1}^{n}|\varepsilon_i|^r \xrightarrow{p} \mathbb{E}(|\varepsilon_i|^r) \tag{5.45}$$

by the Weak Law of Large numbers, one obtains

$$\mathbb{E}^{\mathcal{S}}(|\varepsilon_k^*|^r) = O_P(1).$$

So, remark 22 can be applied directly to (5.28).

In this way, one obtains:

$$\max_i\left(\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k,\boldsymbol{\chi}_i)\varepsilon_k^*\right) = O_{P^{\mathcal{S}}}\left(F(b)^{-1}n^{-1/2+1/r}\log n\right). \tag{5.46}$$

Now, let's study the first term S_{n1}^* in (5.25). The decomposition:

$$\widetilde{X}_{i,j} = \widetilde{g}_{j,b}(\boldsymbol{\chi}_i) + \eta_{i,j} - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_i,\boldsymbol{\chi}_l)\eta_{l,j} \tag{5.47}$$

will be used along the proof, which allows to write the j-component of S_{n1}^*, $S_{n1,j}^*$ as:

$$\begin{aligned} S_{n1,j}^* &= \sum_{i=1}^{n}\widetilde{g}_{j,b}(\boldsymbol{\chi}_i)\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i) + \sum_{i=1}^{n}\eta_{i,j}\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i) - \sum_{i=1}^{n}(\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_i,\boldsymbol{\chi}_l)\eta_{l,j})\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i) \\ &= S_{n1,j1}^* + S_{n1,j2}^* - S_{n1,j3}^* \end{aligned} \tag{5.48}$$

and to study each component separately.

It starts with $S_{n1,j1}^*$:

$$\begin{aligned} &S_{n1,j1}^* = \sum_{i=1}^{n}\widetilde{g}_{j,b}(\boldsymbol{\chi}_i)\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i) \le n\max_i|\widetilde{g}_{j,b}(\boldsymbol{\chi}_i)|\max_i|\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i)| = \\ &n(O(b^\alpha + F(b)^{-1}n^{-1/2+1/r}\log n))^2 = O(nb^{2\alpha} + F(b)^{-2}n^{2/r}\log^2 n)\ a.s. \end{aligned} \tag{5.49}$$

About $S_{n1,j2}^*$, if one applies lemma 21 in Appendix A taking $a_{ik} = \widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i)$ ($\max|a_{ik}| = \max|\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i| = O\left(b^\alpha + F(b)^{-1}n^{-1/2+1/r}\log n\right) = O(a_n)$ a.s.) and $V_k = \eta_{i,j}$, $0.5 < \gamma < 1 - 9/(4a)$, one has that

$$S_{n1,j2}^* = O(b^\alpha n^{1/2+1/r}\log n + F(b)^{-1}n^{2/n}\log^2 n)\ a.s. \tag{5.50}$$

It remains to study the third term $S^*_{n1,j3}$, where

$$S^*_{n1,j3} = \sum_{i=1}^{n}(\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j})\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i) \leq n \max_i |\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j}| \max_i |\widetilde{\widehat{m}}_b(\boldsymbol{\chi}_i)|.$$

Lemma 21 can be applied again, with $a_{il} = w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)$ ($\max |w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)| = O(1/nF(b))$), $V_l = \eta_{l,j}$ and $0.5 < \gamma < 1 - 9/(4a)$, which gives that

$$\max |\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j}| = O(F(b)^{-1}n^{-1/2+1/r}\log n) \text{ a.s.}$$

This result, together with (5.34), implies:

$$S^*_{n1,j3} = O(F(b)^{-1}n^{1/2+1/r}\log n b^{\alpha} + F(b)^{-2}n^{2/r}\log^2 n) \ a.s. \tag{5.51}$$

Considering (5.49), (5.50) and (5.51) together with assumptions (5.14), one obtains:

$$S^*_{n1} = o(n^{1/2}) \ a.s. \tag{5.52}$$

Now, the j-component of S^*_{n2}, $S^*_{n2,j}$, will be studied. Using again the decomposition (5.47), one develops the following expression:

$$\begin{aligned} S^*_{n2,j} &= \sum_{i=1}^{n}\widetilde{g}_{j,b}(\boldsymbol{\chi}_i)\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon^*_k + \sum_{i=1}^{n}\eta_{i,j}\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon^*_k \\ &\quad - \sum_{i=1}^{n}(\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j})\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon^*_k = S^*_{n2,j1} + S^*_{n2,j2} - S^*_{n2,j3} \end{aligned}$$

First, $S^*_{n2,j1}$ is analysed:

$$\begin{aligned} S^*_{n2,j1} &= \sum_{i=1}^{n}\widetilde{g}_{j,b}(\boldsymbol{\chi}_i)\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon^*_k \leq \\ &\quad n \max_i |\widetilde{g}_{j,b}(\boldsymbol{\chi}_i)| \max_i |\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon^*_k| = \\ &\quad nO_{a.s.}\left(b^{\alpha} + F(b)^{-1}n^{-1/2+1/r}\log n\right) O_{P\mathcal{S}}\left(F(b)^{-1}n^{-1/2+1/r}\log n\right) = \\ &\quad O_{P\mathcal{S}}\left(b^{\alpha}F(b)^{-1}n^{1/2+1/r}\log n + F(b)^{-2}n^{2/r}\log^2 n\right) \end{aligned} \tag{5.53}$$

where it has been used (35) in Aneiros and Vieu (2008) and (5.46) above.

Applying Remark 22 on $S^*_{n2,j2}$, with $a_{i,k} = \sum_{k=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon^*_k$ and $V_k = \eta_{i,j}$, one gets:

$$\begin{aligned} S^*_{n2,j2} &= \sum_{i=1}^n \eta_{i,j} \sum_{k=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon^*_k = \\ & O_{P\mathcal{S}}(F(b)^{-1} n^{-1/2+1/r} \log n n^{1/2+1/r} \log n) = \\ & O_{P\mathcal{S}}(F(b)^{-1} n^{2/r} \log^2 n). \end{aligned} \tag{5.54}$$

For the last term of $S^*_{n2,j}$ one has:

$$\begin{aligned} S^*_{n2,j3} &= \sum_{i=1}^n (\sum_{l=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j}) \sum_{k=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon^*_k \leq \\ & n \max_i |\sum_{l=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j}| \max_i |\sum_{k=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_k)\varepsilon^*_k| = \\ & n O_{a.s.}(F(b)^{-1} n^{-1/2+1/r} \log n) O_{P\mathcal{S}}\left(F(b)^{-1} n^{-1/2+1/r} \log n\right) = \\ & O_{P\mathcal{S}}(F(b)^{-2} n^{2/r} \log^2 n) \end{aligned} \tag{5.55}$$

Finally, considering together the boundaries for the three elements of S^*_{n2}, (5.53), (5.54) and (5.55), one obtains:

$$S^*_{n2} = o_{P\mathcal{S}}(n^{1/2}). \tag{5.56}$$

In a similar way, one can study the j-th component of the third term S^*_{n3}, $S^*_{n3,j}$, applying the decomposition (5.47):

$$\begin{aligned} S^*_{n3,j} &= \sum_{i=1}^n \widetilde{g}_{j,b}(\boldsymbol{\chi}_i)\varepsilon^*_i + \sum_{i=1}^n \eta_{i,j}\varepsilon^*_i - \sum_{i=1}^n \left(\sum_{l=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j}\right)\varepsilon^*_i = \\ & S^*_{n3,j1} + S^*_{n3,j2} - S^*_{n3,j3} \end{aligned} \tag{5.57}$$

Remark 22 is applied to

$$S^*_{n3,j1} = \sum_{i=1}^n \widetilde{g}_{j,b}(\boldsymbol{\chi}_i)\varepsilon^*_i,$$

taking $V_i = \varepsilon^*_i$ and $a_{j,i} = \widetilde{g}_{j,b}(\boldsymbol{\chi}_i)$, as

$$\max_i |\widetilde{g}_{j,b}(\boldsymbol{\chi}_i)| = O\left(b^\alpha + F(b)^{-1} n^{-1/2+1/r} \log n\right) \text{ a.s.}$$

Then:

$$S^*_{n3,j1} = O_{P\mathcal{S}}(b^{\alpha} n^{1/2+1/r} \log n + F(b)^{-1} n^{2/r} \log^2 n). \tag{5.58}$$

In the same way, considering again $V_i = \varepsilon_i^*$ but $a_{j,i} = \sum_{l=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j}$ (remember that $\max |\sum_{l=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j}| = O(F(b)^{-1} n^{-1/2+1/r} \log n)$ a.s. was studied in (5.51)) one obtains that:

$$S^*_{n3,j3} = \sum_{i=1}^n \left(\sum_{l=1}^n w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_l)\eta_{l,j} \right) \varepsilon_i^* = O_{P\mathcal{S}}(F(b)^{-1} n^{2/r} \log^2 n). \tag{5.59}$$

Finally, from (5.57), (5.58) and (5.59) one obtains:

$$S^*_{n3} = \sum_{i=1}^n \boldsymbol{\eta}_i \varepsilon_i^* + o_{P\mathcal{S}}(n^{1/2}). \tag{5.60}$$

Now, from (5.52), (5.56) and (5.60) one has that (5.26) holds.

Finally, one must prove (5.27), where $\mathbf{C} = \sigma_\varepsilon^2 \mathbb{E}(\boldsymbol{\eta}_1 \boldsymbol{\eta}_1^T) = \sigma_\varepsilon^2 \mathbf{B}$ (see assumption (5.12)).

Three steps will be used to prove (5.27): first, seeing that its expectation, conditionally on the sample $\mathcal{S}$, is zero. Second, showing that the variance, also conditionally on the sample $\mathcal{S}$, converges to $\mathbf{C}$. Finally, showing that the expression given in (5.27) has, asymptotically, normal distribution.

On the one hand, as the bootstrapped errors in the naïve bootstrap procedure are the centred residuals, one has that

$$\mathbb{E}^{\mathcal{S}}(n^{-1/2} \sum_{i=1}^n \boldsymbol{\eta}_i \varepsilon_i^*) = n^{-1/2} \sum_{i=1}^n \boldsymbol{\eta}_i \mathbb{E}^{\mathcal{S}}(\varepsilon_i^*) = n^{-1/2} \sum_{i=1}^n \boldsymbol{\eta}_i n^{-1} \sum_{j=1}^n (\widehat{\varepsilon}_{j,b} - \overline{\widehat{\varepsilon}}_b) = \mathbf{0}.$$

On the other hand:

$$Var^{\mathcal{S}}(n^{-1/2} \sum_{i=1}^n \boldsymbol{\eta}_i \varepsilon_i^*) = n^{-1} \sum_{i=1}^n \boldsymbol{\eta}_i Var^{\mathcal{S}}(\varepsilon_i^*) \boldsymbol{\eta}_i^T = \widehat{\sigma}_\varepsilon^2 n^{-1} \sum_{i=1}^n \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T \to_P \mathbf{C},$$

where the convergence is obtained using that

$$\widehat{\sigma}_\varepsilon^2 = Var^{\mathcal{S}}(\varepsilon^*) = n^{-1} \sum_{j=1}^n (\widehat{\varepsilon}_{j,b} - \overline{\widehat{\varepsilon}}_b)^2 \to_P \sigma_\varepsilon^2$$

(see (5.43)) and, for the Strong Law of Large Numbers,

$$n^{-1}\sum_{i=1}^{n}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \to \mathbb{E}(\boldsymbol{\eta}_1\boldsymbol{\eta}_1^T) = \mathbf{B} \text{ a.s.}$$

Finally, one needs to prove the asymptotic normality of $n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\eta}_i\varepsilon_i^*$ (conditionally on the sample $\mathcal{S}$). Following the proof of Theorem 4 in Chapter 4, as in the expression (4.31), it will be enough to check Liapunov's condition; that is, we need to prove the following:

$$n^{-3/2}\sum_{i=1}^{n}\mathbb{E}^{\mathcal{S}}(|\mathbf{a}^T\boldsymbol{\eta}_i\varepsilon_i^*|^3) \to_P 0.$$

Taking into account that

$$n^{-3/2}\sum_{i=1}^{n}\mathbb{E}^{\mathcal{S}}(|\mathbf{a}^T\boldsymbol{\eta}_i\varepsilon_i^*|^3) = n^{-3/2}\sum_{i=1}^{n}|\mathbf{a}^T\boldsymbol{\eta}_i|^3\mathbb{E}^{\mathcal{S}}(|\varepsilon_i^*|^3) =$$
$$n^{-3/2}\sum_{i=1}^{n}|\mathbf{a}^T\boldsymbol{\eta}_i|^3 n^{-1}\sum_{k=1}^{n}|\widehat{\varepsilon}_{k,b}-\overline{\widehat{\varepsilon}}|^3 \tag{5.61}$$

the convergence to zero is achieved taking into account that

$$n^{-1}\sum_{i=1}^{n}|\mathbf{a}^T\boldsymbol{\eta}_i|^3 \to \mathbb{E}(|\mathbf{a}^T\boldsymbol{\eta}|^3) \text{ a.s.}$$

and

$$n^{-1}\sum_{k=1}^{n}|\widehat{\varepsilon}_{k,b}-\overline{\widehat{\varepsilon}}|^3 = O_P(1)$$

(see (5.43) and (5.45).

So, one has asymptotic normality of $n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\eta}_i\varepsilon_i^*$ in probability, conditionally on the sample $\mathcal{S}$, for the naïve bootstrap.

This concludes the proof of the theorem. □

Proof Theorem 11

Firsts steps of the proof of Theorem 10 can be followed in this case, obtaining the same decomposition in $T_1(y)$ and $T_2(y)$ (5.18). As in Theorem 10 for the naïve bootstrap, now with wild bootstrap it is needed to prove that

$$T_1(y) \longrightarrow_P 0 \text{ for any fixed value of } y \tag{5.62}$$

and

$$T_2(y) \longrightarrow 0 \text{ for any fixed value of } y. \tag{5.63}$$

Note that $T_2(y)$ does not depend on the bootstrap procedure. So, from the study given in the proof of Theorem 10 one has that (5.63) is already verified. Then, it is only needed to prove (5.62) and this will be done by the same technique as (5.24). Thus, one needs to prove the following convergence:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}}_b) \xrightarrow{d} N(0, \mathbf{A}), \text{ in probability conditionally on } \mathcal{S}. \tag{5.64}$$

One can apply in (5.64) the same decomposition used in Theorem 10 to obtain (5.26), and thus, proving that:

$$S_{n1}^* - S_{n2}^* + S_{n3}^* = \sum_{i=1}^{n} \boldsymbol{\eta}_i \varepsilon_i^* + o_{Ps}(n^{1/2}) \tag{5.65}$$

and

$$n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\eta}_i \varepsilon_i^* \xrightarrow{d} N(0, \mathbf{C}), \text{ in probability conditionally on } \mathcal{S}, \tag{5.66}$$

where $\mathbf{C} = \sigma_\varepsilon^2 \mathbb{E}(\boldsymbol{\eta}_1 \boldsymbol{\eta}_1^T) = \sigma_\varepsilon^2 \mathbf{B}$ (see Assumption (5.12)), one obtains, together with Lemma 24 in Appendix A, the asymptotic normality of (5.64).

In order to study (5.65), the order in probability, conditionally on the sample $\mathcal{S}$, of

$$\max_i \left(\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i) \varepsilon_k^* \right)$$

plays, again, a main role. So, first one needs to prove that:

$$\max_i \left(\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i) \varepsilon_k^* \right) = O\left(F(b)^{-1} n^{-1/2+1/r} \log n \sqrt{\log \log n} \right) \text{ a.s.,} \tag{5.67}$$

conditionally on the sample $\mathcal{S}$.

Decomposition given in (5.36) can be included into the expression of (5.67), resulting:

$$\begin{aligned}
&\max_i \left(|\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i)\varepsilon_k^*| \right) = \max_i \left(|\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i) V_k \widehat{\varepsilon}_{k,b}| \right) \\
&\leq \max_i |\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i) V_k \varepsilon_k| + \max_i |\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i) V_k| \times \\
&\times \left(\max_k |\widetilde{\boldsymbol{X}}_k^T (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b)| + \max_k |\widetilde{m}_b(\boldsymbol{\chi}_k)| + \max_k |\sum_{j=1}^{n} w_b(\boldsymbol{\chi}_j, \boldsymbol{\chi}_k)\varepsilon_j| \right) \\
&= \max_i |\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i) V_k \varepsilon_k| + \max_i |\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i) V_k| \times D_b \qquad (5.68)
\end{aligned}$$

where D_b has been defined in (5.37).

When dealing with D_b in Theorem 10, the result (A.1) in Theorem 25 in Appendix A will be used in order to obtain that $D_b = o_P(1)$ and so, $\widehat{\varepsilon}_{k,b} = \varepsilon_k + o_P(1)$ uniformly in k. Meanwhile, in this case, we can apply result (A.2) in the same Theorem 25 in Appendix A giving

$$|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b| = O(n^{-1/2}\sqrt{\log\log n}) \text{ a.s.}$$

This result, together with the same study of the other terms in (5.37), allows to bound D_b but using almost sure convergence instead of convergence in probability:

$$D_b = o(\sqrt{\log\log n}) \text{ a.s.} \qquad (5.69)$$

In addition, one has that:

$$\max_k \mathbb{E}^{\mathcal{S}}(|V_k|^r) = \max_k \mathbb{E}(|V_k|^r) < C < \infty,$$

(see assumptions on V_k), which, together with the fact that the errors ε_k are bounded, gives

$$\max_k \mathbb{E}^{\mathcal{S}}(|V_k\varepsilon_k|^r) = \max_k \mathbb{E}(|V_k|^r)|\varepsilon_k|^r < C < \infty$$

Thus, from Lemma 21 in Appendix A one obtains that

$$\begin{aligned}
&\max_i |\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i) V_k| = O\left(F(b)^{-1} n^{-1/2+1/r} \log n\right) \text{ a.s.}, \\
&\text{conditionally on the sample } \mathcal{S}. \qquad (5.70)
\end{aligned}$$

and

$$\max_i |\sum_{k=1}^{n} w_b(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i) V_k \varepsilon_k| = O\left(F(b)^{-1} n^{-1/2+1/r} \log n\right) \text{ a.s.,}$$
conditionally on the sample $\mathcal{S}$. (5.71)

Now, from (5.68)-(5.71), one obtains (5.67).

Focus now on (5.65). As S_{n1}^* is not affected by the bootstrap procedure, one obtains directly from (5.52):

$$S_{n1}^* = o(n^{1/2}) \text{ a.s.} \tag{5.72}$$

Now, considering (5.67) and using similar techniques as those for S_{n2} in Theorem 10, it holds that

$$S_{n2}^* = o_{P^{\mathcal{S}}}(n^{1/2}) \tag{5.73}$$

Note that for the case of the wild bootstrap, in order to obtain (5.73), some slight modification in the assumptions on $(n, b, F(b))$ must be done because the term $\sqrt{\log \log n}$ in (5.67), as it was stated in the enunciate of the Theorem.

In a similar way (and considering, again, some slight modification in the assumptions on $(n, b, F(b))$), one obtains that

$$S_{n3}^* = \sum_{i=1}^{n} \boldsymbol{\eta}_i \varepsilon_i^* + o_{P^{\mathcal{S}}}(n^{1/2}) \tag{5.74}$$

Now, from (5.72), (5.73) and (5.74) one has that (5.65) holds.

Finally, one must prove (5.66).

Here, the same outline that was used in Theorem 10 to prove the normality of (5.27) can be followed. That is, to prove, for the expression given in (5.66), that its expectation (conditionally on $\mathcal{S}$) is zero, its variance also conditionally on $\mathcal{S}$ converges in probability to $\mathbf{C}$ and that it is asymptotically normally distributed using Liapunov's condition.

On the one hand, as $\mathbb{E}^{\mathcal{S}}(\varepsilon_i^*) = \mathbb{E}^{\mathcal{S}}(\widehat{\varepsilon}_{i,b} V_i) = \widehat{\varepsilon}_{i,b} \mathbb{E}^{\mathcal{S}}(V_i)$ and $\mathbb{E}^{\mathcal{S}}(V_i) = \mathbb{E}(V_i) = 0$, one has that

$$\mathbb{E}^{\mathcal{S}}(n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\eta}_i \varepsilon_i^*) = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\eta}_i \mathbb{E}^{\mathcal{S}}(\varepsilon_i^*) = \mathbf{0}.$$

On the other hand, it is necessary to prove that:

$$Var^{\mathcal{S}}(n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\eta}_i\varepsilon_i^*) = n^{-1}\sum_{i=1}^{n}\boldsymbol{\eta}_i Var^{\mathcal{S}}(\varepsilon_i^*)\boldsymbol{\eta}_i^T = n^{-1}\sum_{i=1}^{n}\widehat{\varepsilon}_{i,b}^2\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \rightarrow_P \mathbf{C}\ . \tag{5.75}$$

Noting that

$$\widehat{\varepsilon}_{i,b} = \widetilde{\boldsymbol{X}}_i^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b) + \widetilde{m}_b(\boldsymbol{\chi}_i) - \sum_{j=1}^{n} w_b(\boldsymbol{\chi}_j, \boldsymbol{\chi}_i)\varepsilon_j + \varepsilon_i = \delta_i + \varepsilon_i,$$

where it is denoted $\delta_i = \widetilde{\boldsymbol{X}}_i^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b) + \widetilde{m}_b(\boldsymbol{\chi}_i) - \sum_{j=1}^{n} w_b(\boldsymbol{\chi}_j, \boldsymbol{\chi}_i)\varepsilon_j$, one can write

$$n^{-1}\sum_{i=1}^{n}\widehat{\varepsilon}_{i,b}^2\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T = n^{-1}\sum_{i=1}^{n}(\delta_i+\varepsilon_i)^2\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T = n^{-1}\sum_{i=1}^{n}(\delta_i^2 + 2\delta_i\varepsilon_i + \varepsilon_i^2)\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T =$$
$$n^{-1}\sum_{i=1}^{n}\delta_i^2\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T + n^{-1}\sum_{i=1}^{n}2\delta_i\varepsilon_i\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T + n^{-1}\sum_{i=1}^{n}\varepsilon_i^2\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T = \Delta_1 + \Delta_2 + \Delta_3.$$

Now, taking into account that $\max_i|\delta_i| \leq D_b = o_P(1)$ (see (5.42)) and applying the Strong Law of Large Numbers (note that $\boldsymbol{\eta}_i$ has $r > 4$ finite moments), one obtains

$$\begin{aligned} |\Delta_1| &\leq n^{-1}\sum_{i=1}^{n}|\delta_i|^2|\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T| \leq \\ &\max_k|\delta_k|^2 n^{-1}\sum_{i=1}^{n}|\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T| = o_P(1)O_P(1) = o_P(1). \end{aligned}$$

If, in addition, one uses the fact that $|\varepsilon_i| \leq$C$< \infty$, then

$$\begin{aligned} |\Delta_2| &\leq n^{-1}\sum_{i=1}^{n}|2\delta_i\varepsilon_i||\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T| \leq \\ &2\max_k|\delta_k|Cn^{-1}\sum_{i=1}^{n}|\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T| = o_P(1)O_P(1) = o_P(1). \end{aligned}$$

Finally, the Strong Law of Large numbers gives

$$\begin{aligned} \Delta_3 &= n^{-1}\sum_{i=1}^{n}\varepsilon_i^2\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \longrightarrow_P \mathbb{E}(\varepsilon_1^2\boldsymbol{\eta}_1\boldsymbol{\eta}_1^T) = \\ &\mathbb{E}(\varepsilon_1^2)\mathbb{E}(\boldsymbol{\eta}_1\boldsymbol{\eta}_1^T) = \sigma_\varepsilon^2\mathbb{E}(\boldsymbol{\eta}_1\boldsymbol{\eta}_1^T) = \boldsymbol{C} \text{ a.s.} \end{aligned}$$

where in the last but one equality, the independence between ε_i and $\boldsymbol{\eta}_i$ was used (see assumption (5.8)). So, it has been proved that (5.75) is true.

Finally, one needs to prove that $n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\eta}_i\varepsilon_i^*$ is, asymptotically, normal distributed in probability, conditionally on the sample $\mathcal{S}$. Again, Liapunov's condition is used. Then, it must be verified that:

$$n^{-3/2}\sum_{i=1}^{n}\mathbb{E}^{\mathcal{S}}(|\mathbf{a}^T\boldsymbol{\eta}_i\varepsilon_i^*|^3) = n^{-3/2}\sum_{i=1}^{n}|\mathbf{a}^T\boldsymbol{\eta}_i|^3\mathbb{E}^{\mathcal{S}}(|\varepsilon_i^*|^3) \rightarrow_P 0.$$

Using that $\max_i|\widehat{\varepsilon}_{i,b}| = o_P(1)$ (remember that $\widehat{\varepsilon}_{i,b} = \delta_i + \varepsilon_i$ with $|\varepsilon_i| <$C and $\max_i|\delta_i| = o_P(1)$), assumptions on V_i and applying Strong Law of Large Numbers (note that $\boldsymbol{\eta}_i$ has more than 6 finite moments), one has that

$$\begin{aligned} n^{-3/2}\sum_{i=1}^{n}|\mathbf{a}^T\boldsymbol{\eta}_i|^3\mathbb{E}^{\mathcal{S}}(|\varepsilon_i^*|^3) &= \mathbb{E}(|V_1|^3)n^{-3/2}\sum_{i=1}^{n}|\mathbf{a}^T\boldsymbol{\eta}_i|^3|\widehat{\varepsilon}_{i,b}|^3 \\ \leq o_P(1)\mathbb{E}(|V_1|^3)n^{-3/2}\sum_{i=1}^{n}|\mathbf{a}^T\boldsymbol{\eta}_i|^3 &= o_P(1)\mathbb{E}(|V_1|^3)o_P(1) \rightarrow_P 0. \end{aligned} \quad (5.76)$$

So, one obtains the asymptotic normality of $n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\eta}_i\varepsilon_i^*$ in probability, conditionally on the sample $\mathcal{S}$, for the wild bootstrap and thus, (5.66) is verified. This concludes the proof of the theorem. □

Proof of Theorem 14

The proof begins with the study of the second term of the expression in Theorem 14, which is related to the convergence of the estimator $\widehat{m}$ and then, it follows with the study of the term related to the bootstrap estimator.

Following decomposition is developed:

$$
\begin{aligned}
&(nF_\chi(h))^{1/2}(\widehat{m}_h(\chi) - m(\chi)) = \\
&= (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_h) - m(\chi))) = \\
&= (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)(\boldsymbol{X}_i^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \varepsilon_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_h) - m(\chi))) = \\
&= (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)(m(\boldsymbol{\chi}_i) + \varepsilon_i) - m(\chi)) - \\
&-(nF_\chi(h))^{1/2}\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)\boldsymbol{X}_i^T(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}) = \\
&= S_1(\chi) - S_2(\chi)
\end{aligned}
\tag{5.77}
$$

In last expression, $S_1(\chi)$ includes the nonparametric part of the model, meanwhile, $S_2(\chi)$ represents the linear/parametric part. Then, each term can be studied separately.

In order to manage the term $S_1(\chi)$, Theorem 28 in the Appendix A will be applied (remember that the bias term is cancelled by assumption (4.9) but considering the following "auxiliary nonparametric model":

$$\mathcal{Y} = m(\boldsymbol{\chi}) + \varepsilon. \tag{5.78}$$

(Note that $\mathbb{E}(\varepsilon|\boldsymbol{\chi}) = \mathbb{E}(\varepsilon|\boldsymbol{X}, \boldsymbol{\chi}) = 0$ and $\mathbb{E}(\varepsilon^2|\boldsymbol{\chi}) = \mathbb{E}(\varepsilon^2|\mathbf{X}, \boldsymbol{\chi})$).

Denote

$$\widehat{m}_h^{NP}(\chi) = \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)\mathcal{Y}_i = \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)(m(\boldsymbol{\chi}_i) + \varepsilon_i).$$

Then, one obtains

$$S_1(\chi) = (nF_\chi(h))^{1/2}(\widehat{m}_h^{NP}(\chi) - m(\chi)). \tag{5.79}$$

Focus now on the linear part, $S_2(\chi)$. From Lemma 23 in Appendix A, it holds:

$$\max|w_h(\boldsymbol{\chi}_i, \chi)| = O((nF(h))^{-1}) \text{ a.s.} \tag{5.80}$$

Taking into account (4.35) together with (5.80), Theorem 25 in Appendix A and (5.15), one obtains

$$S_2(\chi) = O((F(h)\log\log n)^{1/2}) \text{ a.s.}$$

In addition, from the fact that $F(h)\log\log n \to 0$ (see (4.19)) one has that

$$S_2(\chi) = o(1) \text{ a.s.} \tag{5.81}$$

Finally, (5.77) together with (5.79) and (5.81) gives

$$(nF_\chi(h))^{1/2}(\widehat{m}_h(\chi) - m(\chi)) = (nF_\chi(h))^{1/2}(\widehat{m}_h^{NP}(\chi) - m(\chi)) + o(1) \text{ a.s.} \tag{5.82}$$

Next part of the proof will focus on the bootstrap term of the theorem. Specifically, it will find the relationship between the bootstrap estimator of m from the nonparametric model (5.78) and the corresponding bootstrap estimator from the SFPL regression model (5.1). One can establish the naïve bootstrap procedure for the model (5.78) as in Section 4.2:

Step 1: Construct the residuals $\widehat{e}_{i,b} = \mathcal{Y}_i - \widehat{m}_b^{NP}(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$.
Step 2: Draw e_i^*, for $i = 1, \ldots, n$, randomly from the empirical distribution of $(\widehat{e}_{1,b} - \bar{\widehat{e}}_b, \ldots, \widehat{e}_{n,b} - \bar{\widehat{e}}_b)$, where

$$\bar{\widehat{e}}_b = \frac{1}{n}\sum_{k=1}^{n} \widehat{e}_{k,b}$$

Step 3: Obtain $\mathcal{Y}_i^* = \widehat{m}_b^{NP}(\boldsymbol{\chi}_i) + e_i^*$, $i = 1, \ldots, n$.
Step 4: Define $\widehat{m}_{hb}^{*NP}(\chi) = \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)\mathcal{Y}_i^*$.

Taking into account that

$$\begin{aligned} \widehat{e}_{i,b} &= \mathcal{Y}_i - \widehat{m}_b^{NP}(\boldsymbol{\chi}_i) = m(\boldsymbol{\chi}_i) + \varepsilon_i - \widehat{m}_b^{NP}(\boldsymbol{\chi}_i) = \\ &\quad m(\boldsymbol{\chi}_i) + \varepsilon_i - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l) \end{aligned} \tag{5.83}$$

and denoting $j(i)$ as a value chosen at random in $\{1, \ldots, n\}$, one has that

$$\begin{aligned} e_i^* = &\quad \widehat{e}_{j(i),b} - \frac{1}{n}\sum_{k=1}^{n} \widehat{e}_{k,b} = \\ = &\quad m(\boldsymbol{\chi}_{j(i)}) + \varepsilon_{j(i)} - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_{j(i)})(m(\boldsymbol{\chi}_l) + \varepsilon_l) - \\ &\quad - \frac{1}{n}\sum_{k=1}^{n}(m(\boldsymbol{\chi}_k) + \varepsilon_k - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_k)(m(\boldsymbol{\chi}_l) + \varepsilon_l)). \end{aligned} \tag{5.84}$$

Then,

$$\begin{aligned}\mathcal{Y}_i^* = \quad & \widehat{m}_b^{NP}(\boldsymbol{\chi}_i) + e_i^* = \\ = \quad & \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l) + m(\boldsymbol{\chi}_{j(i)}) + \varepsilon_{j(i)} - \\ & - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_{j(i)})(m(\boldsymbol{\chi}_l) + \varepsilon_l) - \\ & - \frac{1}{n}\sum_{k=1}^{n}(m(\boldsymbol{\chi}_k) + \varepsilon_k - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_k)(m(\boldsymbol{\chi}_l) + \varepsilon_l)). \qquad (5.85)\end{aligned}$$

Therefore

$$\begin{aligned}\widehat{m}_{hb}^{*NP}(\chi) = \quad & \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)\mathcal{Y}_i^* = \\ = \quad & \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l) + m(\boldsymbol{\chi}_{j(i)}) + \varepsilon_{j(i)} - \\ & - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_{j(i)})(m(\boldsymbol{\chi}_l) + \varepsilon_l) - \\ & - \frac{1}{n}\sum_{k=1}^{n}(m(\boldsymbol{\chi}_k) + \varepsilon_k - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_k)(m(\boldsymbol{\chi}_l) + \varepsilon_l))]. \qquad (5.86)\end{aligned}$$

Then, it can be written:

$$\begin{aligned}& (nF_\chi(h))^{1/2}(\widehat{m}_{hb}^{*NP}(\chi) - \widehat{m}_b^{NP}(\chi)) = \\ & (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l) + \\ & + m(\boldsymbol{\chi}_{j(i)}) + \varepsilon_{j(i)} - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_{j(i)})(m(\boldsymbol{\chi}_l) + \varepsilon_l) - \\ & - \frac{1}{n}\sum_{k=1}^{n}(m(\boldsymbol{\chi}_k) + \varepsilon_k - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_k)(m(\boldsymbol{\chi}_l) + \varepsilon_l))] - \\ & \sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i, \chi)(m(\boldsymbol{\chi}_i) + \varepsilon_i)) \qquad (5.87)\end{aligned}$$

Focus now on the bootstrap procedure in the SFPL model (5.1). One has that

$$
\begin{aligned}
&(nF_{\chi}(h))^{1/2}(\widehat{m}^{*}_{hb}(\chi) - \widehat{m}_{b}(\chi)) = \\
&(nF_{\chi}(h))^{1/2}(\sum_{i=1}^{n} w_{h}(\boldsymbol{\chi}_{i},\chi)(Y_{i}^{*} - \boldsymbol{X}_{i}^{T}\widehat{\boldsymbol{\beta}}_{b}^{*}) - \widehat{m}_{b}(\chi)) = \\
&= (nF_{\chi}(h))^{1/2}(\sum_{i=1}^{n} w_{h}(\boldsymbol{\chi}_{i},\chi)(\boldsymbol{X}_{i}^{T}\widehat{\boldsymbol{\beta}}_{b} + \widehat{m}_{b}(\boldsymbol{\chi}_{i}) + \varepsilon_{i}^{*} - \boldsymbol{X}_{i}^{T}\widehat{\boldsymbol{\beta}}_{b}^{*}) - \widehat{m}_{b}(\chi)) = \\
&= (nF_{\chi}(h))^{1/2}(\sum_{i=1}^{n} w_{h}(\boldsymbol{\chi}_{i},\chi)(\widehat{m}_{b}(\boldsymbol{\chi}_{i}) + \varepsilon_{i}^{*}) - \widehat{m}_{b}(\chi)) - \\
&-(nF_{\chi}(h))^{1/2}\sum_{i=1}^{n} w_{h}(\boldsymbol{\chi}_{i},\chi)\boldsymbol{X}_{i}^{T}(\widehat{\boldsymbol{\beta}}_{b}^{*} - \widehat{\boldsymbol{\beta}}_{b}) = \\
&= S_{1}^{*}(\chi) - S_{2}^{*}(\chi)
\end{aligned}
\tag{5.88}
$$

Note that the bootstrapped errors, ε_{i}^{*}, will be drawn from the empirical distribution function of the following terms:

$$
\begin{aligned}
\widehat{\varepsilon}_{j,b} - \bar{\widehat{\varepsilon}} = \quad & Y_{j} - \boldsymbol{X}_{j}^{T}\widehat{\boldsymbol{\beta}}_{b} - \widehat{m}_{b}(\boldsymbol{\chi}_{j}) - \frac{1}{n}\sum_{k=1}^{n}(Y_{k} - \boldsymbol{X}_{k}^{T}\widehat{\boldsymbol{\beta}}_{b} - \widehat{m}_{b}(\boldsymbol{\chi}_{k})) = \\
= \quad & Y_{j} - \boldsymbol{X}_{j}^{T}\widehat{\boldsymbol{\beta}}_{b} - \sum_{l=1}^{n} w_{b}(\boldsymbol{\chi}_{l},\boldsymbol{\chi}_{j})(Y_{l} - \boldsymbol{X}_{l}^{T}\widehat{\boldsymbol{\beta}}_{b}) - \\
& -\frac{1}{n}\sum_{k=1}^{n}(Y_{k} - \boldsymbol{X}_{k}^{T}\widehat{\boldsymbol{\beta}}_{b} - \sum_{l=1}^{n} w_{b}(\boldsymbol{\chi}_{l},\boldsymbol{\chi}_{k})(Y_{l} - \boldsymbol{X}_{l}^{T}\widehat{\boldsymbol{\beta}}_{b}))
\end{aligned}
$$

Then, denoting $\varepsilon_i^* = \widehat{\varepsilon}_{j(i),b} - \bar{\widehat{\varepsilon}}$ one can develop the term $S_1^*(\chi)$ as follows:

$$\begin{aligned}
S_1^*(\chi) = \quad & (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)(\widehat{m}_b(\boldsymbol{\chi}_i)+\varepsilon_i^*) - \widehat{m}_b(\chi)) = \\
= \quad & (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)(\widehat{m}_b(\boldsymbol{\chi}_i)+\widehat{\varepsilon}_j - \bar{\widehat{\varepsilon}}) - \widehat{m}_b(\chi)) = \\
= \quad & (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)(Y_l - \boldsymbol{X}_l^T\widehat{\boldsymbol{\beta}}_b) + \\
& + Y_{j(i)} - \boldsymbol{X}_{j(i)}^T\widehat{\boldsymbol{\beta}}_b - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_{j(i)})(Y_l - \boldsymbol{X}_l^T\widehat{\boldsymbol{\beta}}_b) - \\
& - \frac{1}{n}\sum_{k=1}^{n}(Y_k - \boldsymbol{X}_k^T\widehat{\boldsymbol{\beta}}_b - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_k)(Y_l - \boldsymbol{X}_l^T\widehat{\boldsymbol{\beta}}_b))] - \\
& - \sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i,\chi)(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b))
\end{aligned}$$

Using the definition of Y_i (5.1) in last expressions it holds:

$$\begin{aligned}
& S_1^*(\chi) = \\
& (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)(\boldsymbol{X}_l^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_l) + \varepsilon_l - \boldsymbol{X}_l^T\widehat{\boldsymbol{\beta}}_b) + \\
& + \boldsymbol{X}_{j(i)}^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_{j(i)}) + \varepsilon_{j(i)} - \boldsymbol{X}_{j(i)}^T\widehat{\boldsymbol{\beta}}_b - \\
& \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_{j(i)})(\boldsymbol{X}_l^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_l) + \varepsilon_l - \boldsymbol{X}_l^T\widehat{\boldsymbol{\beta}}_b) - \\
& - \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{X}_k^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_k) + \varepsilon_k - \boldsymbol{X}_k^T\widehat{\boldsymbol{\beta}}_b - \\
& \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_k)(\boldsymbol{X}_l^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_l) + \varepsilon_l - \boldsymbol{X}_l^T\widehat{\boldsymbol{\beta}}_b))] - \\
& - \sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i,\chi)(\boldsymbol{X}_i^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \varepsilon_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b))
\end{aligned}$$

Now, last expressions can be divided into two different parts, grouping terms related to $m(\chi) + \varepsilon$ (from the "auxiliary nonparametric model") and

the terms related to the linear part of the model ($\boldsymbol{\beta}$), resulting:

$$
\begin{aligned}
S_1^*(\chi) = \quad & (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l)+\varepsilon_l)+ \\
& +m(\boldsymbol{\chi}_{j(i)})+\varepsilon_{j(i)}-\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_{j(i)})(m(\boldsymbol{\chi}_l)+\varepsilon_l) \\
& -\frac{1}{n}\sum_{k=1}^{n}(m(\boldsymbol{\chi}_k)+\varepsilon_k-\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_k)(m(\boldsymbol{\chi}_l)+\varepsilon_l))]- \\
& -\sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i,\chi)(m(\boldsymbol{\chi}_i)+\varepsilon_i))+ \\
& +(nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)+ \\
& \boldsymbol{X}_{j(i)}^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)-\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_{j(i)})\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)- \\
& \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{X}_k^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)-\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_k)\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b))]- \\
& \sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i,\chi)\boldsymbol{X}_i^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b))= \\
& S_{1,1}^*(\chi)+S_{1,2}^*(\chi)
\end{aligned}
\tag{5.89}
$$

where,

$$
S_{1,1}^*(\chi) = \quad (nF_\chi(h))^{1/2}(\widehat{m}_{hb}^{*NP}(\chi)-\widehat{m}_b^{NP}(\chi)) \tag{5.90}
$$

and

$$
\begin{aligned}
S_{1,2}^*(\chi) = \quad & (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)+ \\
& \boldsymbol{X}_{j(i)}^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)-\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_{j(i)})\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)- \\
& -\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{X}_k^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)-\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_k)\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b))]- \\
& \sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i,\chi)\boldsymbol{X}_i^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b))
\end{aligned}
$$

From (5.80), (4.35), (5.15) and (A.2) in Theorem 25 in the Appendix A, it is obtained that

$$S^*_{1,2}(\chi) = O((F(h)\log\log n)^{1/2}) \text{ a.s.}$$

Then, taking into account that $F(h)\log\log n \to 0$ (see (4.19)) one has that

$$S^*_{1,2}(\chi) = o(1) \text{ a.s.} \tag{5.91}$$

Finally, the term $S^*_2(\chi)$ is studied. From Theorem 10 and Theorem 25 one has

$$\widehat{\boldsymbol{\beta}}^*_b - \widehat{\boldsymbol{\beta}}_b = O_{P\mathcal{S}}(n^{-1/2}).$$

This fact allows to follow similar reasoning as those used in the study of $S_2(\chi)$, giving

$$S^*_2(\chi) = o_{P\mathcal{S}}(1). \tag{5.92}$$

From (5.88), (5.89), (5.90), (5.91) and (5.92), one obtains

$$(nF_\chi(h))^{1/2}(\widehat{m}^*_{hb}(\chi) - \widehat{m}_b(\chi)) = (nF_\chi(h))^{1/2}(\widehat{m}^{*NP}_{hb}(\chi) - \widehat{m}^{NP}_b(\chi)) + o_{P\mathcal{S}}(1). \tag{5.93}$$

Finally, (5.82), (5.93) and Theorem 4 give the result of Theorem 14. □

Proof of Theorem 15

The proof for this Theorem will follow the same outline as in Theorem 14.

First part of the proof for the Theorem 14 can be applied in this case, since bootstrap methodology does not affect the study of the term

$$(nF_\chi(h))^{1/2}(\widehat{m}_h(\chi) - m(\chi)),$$

so, one can obtain (5.82) following the same procedure as in Theorem 14.

Again, the "auxiliary nonparametric model" defined in (5.78) is used and the proof will focus on the bootstrap part of the Theorem, analysing the relationship between the bootstrap estimator of m from the nonparametric model (5.78) and the corresponding bootstrap estimator from the SFPL regression model (5.1).

Wild bootstrap procedure developed in Section 4.2 can be established for the model (5.78) , changing only the second step with respect to the naïve bootstrap procedure:

Step 1: Construct the residuals $\widehat{e}_{i,b} = \mathcal{Y}_i - \widehat{m}_b^{NP}(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$.

Step 2: Define $e_i^* = V_i \widehat{e}_{i,b}$, $i = 1, \ldots, n$, where V_i verify the assumptions established in Section 4.2.

Step 3: Obtain $\mathcal{Y}_i^* = \widehat{m}_b^{NP}(\boldsymbol{\chi}_i) + e_i^*$, $i = 1, \ldots, n$.

Step 4: Define $\widehat{m}_{hb}^{*NP}(\chi) = \sum_{i=1}^n w_h(\boldsymbol{\chi}_i, \chi)\mathcal{Y}_i^*$.

Taking into account the development for the term $\widehat{e}_{i,b}$ in (5.83), one has:

$$e_i^* = V_i \widehat{e}_{i,b} = V_i(m(\boldsymbol{\chi}_i) + \varepsilon_i - \sum_{l=1}^n w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l)).$$

Then,

$$\mathcal{Y}_i^* = \widehat{m}_b^{NP}(\boldsymbol{\chi}_i) + e_i^* =$$
$$\sum_{l=1}^n w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l) + V_i(m(\boldsymbol{\chi}_i) + \varepsilon_i - \sum_{l=1}^n w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l)).$$

Therefore

$$\widehat{m}_{hb}^{*NP}(\chi) = \sum_{i=1}^n w_h(\boldsymbol{\chi}_i, \chi)\mathcal{Y}_i^* =$$
$$= \sum_{i=1}^n w_h(\boldsymbol{\chi}_i, \chi)[\sum_{l=1}^n w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l) +$$
$$+V_i(m(\boldsymbol{\chi}_i) + \varepsilon_i - \sum_{l=1}^n w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l))].$$

Then, one has the following expression:

$$(nF_\chi(h))^{1/2}(\widehat{m}_{hb}^{*NP}(\chi) - \widehat{m}_b^{NP}(\chi)) =$$
$$= (nF_\chi(h))^{1/2}(\sum_{i=1}^n w_h(\boldsymbol{\chi}_i, \chi)[\sum_{l=1}^n w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l) +$$
$$+V_i(m(\boldsymbol{\chi}_i) + \varepsilon_i - \sum_{l=1}^n w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l) + \varepsilon_l))] -$$
$$- \sum_{i=1}^n w_b(\boldsymbol{\chi}_i, \chi)(m(\boldsymbol{\chi}_i) + \varepsilon_i))$$

Focussing on the bootstrap procedure in the SFPL regression model (5.1), one has

$$
\begin{aligned}
&(nF_{\chi}(h))^{1/2}(\widehat{m}^{*}_{hb}(\chi) - \widehat{m}_{b}(\chi)) = \\
&(nF_{\chi}(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)(Y_i^{*} - \boldsymbol{X}_i^{T}\widehat{\boldsymbol{\beta}}_b^{*}) - \widehat{m}_b(\chi)) = \\
&= (nF_{\chi}(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)(\boldsymbol{X}_i^{T}\widehat{\boldsymbol{\beta}}_b + \widehat{m}_b(\boldsymbol{\chi}_i) + \varepsilon_i^{*} - \boldsymbol{X}_i^{T}\widehat{\boldsymbol{\beta}}_b^{*}) - \widehat{m}_b(\chi)) = \\
&= (nF_{\chi}(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)(\widehat{m}_b(\boldsymbol{\chi}_i) + \varepsilon_i^{*}) - \widehat{m}_b(\chi)) - \\
&-(nF_{\chi}(h))^{1/2}\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)\boldsymbol{X}_i^{T}(\widehat{\boldsymbol{\beta}}_b^{*} - \widehat{\boldsymbol{\beta}}_b) = \\
&= S_1^{*}(\chi) - S_2^{*}(\chi),
\end{aligned}
\tag{5.94}
$$

where,

$$
\begin{aligned}
S_1^{*}(\chi) = \quad & (nF_{\chi}(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)(\widehat{m}_b(\boldsymbol{\chi}_i) + \varepsilon_i^{*}) - \widehat{m}_b(\chi)) \\
= \quad & (nF_{\chi}(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)(Y_l - \boldsymbol{X}_l^{T}\widehat{\boldsymbol{\beta}}_b) + \\
& + V_i(Y_i - \boldsymbol{X}_i^{T}\widehat{\boldsymbol{\beta}}_b - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)(Y_l - \boldsymbol{X}_l^{T}\widehat{\boldsymbol{\beta}}_b)] - \\
& - \sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i,\chi)(Y_i - \boldsymbol{X}_i^{T}\widehat{\boldsymbol{\beta}}_b)).
\end{aligned}
$$

Replacing the definition for Y given by (5.1) into last expression, it is obtained that:

$$
\begin{aligned}
S_1^{*}(\chi) = \quad & (nF_{\chi}(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)(\boldsymbol{X}_l^{T}\boldsymbol{\beta} + m(\boldsymbol{\chi}_l) + \\
& \varepsilon_l - \boldsymbol{X}_l^{T}\widehat{\boldsymbol{\beta}}_b) + V_i(\boldsymbol{X}_i^{T}\boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \varepsilon_i - \boldsymbol{X}_i^{T}\widehat{\boldsymbol{\beta}}_b - \\
& \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)(\boldsymbol{X}_l^{T}\boldsymbol{\beta} + m(\boldsymbol{\chi}_l) + \varepsilon_l - \boldsymbol{X}_l^{T}\widehat{\boldsymbol{\beta}}_b)] - \\
& - \sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i,\chi)(\boldsymbol{X}_i^{T}\boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \varepsilon_i - \boldsymbol{X}_i^{T}\widehat{\boldsymbol{\beta}}_b)).
\end{aligned}
$$

Again, one can divide the expression grouping the terms for $m(\chi)+\varepsilon$ (from the "auxiliary nonparametric model") and the terms for the linear part ($\boldsymbol{\beta}$):

$$\begin{aligned} S_1^*(\chi) = \quad & (nF_\chi(h))^{1/2}(\sum_{i=1}^n w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^n w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l)+\varepsilon_l)+ \\ & +V_i(m(\boldsymbol{\chi}_i)+\varepsilon_i-\sum_{l=1}^n w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)(m(\boldsymbol{\chi}_l)+\varepsilon_l)]- \\ & \sum_{i=1}^n w_b(\boldsymbol{\chi}_i,\chi)(m(\boldsymbol{\chi}_i)+\varepsilon_i))+ \\ & +(nF_\chi(h))^{1/2}(\sum_{i=1}^n w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^n w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)+ \\ & +V_i(\boldsymbol{X}_i^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)-\sum_{l=1}^n w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b))]- \\ & \sum_{i=1}^n w_b(\boldsymbol{\chi}_i,\chi)\boldsymbol{X}_i^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)) \end{aligned}$$

which follows to

$$\begin{aligned} S_1^*(\chi) = \quad & (nF_\chi(h))^{1/2}(\widehat{m}_{hb}^{*NP}(\chi)-\widehat{m}_b^{NP}(\chi))+ \\ & +(nF_\chi(h))^{1/2}(\sum_{i=1}^n w_h(\boldsymbol{\chi}_i,\chi)[\sum_{l=1}^n w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)+ \\ & +V_i(\boldsymbol{X}_i^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b)-\sum_{l=1}^n w_b(\boldsymbol{\chi}_l,\boldsymbol{\chi}_i)\boldsymbol{X}_l^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b))]- \\ & \sum_{i=1}^n w_b(\boldsymbol{\chi}_i,\chi)\boldsymbol{X}_i^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_b))= \\ = \quad & S_{1,1}^*(\chi)+S_{1,2}^*(\chi), \end{aligned} \tag{5.95}$$

where

$$S_{1,1}^*(\chi) = \quad (nF_\chi(h))^{1/2}(\widehat{m}_{hb}^{*NP}(\chi)-\widehat{m}_b^{NP}(\chi)). \tag{5.96}$$

Considering now the second term in (5.95), which is

$$S^*_{1,2}(\chi) = (nF_\chi(h))^{1/2}(\sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)[\sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)\boldsymbol{X}_l^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b) +$$
$$+V_i(\boldsymbol{X}_i^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b) - \sum_{l=1}^{n} w_b(\boldsymbol{\chi}_l, \boldsymbol{\chi}_i)\boldsymbol{X}_l^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b))] -$$
$$\sum_{i=1}^{n} w_b(\boldsymbol{\chi}_i, \chi)\boldsymbol{X}_i^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_b)) \tag{5.97}$$

and taking into account (5.80), (4.35), (5.15) and (A.2) in Theorem 25 in the Appendix A, one obtains

$$S^*_{1,2}(\chi) = O((F(h)\log\log n)^{1/2}) \text{ a.s.}$$

Then, from the fact that $F(h)\log\log n \to 0$ (see (4.19)) it is true that

$$S^*_{1,2}(\chi) = o(1) \text{ a.s.}, \tag{5.98}$$

which is also equivalent to (5.91) in Theorem 14.

Finally, it remains to study the term $S^*_2(\chi)$. From Theorem 11 and Theorem 25 in Appendix A one has

$$\widehat{\boldsymbol{\beta}}^*_b - \widehat{\boldsymbol{\beta}}_b = O_{P\mathcal{S}}(n^{-1/2}).$$

This fact allows to follow similar reasoning as those used in the study of $S_2(\chi)$ (explained in Theorem 14), giving

$$S^*_2(\chi) = o_{P\mathcal{S}}(1), \tag{5.99}$$

which is, again, analogous to (5.92) in Theorem 14.

From (5.94), (5.95), (5.96), (5.97) and (5.99), one gets that

$$(nF_\chi(h))^{1/2}(\widehat{m}^*_{hb}(\chi) - \widehat{m}_b(\chi)) = (nF_\chi(h))^{1/2}(\widehat{m}^{*NP}_{hb}(\chi) - \widehat{m}^{NP}_b(\chi)) + o_{P\mathcal{S}}(1). \tag{5.100}$$

Finally, (5.82), (5.100) and Theorem 5 give the result of Theorem 15. □

5.4 Simulacion study

This section presents a simulation study designed to analyse the behaviour of the bootstrap procedures applied to the construction of confidence intervals for the SFPL model. It contains a description of the procedure developed to build the confidence intervals for the SFPL model (5.1) and the results obtained with the simulated models, following the ones used previously, in Section 4.4, for the FNP regression model. Again, because of its generality, the wild bootstrap procedure will de considered.

5.4.1 Building the confidence intervals

Given $\{X, \chi\}$, and a model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{5.101}$$

where the sequence $\{(\boldsymbol{X}_i, \boldsymbol{\chi}_i, Y_i)\}$ is α-mixing and identically distributed as $(\boldsymbol{X}, \boldsymbol{\chi}, Y)$, and $\{X, \chi\}$ are observed from $\{\boldsymbol{X}, \boldsymbol{\chi}\}$, true and bootstrap $(1-\alpha)$-confidence intervals can be built for each part of the model: linear (β), nonparametric $(m(\chi))$ and the explanatory part together $(X^T\beta + m(\chi))$.

1. CI for the linear part, assuming for simplicity that $\beta \in \mathbb{R}$:
   - True CI: $I_{1-\alpha}^{\beta,true} = (\widehat{\beta}_h + q_{\alpha/2}^{\beta,true}, \widehat{\beta}_h + q_{1-\alpha/2}^{\beta,true})$
   - Bootstrap CI: $I_{1-\alpha}^{\beta,*} = (\widehat{\beta}_h + q_{\alpha/2}^{\beta,*}, \widehat{\beta}_h + q_{1-\alpha/2}^{\beta,*})$
2. CI for the nonparametric part:
   - True CI: $I_{\chi,1-\alpha}^{m,true} = (\widehat{m}_h(\chi) + q_{\alpha/2}^{m,true}(\chi), \widehat{m}_h(\chi) + q_{1-\alpha/2}^{m,true}(\chi))$
   - Bootstrap CI: $I_{\chi,1-\alpha}^{m,*} = (\widehat{m}_h(\chi) + q_{\alpha/2}^{m,*}(\chi), \widehat{m}_h(\chi) + q_{1-\alpha/2}^{m,*}(\chi))$
3. CI for the explanatory part:
   - True CI: $I_{X,\chi,1-\alpha}^{true} =$ $((X^T\widehat{\beta}_h + \widehat{m}_h(\chi)) + q_{\alpha/2}^{true}(X,\chi), (X^T\widehat{\beta}_h + \widehat{m}_h(\chi)) + q_{1-\alpha/2}^{true}(X,\chi))$
   - Bootstrap CI: $I_{X,\chi,1-\alpha}^{*} =$ $((X^T\widehat{\beta}_h + \widehat{m}_h(\chi)) + q_{\alpha/2}^{*}(X,\chi), (X^T\widehat{\beta}_h + \widehat{m}_h(\chi)) + q_{1-\alpha/2}^{*}(X,\chi))$

where the quantiles involved in the confidence intervals were computed in the following way:

Theoretical quantiles $(q_p^{\beta,true}, q_p^{m,true}(\chi), q_p^{true}(X,\chi))$

1. Generate n_{MC} samples $\{(\boldsymbol{X}_i^s, \boldsymbol{\chi}_i^s, Y_i^s), i = 1, \ldots, n\}_{s=1}^{n_{MC}}$ from Model (5.101).

2. Carry out n_{MC} estimates $\left\{\widehat{\beta}_h^s\right\}_{s=1}^{n_{MC}}$ and $\{\widehat{m}_h^s(\chi)\}_{s=1}^{n_{MC}}$, where $\widehat{\beta}_h^s$ is the estimator of β and $\widehat{m}_h^s(\chi)$ is the functional kernel estimator, derived from the s^{th} sample $\{(\boldsymbol{X}_i^s, \boldsymbol{\chi}_i^s, Y_i^s)\}_{i=1}^{n}$.

3. Compute the approximation errors for each part of the model:

   (a) For the linear part
   $ERROR.MC.\beta_s = \left\{\beta - \widehat{\beta}_h^s\right\}_{s=1}^{n_{MC}}$.

   (b) For the nonparametric part
   $ERROR.MC.m_s = \{m(\chi) - \widehat{m}_h^s(\chi)\}_{s=1}^{n_{MC}}$.

   (c) For all the explanatory part
   $ERROR.MC_s = \left\{X^T(\beta - \widehat{\beta}_h^s) + (m(\chi) - \widehat{m}_h^s(\chi))\right\}_{s=1}^{n_{MC}}$.

4. Compute the theoretical quantiles, which will be used to build the confidence intervals for each part:

   (a) For the linear part: $q_p^{\beta,true}$ from the quantile of order p of $ERROR.MC.\beta_s$.

   (b) For the nonparametric part: $q_p^{m,true}(\chi)$ from the quantile of order p of $ERROR.MC.m_s$.

   (c) For all the explanatory part: $q_p^{true}(X,\chi)$ from the quantile of order p of $ERROR.MC_s$.

Bootstrap quantiles $(q_p^{\beta,*}, q_p^{m,*}(\chi), q_p^{*}(X,\chi))$

1. Generate the sample $\mathcal{S} = \{(\boldsymbol{X}_1, \boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{\chi}_n, Y_n)\}$ from Model (5.101).

2. Compute $\widehat{\beta}_b$ and $\widehat{m}_b(\chi)$ over the dataset $\mathcal{S}$.

3. Repeat B times the bootstrap algorithm over $\mathcal{S}$ by using i.i.d. random variables V_i drawn from the two Dirac distributions

$$0.1(5+\sqrt{5})\delta_{(1-\sqrt{5})/2} + 0.1(5-\sqrt{5})\delta_{(1+\sqrt{5})/2},$$

   giving the B bootstrap estimates $\left\{\widehat{\beta}_b^{*,r}\right\}_{r=1}^{B}$ and $\{\widehat{m}_{hb}^{*,r}(\chi)\}_{r=1}^{B}$.

4. Compute the set of bootstrap errors for each part of the model:

   (a) For the linear part
   $ERROR.BOOT.\beta_r = \left\{\widehat{\beta}_b - \widehat{\beta}_b^{*,r}\right\}_{r=1}^{B}$.

   (b) For the nonparametric part
   $ERROR.BOOT.m_r \left\{\widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi)\right\}_{r=1}^{B}$.

   (c) For the explanatory part
   $ERROR.BOOT_r = \left\{X^T(\widehat{\beta}_b - \widehat{\beta}_b^{*,r}) + (\widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi))\right\}_{r=1}^{B}$.

5. Compute the bootstrap quantiles for each part of the model, using the bootstrap errors:

   (a) For the linear part: $q_p^{\beta,*}$ from the quantile of order p of $ERROR.BOOT.\beta_r$.

   (b) For the nonparametric part: $q_p^{m,*}(\chi)$ from the quantile of order p of $ERROR.BOOT.m_r$.

   (c) For all the explanatory part: $q_p^*(X,\chi)$ from the quantile of order p of $ERROR.BOOT_r$.

Finally, the estimates $\widehat{\beta}_h$ and $\widehat{m}_h(\chi)$ needed for the intervals were obtained from $\mathcal{S}$.

The quadratic kernel, $K(u) = 1.5(1-u^2)1_{[0,1]}(u)$, was considered in the estimates $\widehat{m}_h$ and $\widehat{m}_{hb}^*$, while the bandwidth $b = b_{CV}$ was selected by means of the cross-validation methodology. Then, $h = b_{CV}$ was set.

5.4.2 Model 1: smooth curves

Simulated data will be obtained from Model 1 in Section 4.4, adding a scalar covariate to deal with the SFPL model. The model is compound by:

$$\boldsymbol{\chi}_i(t_j) = \cos(a_i + \pi(2t_j - 1)),$$

where $\{a_i\}$ comes from AR(1) gaussian processes with correlation coefficient $\rho_a = 0.7$ and variance $\sigma_a^2 = 0.05$, while $0 = t_1 < t_2 < \cdots < t_{99} < t_{100} = 1$ are 100 equally spaced measurements. The regression operator was

$$m(\boldsymbol{\chi}) = \frac{1}{2\pi}\int_{1/2}^{3/4} (\boldsymbol{\chi}'(t))^2 dt.$$

The scalar explanatory variables for the linear part of the model can be chosen as $\{a_i^2\}$, if one considers dependent covariates, or an $AR(1)$ gaussian process independent of $\{a_i\}$ (with correlation coefficient $\rho = 0.8$ and variance $\sigma^2 = 0.5$), if one considers independent covariates. The unknown parameter was $\beta = 1$.

The errors $\{\varepsilon_i\}$ were independent centred gaussians of variance equal to 0.1 times the empirical variance of $\{m(\boldsymbol{\chi}_1), \ldots, m(\boldsymbol{\chi}_n)\}$.

True and bootstrap $(1-\alpha)$-confidence intervals for $\beta, m(\chi)$ and $X^T\beta + m(\chi)$, with $(X, \chi) \in \mathcal{D}$, were computed and compared. Test-sample $\mathcal{D} = \{(X_i, \chi_i), i = 1, \ldots n_{\mathcal{D}},\}$ is generated following the generation process explained in Section 4.4 to obtain independent curves (χ_i), and following the same procedure to generate the values for the corresponding scalar covariate (X_i).

Note that, from the procedure explained above, one obtains one $(1-\alpha)$-confidence interval of each type for β, $m(\chi)$ and also for $X^T\beta + m(\chi)$: true ($I_{1-\alpha}^{\beta,true}$, $I_{\chi,1-\alpha}^{m,true}$ and $I_{X,\chi,1-\alpha}^{true}$) and bootstrap ($I_{1-\alpha}^{\beta,*}$, $I_{\chi,1-\alpha}^{m,*}$ and $I_{X,\chi,1-\alpha}^{*}$) confidence intervals, respectively. To compare the accuracy of each type of interval, the empirical coverages are obtained by repeating the procedure M times and computing the proportion of times that each interval contains the value β, $m(\chi)$ and $X^T\beta + m(\chi)$, respectively, that is, to check if:

- $I_{1-\alpha}^{\beta,true}$ and $I_{1-\alpha}^{\beta,*}$ contains β.
- $I_{\chi,1-\alpha}^{m,true}$ and $I_{\chi,1-\alpha}^{m,*}$ contains $m(\chi)$, for each χ in $\mathcal{D}$.
- $I_{X,\chi,1-\alpha}^{true}$ and $I_{X,\chi,1-\alpha}^{*}$ contains $X\beta + m(\chi)$, for each (X, χ) in $\mathcal{D}$.

Values $n_{MC} = 2000$, $B = 500$, $n_{\mathcal{D}} = 100$, $M = 500$, $1-\alpha = 0.95, 0.90$ and $n = 200, 400$ were considered.

Table 5.1 reports the empirical coverage of the confidence intervals obtained for Model 1 considering independent covariates, meanwhile Table 5.2 reports the analogous case when the dependent covariates is included in the model. The coverage when dealing with confidence intervals for $m(\chi)$ and $X^T\beta + m(\chi)$ is the average over $\mathcal{D}$ of the empirical coverage of the computed confidence intervals.

Table 5.1: Empirical coverage of the true and bootstrap confidence intervals for Model 1 with independent covariates, for each part of the SFPL model. For $m(\chi)$ and $X^T\beta + m(\chi)$, the average over $\mathcal{D}$ of the empirical coverages is shown, with the standard deviation in brackets.

n=200			
$1-\alpha$	0.95		
	β	$m(\chi)$	$X^T\beta + m(\chi)$
Coverage (I^{true})	0.946	0.950 (0.015)	0.951 (0.013)
Coverage (I^*)	0.926	0.907 (0.033)	0.911 (0.032)
$1-\alpha$	0.90		
Coverage (I^{true})	0.910	0.899 (0.021)	0.900 (0.020)
Coverage (I^*)	0.872	0.856 (0.037)	0.858 (0.035)
n=400			
$1-\alpha$	0.95		
	β	$m(\chi)$	$X^T\beta + m(\chi)$
Coverage (I^{true})	0.952	0.948 (0.011)	0.949 (0.010)
Coverage (I^*)	0.946	0.909 (0.085)	0.910 (0.082)
$1-\alpha$	0.90		
Coverage (I^{true})	0.902	0.896 (0.016)	0.898 (0.017)
Coverage (I^*)	0.886	0.855 (0.091)	0.858 (0.090)

Tables 5.1 and 5.2 show promising results related to the coverage of the confidence intervals. In both cases, looking at the true confidence intervals, the coverage is almost the confidence level, being also very similar for each part of the model. Regarding the bootstrap confidence intervals, the coverages are a little lower than the confidence level when dealing with the linear part of the model and they are slightly lower in the case of the nonparametric and all the explanatory part together.

One can see in Figure 5.1 the empirical coverages obtained for the true and bootstrap confidence intervals for $X^T\beta + m(\chi)$, considering independent covariates. One can see how the coverage of the true confidence intervals is always around the confidence level, while the corresponding of the bootstrap confidence intervals is lower than the confidence level in general.

Table 5.2: Empirical coverage of the true and bootstrap confidence intervals for Model 1 with dependent covariates, for each part of the SFPL model. For $m(\chi)$ and $X^T\beta + m(\chi)$, the average over $\mathcal{D}$ of the empirical coverages is shown, with the standard deviation in brackets.

n=200			
$1-\alpha$	0.95		
	β	$m(\chi)$	$X^T\beta + m(\chi)$
Coverage (I^{true})	0.960	0.948 (0.008)	0.945 (0.009)
Coverage (I^*)	0.832	0.833 (0.095)	0.826 (0.108)
$1-\alpha$	0.90		
Coverage (I^{true})	0.902	0.896 (0.013)	0.891 (0.013)
Coverage (I^*)	0.778	0.769 (0.102)	0.757 (0.122)
n=400			
$1-\alpha$	0.95		
	β	$m(\chi)$	$X^T\beta + m(\chi)$
Coverage (I^{true})	0.956	0.948 (0.008)	0.950 (0.007)
Coverage (I^*)	0.860	0.861 (0.093)	0.841 (0.135)
$1-\alpha$	0.90		
Coverage (I^{true})	0.922	0.900 (0.009)	0.897 (0.012)
Coverage (I^*)	0.810	0.798 (0.103)	0.774 (0.144)

Figure 5.2 displays the true (points) and bootstrap (lines) confidence intervals obtained for β, $m(\chi)$ and $X^T\beta + m(\chi)$, respectively.

Regarding the differences between both kinds of scalar covariate, one can see that the coverages are better for the case of independent covariates than for the dependent one. Thus, it seems that the latter case (Table 5.2) is more difficult to estimate, but one can see an improvement in the coverages as the sample size increases.

If one compares the results with the corresponding for the FNP model (Section 4.4), one can observe a slight decrease in the coverages, probably due to the higher difficulty on the estimation for the SFPL model. However, the bootstrap procedure seems to have a nice behaviour within this example.

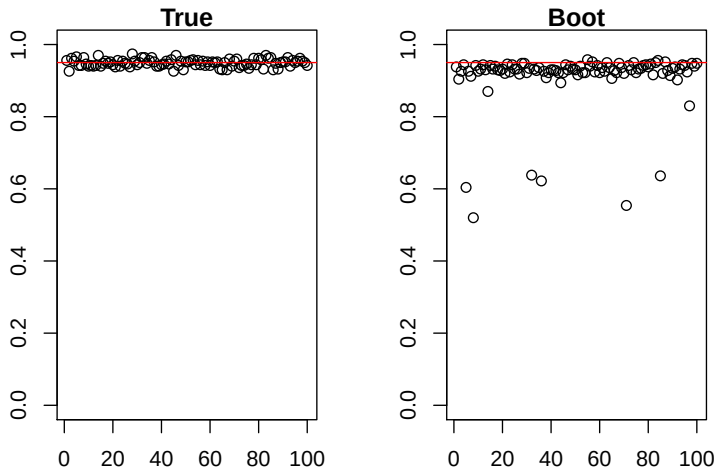


Figure 5.1: Empirical coverage for the true and bootstrap confidence intervals (for the complete explanatory part $X^T\beta + m(\chi)$) for Model 1 for each (X, χ) in $\mathcal{D}$ (considering independent covariates, $1 - \alpha = 0.95$ and $n = 400$). Solid line located at a height $1 - \alpha$.

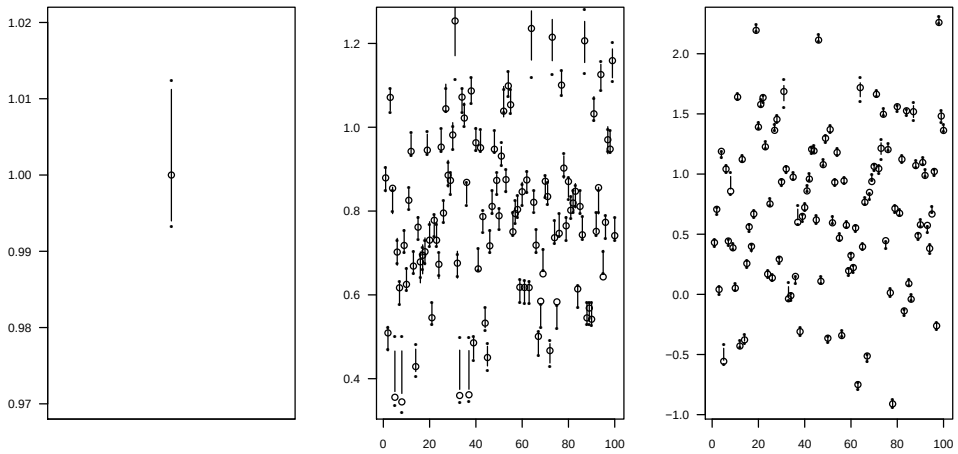


Figure 5.2: From left to right: the lines represent the bootstrap confidence intervals for β and for each $m(\chi)$ and $X^T\beta + m(\chi)$, respectively. The dots delimit the corresponding true confidence interval. Model 1 is considered, with independent covariates, $1 - \alpha = 0.95$ and $n = 400$.

5.4.3 Model 2: rough curves

Simulated data will be obtained based on a modified version of Model 2 in Section 4.4, adding a scalar covariate to deal with the SFPL model. The model is compound by:

$$\boldsymbol{\chi}_i(t_j) = b_{2i}\cos(b_{1i}t_j) + \sum_{k=1}^{j} B_{ik}/b,$$

where $b = 10$, $\{b_{1i}\}$ and $\{b_{2i}\}$ come from MA(1) and AR(1) gaussian processes with parameters $\theta_{b_1} = -0.5$ and $\rho_{b_2} = 0.9$, respectively, and variances $\sigma^2_{b_1} = 0.1$ and $\sigma^2_{b_2} = 0.01$, B_{ik} are i.i.d. realizations of $N(0, \sigma)$ with $\sigma = 0.1$ and $0 = t_1 < t_2 < \cdots < t_{99} < t_{100} = \pi$ are 100 equally spaced measurements. The regression operator was

$$m(\boldsymbol{\chi}) = \int_0^{\pi} (\boldsymbol{\chi}(t))^2 dt.$$

The scalar explanatory variables for the linear part of the model can be chosen as $\{b_{2i}\}$, if one considers dependent covariates, or an $AR(1)$ gaussian process independent of $\{b_{2i}\}$ (with correlation coefficient $\rho = 0.8$ and variance $\sigma^2 = 0.5$), if one considers independent covariates.

The errors $\{\varepsilon_i\}$ were, again, independent centred gaussians of variance equal to 0.1 times the empirical variance of $\{m(\boldsymbol{\chi}_1), \ldots, m(\boldsymbol{\chi}_n)\}$.

The simulations are carried out following the same procedure as in Model 1, using also a test sample and choosing the same parameters

Following the same procedure and tuning parameters as for Model 1, Table 5.3 reports the empirical coverage of the confidence intervals obtained for Model 2 considering independent covariates, meanwhile Table 5.4 reports the analogous case when dependent covariates are included in the model.

Analyzing the results shown in Table 5.3 and 5.4 for the true confidence intervals, one can see that the coverage is almost the confidence level. Regarding the bootstrap confidence intervals, the coverages are now lower than the confidence level, this difference being more remarkable when dealing with the nonparametric and all the explanatory part together. It is important to notice that the coverages improve as the sample size increases in both cases (with independent and dependent scalar covariate) and that this improvement is more outstanding than in Model 1.

Table 5.3: Empirical coverage of the true and bootstrap confidence intervals for Model 2 with independent covariates, for each part of the SFPL model. For $m(\chi)$ and $X^T\beta + m(\chi)$, the average over $\mathcal{D}$ of the empirical coverages is shown, with the standard deviation in brackets.

n=200			
$1-\alpha$	0.95		
	β	$m(\chi)$	$X^T\beta + m(\chi)$
Coverage (I^{true})	0.952	0.949 (0.010)	0.951 (0.010)
Coverage (I^*)	0.850	0.706 (0.063)	0.719 (0.066)
$1-\alpha$	0.90		
Coverage (I^{true})	0.902	0.900 (0.012)	0.902 (0.014)
Coverage (I^*)	0.792	0.663 (0.062)	0.671 (0.063)
n=400			
$1-\alpha$	0.95		
	β	$m(\chi)$	$X^T\beta + m(\chi)$
Coverage (I^{true})	0.962	0.952 (0.011)	0.952 (0.011)
Coverage (I^*)	0.884	0.748 (0.062)	0.756 (0.061)
$1-\alpha$	0.90		
Coverage (I^{true})	0.908	0.902 (0.014)	0.901 (0.015)
Coverage (I^*)	0.784	0.701 (0.060)	0.706 (0.059)

As in Model 1, one can see in Figure 5.3 the empirical coverages obtained for the true and bootstrap confidence intervals for $X^T\beta + m(\chi)$, considering independent covariates. The coverage of the true confidence intervals is again around the confidence level. However, the corresponding of the bootstrap confidence intervals is much lower than the confidence level in general, according to the numerical results shown in Table 5.3. Figure 5.4 displays the true (points) and bootstrap (lines) confidence intervals obtained for β, $m(\chi)$ and $X^T\beta + m(\chi)$, respectively. In this case, there are not many differences between the confidence intervals obtained for $m(\chi)$ and for $X^T\beta + m(\chi)$, which agrees with the similar empirical coverages obtained in both cases.

If one compares again the results with the corresponding for the FNP model (Section 4.4), one can observe a higher descent in the coverages. However, taking into account that Model 2 corresponds to rough curves, the bootstrap procedure seems to behave properly within this example.

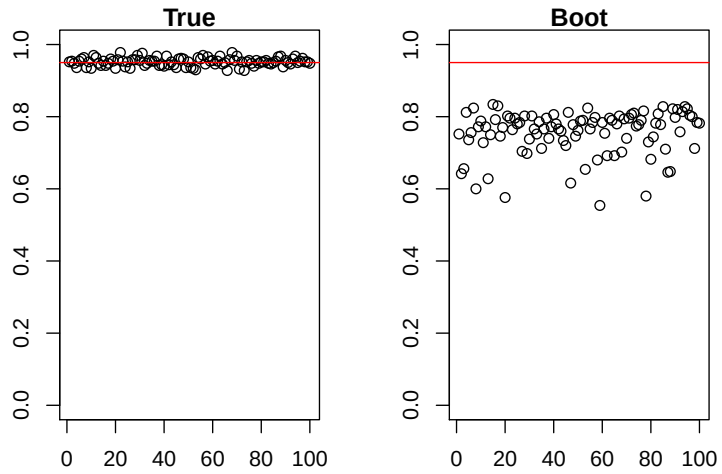


Figure 5.3: Empirical coverage for the true and bootstrap confidence intervals (for the complete explanatory part $X^T\beta + m(\chi)$) for Model 2 for each (X, χ) in $\mathcal{D}$ (considering independent covariates, $1 - \alpha = 0.95$ and $n = 400$). Solid line located at a height $1 - \alpha$.

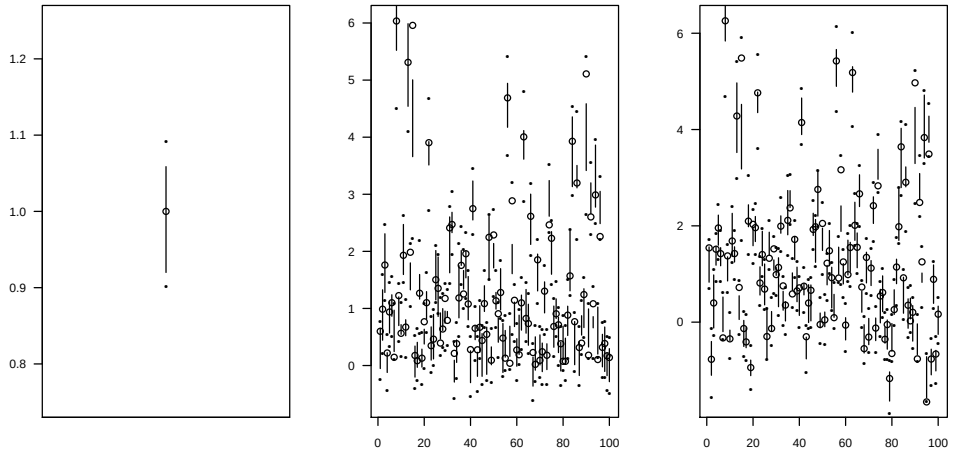


Figure 5.4: From left to right: the lines represent the bootstrap confidence intervals for β and for each $m(\chi)$ and $X^T\beta + m(\chi)$, respectively. The dots delimit the corresponding true confidence interval. Model 2 is considered, with independent covariates, $1 - \alpha = 0.95$ and $n = 400$.

Table 5.4: Empirical coverage of the true and bootstrap confidence intervals for Model 2 with dependent covariates, for each part of the SFPL model. For $m(\chi)$ and $X^T\beta + m(\chi)$, the average over $\mathcal{D}$ of the empirical coverages is shown, with the standard deviation in brackets.

n=200			
$1-\alpha$	0.95		
	β	$m(\chi)$	$X^T\beta + m(\chi)$
Coverage (I^{true})	0.960	0.951 (0.010)	0.951 (0.011)
Coverage (I^*)	0.816	0.736 (0.058)	0.702 (0.076)
$1-\alpha$	0.90		
Coverage (I^{true})	0.912	0.900 (0.014)	0.900 (0.014)
Coverage (I^*)	0.742	0.679 (0.056)	0.657 (0.072)
n=400			
$1-\alpha$	0.95		
	β	$m(\chi)$	$X^T\beta + m(\chi)$
Coverage (I^{true})	0.940	0.945 (0.012)	0.947 (0.011)
Coverage (I^*)	0.864	0.795 (0.055)	0.777 (0.070)
$1-\alpha$	0.90		
Coverage (I^{true})	0.888	0.895 (0.015)	0.897 (0.015)
Coverage (I^*)	0.802	0.738 (0.058)	0.728 (0.068)

5.5 Application to electricity data

This section applies the methodology proposed in this chapter to the construction of confidence intervals for the mean hourly electricity demand/price in Spain given the daily curve of electricity demand/price in the previous day (as functional covariate) and considering also scalar covariates with linear effect over the response: temperature (to be applied in the demand case) and both cumulative demand and mean wind power production (in the case of the price).

The dataset is restricted in this application to the workdays of the second quarter of the year 2012, following the same outline as Section 4.5 for the FNP model. As in the simulation study presented in Section 5.4, the wild bootstrap procedure will be considered, using 1000 bootstrap replicates.

5.5.1 Case study: electricity demand

The first application within the electricity demand case will consists on predict the 24 hours for a fixed day, corresponding to the last day in the database (Friday, June 29, 2012) given the daily curve of electricity demand in the previous day and also the two temperature derived variables (see Section 1.5.1 for details), which will compound the linear part of the SFPL model.

Specifically, the considered SFPL model is:

$$\boldsymbol{\chi}_{i+1}(t) = \boldsymbol{X}_i^T \boldsymbol{\beta}^{(1)} + m_t^{(1)}(\boldsymbol{\chi}_i) + \varepsilon_{i,t}^{(1)} \ (t = 1, \ldots, 24, \ i = 1, \ldots, n),$$

where the temperature covariates are $\boldsymbol{X}_i = (X_{i1}, X_{i2})^T = (HDD_i, CDD_i)^T$ (for the definition of HDD and CDD see (1.12) and (1.13), respectively).

Table 5.5 shows a brief comparison of the confidence intervals obtained for the FNP and the SFPL model (in which all the explanatory part of the model is computed together), through a comparison of the bootstrap confidence intervals lengths. Results show that the length of the confidence intervals for both functional regression models is similar.

Figure 5.5, left panel, shows the bootstrap confidence intervals obtained for the demand within this SFPL model.

Table 5.5: Comparison between the Confidence Interval length in FNP and SFPL model for electricity demand, predicting June 29, 2012.

Model	length: mean (sd)
FNP	1045.92 (353.44)
SFPL	1081.54 (276.23)

A second application is performed, considering the case in which one predicts a fixed hour, 20:00, for the last 21 days of the dataset. In this case, the SFPL model is

$$\boldsymbol{\chi}_{i+1,d}(20) = \boldsymbol{X}_i^T \boldsymbol{\beta}^{(2)} + m_d^{(2)}(\boldsymbol{\chi}_{i,d}) + \varepsilon_{i,d}^{(2)} \ (d = 1, \ldots, 21, \ i = 1, \ldots, n),$$

where the covariates are again $\boldsymbol{X}_i = (X_{i1}, X_{i2})^T = (HDD_i, CDD_i)^T$

The comparison between the length of the confidence intervals, in Table 5.6, shows that the confidence intervals for the SFPL model are shorter than the ones for the FNP model. Also, the confidence intervals are now longer than in the previous case, when one considers the 24 hours of the same day.

Figure 5.5, right panel, represents the obtained confidence intervals for the mean hourly demand in the weekdays in June at the fixed hour 20:00. One can see how the demand remains stable along the month for the same hour of the day and also that, as mentioned above, now the intervals are larger.

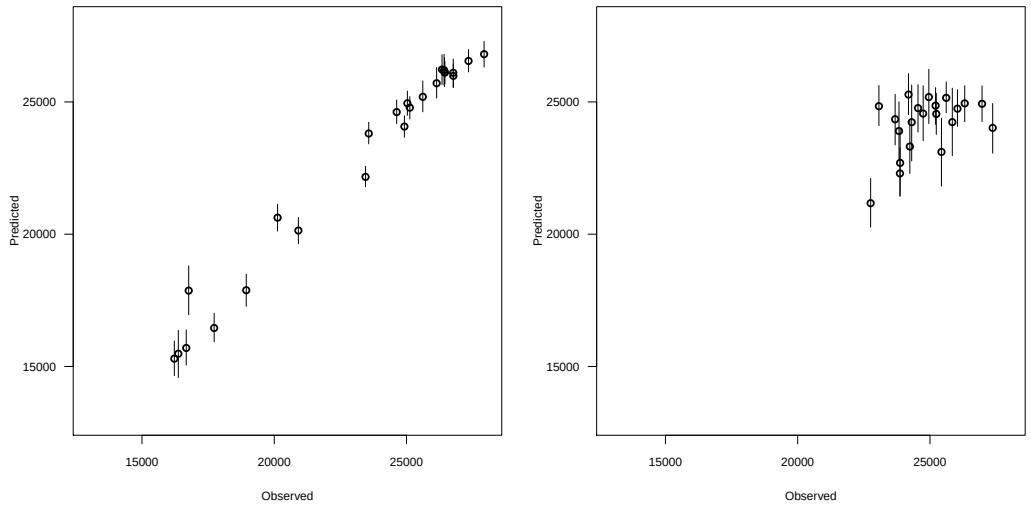


Figure 5.5: Left panel: Bootstrap confidence intervals computed for the electricity demand, for the 24 hours of Friday, June 29, 2012. Right panel: Bootstrap confidence intervals computed for the electricity demand, for the workdays in June, 2012 (fixed hour: 20:00).

Table 5.6: Comparison between the Confidence Interval length in FNP and SFPL model for electricity demand, predicting a fixed hour, 20:00, along the weekdays in June, 2012.

Model	length: mean (sd)
FNP	2107.47 (813.80)
SFPL	1873.15 (468.12)

5.5.2 Case study: electricity price

Application for the electricity price will be also divided into two parts. First one consists in predicting one day (24 hours in June 29, 2012), including the cumulative daily demand (D) and the daily mean wind power production (WPP). The considered SFPL model is

$$\boldsymbol{\chi}_{i+1}(t) = \boldsymbol{X}_i^T \boldsymbol{\beta}^{(1)} + m_t^{(1)}(\boldsymbol{\chi}_i) + \varepsilon_{i,t}^{(1)} \ (t = 1, \ldots, 24, \ i = 1, \ldots, n),$$

where the covariates are $\boldsymbol{X}_i = (X_{i1}, X_{i2})^T = (D_i, WPP_i)^T$

Also, a comparison on the confidence intervals lengths from the FNP and the SFPL model is given in Table 5.7. One can see that including only the demand as scalar covariate in the SFPL model produces similar intervals as the FNP, while adding also the wind power production reduces the length of the confidence intervals. Left panel in Figure 5.6 shows the bootstrap confidence intervals obtained for the mean hourly electricity price within this SFPL model, for the 24 hours in June 29, 2012.

Table 5.7: Comparison between the Confidence Interval length in FNP and SFPL model for electricity price, predicting June 29, 2012.

Model	length: mean (sd)
FNP	7.44 (1.63)
SFPL (D)	7.07 (1.56)
SFPL (D+WPP)	5.53 (1.14)

A second application is performed, predicting one fixed hour, 20:00, for the last 21 days of the dataset, which are the weekdays in June, 2012. In this case, the SFPL model is

$$\boldsymbol{\chi}_{i+1,d}(20) = \boldsymbol{X}_i^T \boldsymbol{\beta}^{(2)} + m_d^{(2)}(\boldsymbol{\chi}_{i,d}) + \varepsilon_{i,d}^{(2)} \ (d = 1, \ldots, 21, \ i = 1, \ldots, n),$$

where the covariates are again $\boldsymbol{X}_i = (X_{i1}, X_{i2})^T = (D_i, WPP_i)^T$

Comparison of lengths in this case (see Table 5.8) shows that SFPL model produces longer confidence intervals than the FNP model if one considers only demand as scalar covariate. However, the length of the confidence intervals decreases when both demand and wind power production are included, begin in that case shorter than the ones for the FNP model. In any case, is

important to include both scalar covariates for the sake of the improvement obtained in the predictions seen in Chapter 3. Right panel in Figure 5.6 shows again the bootstrap confidence intervals obtained for the price within this SFPL model. The confidence intervals represented in this Figure are now similar in terms of length.

Table 5.8: Comparison between the Confidence Interval length in FNP and SFPL model for electricity price, predicting a fixed hour, 20:00, along the weekdays in June, 2012.

Model	length: mean (sd)
FNP	8.21 (3.15)
SFPL (D)	9.68 (4.14)
SFPL (D+WPP)	6.34 (2.30)

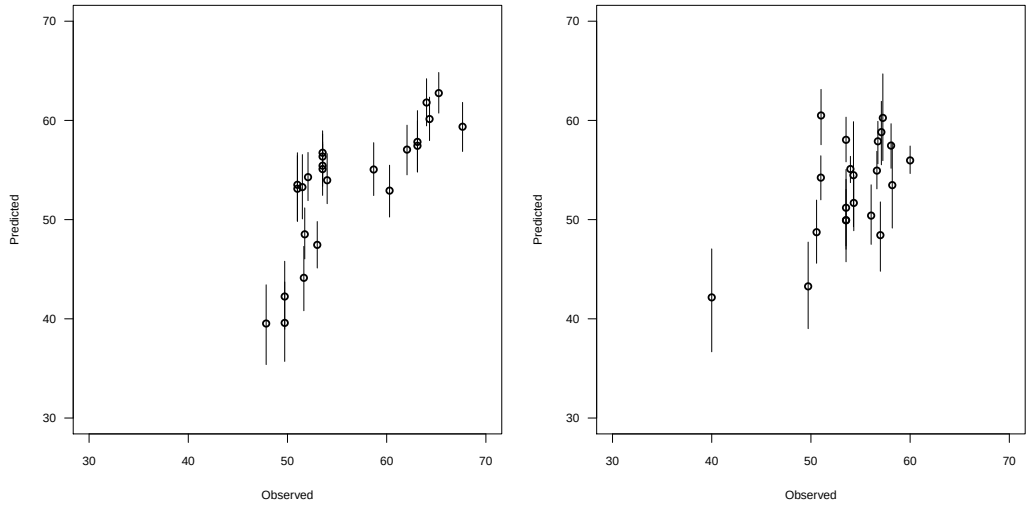


Figure 5.6: Left panel: Bootstrap confidence intervals computed for the electricity price, for the 24 hours of Friday, June 29, 2012. Right panel: Bootstrap confidence intervals computed for the electricity price, for the workdays in June, 2012 (fixed hour: 20:00).

5.6 Conclusions

This chapter proposes two bootstrap procedures to construct pointwise confidence intervals for the SFPL regression model, considering scalar response and functional predictor and in which one adds linear effect of scalar covariates. By means of these two bootstrap procedures one can approximate the asymptotic distribution of the estimators in both components (linear and nonparametric) of the regression model. The validity of these two procedures has been proved theoretically in the setting of dependent data, assuming α–mixing conditions on the sample, and can be also apply to the setting of independent data as a particular case.

There is no preceding in the literature about bootstrap procedures in this kind of SFPL regression models. It is even difficult to find applications of this kind of bootstrap procedures in scalar partial linear regression. One can find in Liang et al. (2000) and You and Chen (2006) proposals for bootstrap approximation in partial linear regression but in the case of fixed design, independent data and regarding the linear component of the model.

A simulation study was carried out to show the performance of the proposed procedures, in addition to an application to a real dataset. Specifically, applications to electrical data from the Spanish Electricity Market illustrate its usefulness in practice.

Chapter 6

Prediction intervals in functional regression

6.1 Introduction

Chapters 4 and 5 proposed bootstrap procedures to build confidence intervals in the context of functional regression under dependence conditions. Specifically, for the FNP model and the SFPL model. Validity of those bootstrap procedures has been proved for both functional regression models, allowing to approximate the true asymptotic distribution of the regression estimator by its bootstrapped version.

When dealing with forecasting, it is important to consider also prediction intervals. Since the confidence intervals are designed to deal with the regression function (or its unknown components), they cannot be applied to the response of the model which is of main interest in the predictions. It is then necessary to include the variability from the error of the model (ε) and not only the variability of the regression estimator, as it is made in confidence intervals.

It is of main importance to distinguish between confidence intervals and prediction intervals. Consider, for instance, the FNP model (see 4.2 in Chapter 4):

$$Y = m(\boldsymbol{\chi}) + \varepsilon,$$

where $\boldsymbol{\chi}$ is the functional predictor, Y is the scalar response and ε is the error of the model. $m(\cdot)$ is the regression operator which is, for a fixed χ, the expectation of the response conditionally on this fixed predictor, that is:

$$\mathbb{E}(Y|\boldsymbol{\chi} = \chi) = m(\chi).$$

Then, one can obtain a confidence interval (level $1-\alpha$) for this conditional expectation based on its estimator $\widehat{m}(\chi)$. This interval will be built as:

$$\left(\widehat{m}(\chi) + q_{\alpha/2}(\chi), \widehat{m}(\chi) + q_{1-\alpha/2}(\chi)\right),$$

where $q_{\alpha/2}(\chi)$ and $q_{1-\alpha/2}(\chi)$ are the quantiles from the distribution of $m(\chi) - \widehat{m}(\chi)$.

This confidence interval is devoted to cover the true value of the regression operator, $m(\chi)$, and it deals only with the variability due to its estimation.

The procedure to obtain a prediction interval is similar, but changing its "philosophy". Prediction intervals are devoted to cover the response and not the regression operator and so, they include, not only the variability due to the estimation, but also the error of the model. That is, instead of working with the expectation of Y conditionally on χ, one deals with the distribution of Y conditionally on χ.

The procedure to build a prediction interval follows the same idea as the confidence interval, but now this variability due to the error of the model has to be added. The starting point is that now, one looks for an interval (a, b) such that

$$P(Y|\chi \in (a, b)) = 1 - \alpha.$$

In the theory, assuming $\varepsilon/\chi \sim N(0, \sigma_\varepsilon^2)$, as $Y/\chi = m(\chi) + \varepsilon/\chi$ then $Y/\chi \sim N(m(\chi), \sigma_{\varepsilon|\chi}^2)$ and one can obtain the desired interval from this normal distribution.

In practice, some of the terms involved in the considered distributions are unknown (for instance $m(\chi)$ and $\sigma_{\varepsilon|\chi}^2$), and this is the reason why one needs to approximate them applying the bootstrap procedures developed in chapters 4 and 5, together with the flexibility of the bootstrap procedures in which normality is not assumed.

In the context in which this memory is stated, electricity data, it is also important to join the point forecasts with prediction intervals, generalizing the results obtained in this real application. As Weron (2006) pointed out, there is a big variety of studies devoted to evaluate point forecasts in electricity markets, meanwhile it is difficult to find references dealing with prediction intervals. Misiorek et al. (2006) were the first in consider prediction intervals within this context, applying them to electricity prices using time series

models.

Construction of prediction intervals allows to connect with Chapter 3 to complement the pointwise prediction results obtained for the electricity demand and price with those prediction intervals and to extend the applications in chapters 4 and 5 from confidence to prediction intervals.

The algorithms developed to build those prediction intervals are adapted to deal with both homoscedasticity and heteroscedasticity, considering each one of the functional regression models (FNP and SFPL regression models).

Pointwise forecasts are not always enough when one wants to obtain predictions for data including disturbances. In that sense, as mentioned above, prediction intervals are a nice tool to study how the forecasts may fluctuate. However, it could be also very useful to obtain the prediction density, which helps to understand the behaviour of the forecasts in a deeper way. Estimation of the prediction density is less common that the estimation of prediction intervals and it will be considered also along this chapter. Specifically, the bootstrap procedures proposed in this memory allow to estimate that prediction density for electricity demand and price.

The rest of the chapter is organized as follows: Section 6.2 contains the procedure developed to build the prediction intervals. Section 6.3 includes an application to the electricity data, showing the accuracy of the obtained prediction intervals. Finally, Section 6.4 gives some conclusions to this chapter.

6.2 Building prediction intervals

The procedure to build prediction intervals will take use of the bootstrap procedures developed in chapters 4 and 5 for the FNP and SFPL model, respectively.

6.2.1 Prediction intervals for Functional Nonparametric model

First, the FNP case is addressed. As indicated in the introduction to this chapter, the prediction intervals focus on the distribution of Y conditionally

on a fixed χ. Thus, in practice one disposes of a sample $\mathcal{S}$, where

$$\mathcal{S} = \{(\boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{\chi}_n, Y_n)\}.$$

Based on this sample for the FNP model introduced in (4.2), the predictor for Y/χ is $\widehat{m}_h(\chi)$ (see 4.3 for details about this estimation) and one has the following decomposition:

$$Y/\chi = m(\chi) + \varepsilon/\chi = \widehat{m}_h(\chi) + m(\chi) - \widehat{m}_h(\chi) + \varepsilon/\chi. \tag{6.1}$$

Hence, as the true regression function $m(\chi)$ is unknown in the practice, one needs to approximate it using the bootstrap procedures developed in Section 4.2. This bootstrap procedures together with Theorems 4 and 5 (for naive and wild bootstrap procedures, respectively) allows to approximate the quantity

$$m(\chi) - \widehat{m}_h(\chi)$$

in (6.1) by its bootstrap version

$$\widehat{m}_b(\chi) - \widehat{m}^*_{hb}(\chi).$$

Moreover, as the objective is to build prediction intervals, also the error term ε/χ has to be considered and it will be approximated from the bootstrap procedure too.

As the bootstrap procedure can be applied as many times as one desires, several values for Y/χ can be obtained to build the prediction intervals.

Two situations will be dealt to build prediction intervals: under homoscedasticity, in which the error of the model has constant variance, or heteroscedasticity, where that variance is variable.

Homoscedasticity

Specifically, the procedure to build the prediction intervals for the homoscedastic FNP model is the following:

Given a curve χ, which is a fixed value of the functional predictor, and a sample $\mathcal{S} = \{(\boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{\chi}_n, Y_n)\}$, which is assumed to follow the FNP model

$$Y_i = m(\boldsymbol{\chi}_i) + \varepsilon_i \ (i = 1, \ldots, n),$$

where $\mathcal{S}$ is α-mixing and identically distributed as $(\boldsymbol{\chi}, Y)$, χ is observed from $\boldsymbol{\chi}$ and ε_i are iid, $\mathbb{E}(\varepsilon|\boldsymbol{\chi}) = 0$. As the model is considered to be homoscedastic,

the error of the model has constant variance (that is, $\mathbb{E}(\varepsilon^2|\boldsymbol{\chi}) = \sigma^2_\varepsilon(\boldsymbol{\chi}) = \sigma^2_\varepsilon = \nu$, where ν denotes the conditional variance). Then, the bootstrap $(1-\alpha)$-prediction intervals for Y/χ were constructed as

$$I^*_{\chi,1-\alpha} = (\widehat{m}_h(\chi) + q^*_{\alpha/2}(\chi), \widehat{m}_h(\chi) + q^*_{1-\alpha/2}(\chi)),$$

where the bootstrap quantiles $q^*_p(\chi)$ were computed in the following way:

Step 1. Compute $\widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$, over the dataset $\mathcal{S}$.

Step 2. Compute the residuals $\widehat{\varepsilon}_{i,b} = Y_i - \widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$.

Step 3. Apply the naive bootstrap procedure to obtain the bootstrap errors: Draw n i.i.d random variables $\varepsilon^*_1, \ldots, \varepsilon^*_n$ from the empirical distribution function of $(\widehat{\varepsilon}_{1,b} - \overline{\widehat{\varepsilon}}_b, \ldots, \widehat{\varepsilon}_{n,b} - \overline{\widehat{\varepsilon}}_b)$, where $\overline{\widehat{\varepsilon}}_b = n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_{i,b}$.

Step 4. Obtain $Y^*_i = \widehat{m}_b(\boldsymbol{\chi}_i) + \varepsilon^*_i$, $i = 1, \ldots, n$ and

$$\widehat{m}^*_{hb}(\chi) = \frac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h) Y^*_i}{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h)}.$$

Step 5. Repeat B times Steps 3–4, giving the B estimates $\{\widehat{m}^{*,r}_{hb}(\chi)\}_{r=1}^B$.

Step 6. Draw B i.i.d random variables $\widetilde{\varepsilon}^1, \ldots, \widetilde{\varepsilon}^B$ from the empirical distribution function of the centred residuals from Step 2. $\widetilde{\varepsilon}$ approximate the error of the model.

Step 7. Compute the set of bootstrap errors:

$$ERRORS.BOOT = \{\widehat{m}_b(\chi) - \widehat{m}^{*,r}_{hb}(\chi) + \widetilde{\varepsilon}^r\}_{r=1}^B.$$

Step 8. Compute the bootstrap quantile, $q^*_p(\chi)$, from the quantile of order p of $ERRORS.BOOT$.

Finally, the estimate $\widehat{m}_h(\chi)$ in each one of the intervals was obtained from $\mathcal{S}$. Note that in this situation, the empirical coverage of the prediction intervals will compute the proportion of times that each interval contains the value Y (in contradistinction to the confidence intervals that consider $m(\chi)$).

Note that, in addition, from the algorithm above, one can consider

$$Y^{*,r}|\chi = \widehat{m}_h(\chi) + \widehat{m}_b(\chi) - \widehat{m}^{*,r}_{hb}(\chi) + \widetilde{\varepsilon}^r, r = 1, \ldots, B.$$

Now, using the bootstrap responses $\{Y^{*,r}|\chi\}_{r=1}^B$ one can obtain an estimation for the prediction density of $Y|\chi$ applying, for instance, the kernel density estimator.

Heteroscedasticity

A second case, when the model is heteroscedastic, is dealt. That is, given a curve χ, which is a fixed value of the functional predictor, and a sample $\mathcal{S} = \{(\boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{\chi}_n, Y_n)\}$, which is assumed to follow the FNP model

$$Y_i = m(\boldsymbol{\chi}_i) + \eta_i \ (i = 1, \ldots, n),$$

where $\mathcal{S}$ is α-mixing and identically distributed as $(\boldsymbol{\chi}, Y)$, and χ is observed from $\boldsymbol{\chi}$. Considering that the model has heteroscedatic errors $\eta_i = \sigma(\boldsymbol{\chi}_i)\varepsilon_i, i = 1, \ldots, n$, where ε_i are iid, $\mathbb{E}(\varepsilon_i|\boldsymbol{\chi}_i) = 0$ and $Var(\varepsilon_i|\boldsymbol{\chi}_i) = 1$. Then, $Var(Y|\boldsymbol{\chi}) = Var(\eta|\boldsymbol{\chi}) = \sigma^2(\boldsymbol{\chi}) = \nu(\boldsymbol{\chi})$, where $\nu(\boldsymbol{\chi})$ denotes the error conditional variance.

Under that situation, an algorithm that includes the estimation of the conditional variance $\nu(\boldsymbol{\chi})$ is proposed. This estimation is made following the ideas in Fan and Yao (1998), but adapted to functional data. In their study, Fan and Yao developed a residual based estimator for that conditional variance, which is based on apply a local linear regression model to the squared residuals, and they proved that the same bandwidth selector used to deal with the original regression model can be applied to deal with the squared residuals regression used to estimate the conditional variance. Here, this procedure will be adapted to functional data using the FNP regression model over the squared residuals (see Step 3 below).

The bootstrap $(1-\alpha)$-prediction intervals for Y/χ were constructed again as

$$I^*_{\chi,1-\alpha} = (\widehat{m}_h(\chi) + q^*_{\alpha/2}(\chi), \widehat{m}_h(\chi) + q^*_{1-\alpha/2}(\chi)),$$

where the bootstrap quantiles $q^*_p(\chi)$ were computed in the following way:

Step 1. Compute $\widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$, over the dataset $\mathcal{S}$.

Step 2. Compute the residuals $\widehat{\eta}_i = Y_i - \widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$.

Step 3. Based on the sample $\mathcal{S}_\eta = \{(\boldsymbol{\chi}_1, \widehat{\eta}_1^2), \ldots, (\boldsymbol{\chi}_n, \widehat{\eta}_n^2)\}$, using Nadaraya-Watson estimator for functional data, the estimator for the error conditional variance $\nu(\chi) = \sigma^2(\chi)$ is obtained as:

$$\hat{\nu}_g(\chi) = \frac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/g)\widehat{\eta}_i^2}{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/g)},$$

where g is the bandwidth. In that way, one obtains the estimators for $\nu(\boldsymbol{\chi}_i) = \sigma^2(\boldsymbol{\chi}_i)$ and $\nu(\chi) = \sigma^2(\chi)$, which are denoted as $\hat{\nu}_i = \hat{\sigma}_i^2 = \hat{\nu}_g(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$ and $\hat{\nu}_g(\chi) = \hat{\sigma}_g^2(\chi)$, respectively.

One can now obtain

$$\widehat{\varepsilon}_i = \frac{\widehat{\eta}_i}{\hat{\sigma}_i}, \quad i = 1, \ldots, n,$$

which are the standardize residuals of the model.

Step 4. Apply the naive bootstrap procedure to obtain the bootstrap errors: Draw n i.i.d random variables $\varepsilon_1^*, \ldots, \varepsilon_n^*$ from the empirical distribution function of $(\widehat{\varepsilon}_1 - \overline{\widehat{\varepsilon}}, \ldots, \widehat{\varepsilon}_n - \overline{\widehat{\varepsilon}})$, where $\overline{\widehat{\varepsilon}} = n^{-1} \sum_{i=1}^n \widehat{\varepsilon}_i$.

Step 5. Obtain $Y_i^* = \widehat{m}_b(\boldsymbol{\chi}_i) + \hat{\sigma}_i \varepsilon_i^*$, $i = 1, \ldots, n$ and

$$\widehat{m}_{hb}^*(\chi) = \frac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h) Y_i^*}{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i, \chi)/h)}.$$

Step 6. Repeat B times Steps 4-5, giving the B estimates $\{\widehat{m}_{hb}^{*,r}(\chi)\}_{r=1}^B$.

Step 7. Draw B i.i.d random variables $\widetilde{\varepsilon}^1, \ldots, \widetilde{\varepsilon}^B$ from the empirical distribution function of the centred residuals from Step 4, and compute $\widetilde{\eta}^r = \hat{\sigma}_g(\chi)\widetilde{\varepsilon}^r$, $r = 1, \ldots, B$. $\widetilde{\eta} = \hat{\sigma}_g(\chi)\widetilde{\varepsilon}$ approximate the error of the model.

Step 8. Compute the set of bootstrap errors:

$$ERRORS.BOOT = \{\widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi) + \widetilde{\eta}^r\}_{r=1}^B.$$

Step 9. Compute the bootstrap quantile, $q_p^*(\chi)$, from the quantile of order p of $ERRORS.BOOT$.

The bootstrap procedures developed in Chapters 4 and 5 for heteroscedastic models are based on the use of wild bootstrap. In the algorithm above, wild bootstrap can be applied to approximate $m(\chi) - \widehat{m}_h(\chi)$ in (6.1) by its bootstrap version $\widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi)$. However, one cannot use wild bootstrap to resample the error of the model, $\widetilde{\eta}^r$, which has to be obtained as indicated in the step 7 above.

Note that, again, from the algorithm above, one can consider

$$Y^{*,r}|\chi = \widehat{m}_h(\chi) + \widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi) + \widetilde{\eta}^r, r = 1, \ldots, B$$

and, using the bootstrap responses $\{Y^{*,r}|\chi\}_{r=1}^B$, one can obtain an estimation for the prediction density of $Y|\chi$ applying, for instance, the kernel density estimator.

6.2.2 Prediction intervals for Semi-Functional Partial Linear model

Moving to the SFPL model, the procedure is analogous. In this case, the sample will be $\mathcal{S}$, where

$$\mathcal{S} = \{(\boldsymbol{X}_1, \boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{\chi}_n, Y_n)\}.$$

Based on this sample for the SFPL model (5.1), the predictor for $Y/\{\boldsymbol{X}, \chi\}$ is $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_h + \widehat{m}_h(\chi)$ (see Subsection 5.2.1 for details about this estimators), where $\widehat{m}_h(\chi)$ differs from the one used in the FNP model as:

$$\widehat{m}_h(\chi) = \sum_{i=1}^{n} w_h(\boldsymbol{\chi}_i, \chi)(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_h).$$

Note that now, the distribution of interest is Y conditionally on both the scalar-linear covariate $\boldsymbol{X}$ and the functional-nonparametric one χ.

One has the following decomposition:

$$Y/\{\boldsymbol{X}, \chi\} = \boldsymbol{X}^T\boldsymbol{\beta} + m(\chi) + \varepsilon/\{\boldsymbol{X}, \chi\} =$$
$$\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_h + \boldsymbol{X}^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_h) + \widehat{m}_h(\chi) + m(\chi) - \widehat{m}_h(\chi) + \varepsilon/\{\boldsymbol{X}, \chi\}. \tag{6.2}$$

Hence, as the true values for the regression function $m(\chi)$ and the parameters vector $\boldsymbol{\beta}$ are unknown in practice, one needs to approximate it using the bootstrap procedures developed in Section 5.2.2. This bootstrap procedures together with Theorems 10–11 for the linear component (for naive and wild bootstrap procedures, respectively) and Theorems 14–15 for the nonparametric component (again for naive and wild bootstrap procedures, respectively) allows to approximate the quantities

$$\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_h \text{ and } m(\chi) - \widehat{m}_h(\chi)$$

in (6.2) by its bootstrap version

$$\widehat{\boldsymbol{\beta}}_h - \widehat{\boldsymbol{\beta}}_h^* \text{ and } \widehat{m}_b(\chi) - \widehat{m}_{hb}^*(\chi).$$

Also, as the objective is to build prediction intervals, the error of the model has to be considered and it will be approximated from the bootstrap procedure too. Again, as the bootstrap procedure can be applied as many times as one desires, several values for $Y/\{\boldsymbol{X}, \chi\}$ can be obtained to build the prediction intervals.

The algorithms will be developed to deal with both homoscedastic and heteroscedastic models, following the ideas in Subsection 6.2.1.

Homoscedasticity

Specifically, the procedure to build the prediction intervals for the homoscedastic SFPL model is the following:

Given $\{\boldsymbol{X}, \chi\}$, which are fixed values of the predictors, and a sample $\mathcal{S}' = \{(\boldsymbol{X}_1, \boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{\chi}_n, Y_n)\}$, which is assumed to follow the SFPL model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

where the process $\{(\boldsymbol{X}_i, \boldsymbol{\chi}_i, Y_i)\}$ is α-mixing and identically distributed as $(\boldsymbol{X}, \boldsymbol{\chi}, Y)$, ε_i are iid, $\mathbb{E}(\varepsilon|\boldsymbol{\chi}) = 0$ and, under homoscedasticity, $\mathbb{E}(\varepsilon^2|\boldsymbol{\chi}) = \sigma_\varepsilon^2(\boldsymbol{\chi}) = \sigma_\varepsilon^2 = \nu$. Then, the bootstrap $(1-\alpha)$-prediction intervals for $Y/\{\boldsymbol{X}, \chi\}$ were constructed as

$$I^*_{\boldsymbol{X},\chi,1-\alpha} = (\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_h + \widehat{m}_h(\chi) + q^*_{\alpha/2}(\boldsymbol{X}, \chi), \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_h + \widehat{m}_h(\chi) + q^*_{1-\alpha/2}(\boldsymbol{X}, \chi)),$$

where the bootstrap quantiles $q^*_p(\boldsymbol{X}, \chi)$ were computed in the following way:

Step 1. Compute $\widehat{\boldsymbol{\beta}}_b$ and $\widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$, over the dataset $\mathcal{S}'$.

Step 2. Compute the residuals $\widehat{\varepsilon}_{i,b} = Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b - \widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1, \ldots, n$.

Step 3. Apply the naive bootstrap procedure to obtain the bootstrap errors: Draw n i.i.d. random variables $\varepsilon_1^*, \ldots, \varepsilon_n^*$ from the empirical distribution function of $(\widehat{\varepsilon}_{1,b} - \overline{\widehat{\varepsilon}}_b, \ldots, \widehat{\varepsilon}_{n,b} - \overline{\widehat{\varepsilon}}_b)$, where $\overline{\widehat{\varepsilon}}_b = n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_{i,b}$.

Step 4. Obtain $Y_i^* = \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b + \widehat{m}_b(\boldsymbol{\chi}_i) + \varepsilon_i^*$, $i = 1, \ldots, n$ and the bootstrap estimators

$$\widehat{\boldsymbol{\beta}}_b^* = (\widetilde{\mathbf{X}}_b^T\widetilde{\mathbf{X}}_b)^{-1}\widetilde{\mathbf{X}}_b^T\widetilde{\mathbf{Y}^*}_b$$

and

$$\widehat{m}^*_{hb}(\chi) = \sum_{i=1}^n w_h(\boldsymbol{\chi}_i, \chi)(Y_i^* - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b^*),$$

Step 5. Repeat B times Steps 3–4, giving the B estimates

$$\{\widehat{\boldsymbol{\beta}}_b^{*,r}\}_{r=1}^B \text{ and } \{\widehat{m}^{*,r}_{hb}(\chi)\}_{r=1}^B.$$

Step 6. Draw B i.i.d random variables $\widetilde{\varepsilon}^1, \ldots, \widetilde{\varepsilon}^B$ from the empirical distribution function of the centred residuals from Step 2. $\widetilde{\varepsilon}$ approximates the error of the model.

Step 7. Compute the set of bootstrap errors:

$$ERRORS.BOOT = \left\{ \boldsymbol{X}^T(\widehat{\boldsymbol{\beta}}_b - \widehat{\boldsymbol{\beta}}_b^{*,r}) + (\widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi)) + \widetilde{\varepsilon}^r \right\}_{r=1}^{B},$$

Step 8. Compute the bootstrap quantile, $q_p^*(\boldsymbol{X}, \chi)$, from the quantile of order p of $ERRORS.BOOT$.

Following the same idea as for the FNP model, one can get, from the algorithm above:

$$Y^{*,r}/\{\boldsymbol{X}, \chi\} = \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_h + \widehat{m}_h(\chi) + \boldsymbol{X}^T(\widehat{\boldsymbol{\beta}}_b - \widehat{\boldsymbol{\beta}}_b^{*,r}) + \widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi) + \widetilde{\varepsilon}^r,$$

where $r = 1, \ldots, B$. Now, the prediction density of $Y/\{\boldsymbol{X}, \chi\}$ can be obtained from the set of bootstrap responses $\{Y^{*,r}/\{\boldsymbol{X}, \chi\}\}_{r=1}^{B}$ using, for example, the kernel density estimator.

Heteroscedasticity

Finally, the procedure to build the prediction intervals for the heteroscedastic SFPL model is the following:

Given $\{\boldsymbol{X}, \chi\}$, which are fixed values of the predictors, and a sample $\mathcal{S}' = \{(\boldsymbol{X}_1, \boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{\chi}_n, Y_n)\}$, which is assumed to follow the SFPL model

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + m(\boldsymbol{\chi}_i) + \eta_i,\ i = 1, \ldots, n,$$

where the process $\{(\boldsymbol{X}_i, \boldsymbol{\chi}_i, Y_i)\}$ is α-mixing and identically distributed as $(\boldsymbol{X}, \boldsymbol{\chi}, Y)$. The errors of the model are heteroscedastic, that is, $\eta_i = \sigma(\boldsymbol{\chi}_i)\varepsilon_i$, $i = 1, \ldots, n$, where ε_i are iid, $\mathbb{E}(\varepsilon_i|\boldsymbol{\chi}_i) = 0$ and $Var(\varepsilon_i|\boldsymbol{\chi}_i) = 1$. Then, $Var(Y|\boldsymbol{\chi}) = Var(\eta|\boldsymbol{\chi}) = \sigma^2(\boldsymbol{\chi}) = \nu(\boldsymbol{\chi})$, where $\nu(\boldsymbol{\chi})$ denotes the error conditional variance.

In the last paragraph, one assumes that the conditional variance, $\nu(\boldsymbol{\chi})$, only depends on the functional explanatory variable $\boldsymbol{\chi}$ and not on the scalar one, $\boldsymbol{X}$ (see assumption (5.8) in Chapter 5). Hence, it can be estimated following a nonparametric estimator similar than in Fan and Yao (1998) but adapted to functional data, as it was applied in the previous algorithm for the FNP model.

The bootstrap $(1-\alpha)$-prediction intervals for $Y/\{\boldsymbol{X}, \chi\}$ were constructed as

$$I^*_{\boldsymbol{X},\chi,1-\alpha} = (\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_h + \widehat{m}_h(\chi) + q^*_{\alpha/2}(\boldsymbol{X},\chi), \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_h + \widehat{m}_h(\chi) + q^*_{1-\alpha/2}(\boldsymbol{X},\chi)),$$

where the bootstrap quantiles $q^*_p(\boldsymbol{X},\chi)$ were computed in the following way:

Step 1. Compute $\widehat{\boldsymbol{\beta}}_b$ and $\widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1,\ldots,n$, over the dataset $\mathcal{S}'$.

Step 2. Compute the residuals $\widehat{\eta}_i = Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b - \widehat{m}_b(\boldsymbol{\chi}_i)$, $i = 1,\ldots,n$.

Step 3. Based on the sample $\mathcal{S}_\eta = \{(\boldsymbol{\chi}_1,\widehat{\eta}_1^2),\ldots,(\boldsymbol{\chi}_n,\widehat{\eta}_n^2)\}$, using Nadaraya-Watson estimator for functional data, the estimator for the error conditional variance $\nu(\chi) = \sigma^2(\chi) = Var(Y/\chi)$ is obtained as:

$$\hat{\nu}_g(\chi) = \frac{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i,\chi)/g)\widehat{\eta}_i^2}{\sum_{i=1}^n K(d(\boldsymbol{\chi}_i,\chi)/g)},$$

where g is the bandwidth. In that way, one obtains the estimators for $\nu(\boldsymbol{\chi}_i) = \sigma^2(\boldsymbol{\chi}_i)$ and $\nu(\chi) = \sigma^2(\chi)$, which are denoted as $\hat{\nu}_i = \hat{\sigma}_i^2 = \hat{\nu}_g(\boldsymbol{\chi}_i)$, $i = 1,\ldots,n$ and $\hat{\nu}_g(\chi) = \hat{\sigma}_g^2(\chi)$, respectively.

One can now obtain

$$\widehat{\varepsilon}_i = \frac{\widehat{\eta}_i}{\hat{\sigma}_i}, \quad i = 1,\ldots,n,$$

which are the standardize residuals of the model.

Step 4. Apply the naive bootstrap procedure to obtain the bootstrap errors: Draw n i.i.d. random variables $\varepsilon_1^*,\ldots,\varepsilon_n^*$ from the empirical distribution function of $(\widehat{\varepsilon}_1 - \overline{\widehat{\varepsilon}},\ldots,\widehat{\varepsilon}_n - \overline{\widehat{\varepsilon}})$, where $\overline{\widehat{\varepsilon}} = n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i$.

Step 5. Obtain

$$Y_i^* = \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b + \widehat{m}_b(\boldsymbol{\chi}_i) + \hat{\sigma}_i\varepsilon_i^*, \quad i = 1,\ldots,n$$

and the bootstrap estimators

$$\widehat{\boldsymbol{\beta}}_b^* = (\widetilde{\mathbf{X}}_b^T\widetilde{\mathbf{X}}_b)^{-1}\widetilde{\mathbf{X}}_b^T\widetilde{\mathbf{Y}^*}_b$$

and

$$\widehat{m}^*_{hb}(\chi) = \sum_{i=1}^n w_h(\boldsymbol{\chi}_i,\chi)(Y_i^* - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_b^*),$$

Step 6. Repeat B times Steps 4-5, giving the B estimates

$$\{\widehat{\boldsymbol{\beta}}_b^{*,r}\}_{r=1}^B \text{ and } \{\widehat{m}^{*,r}_{hb}(\chi)\}_{r=1}^B.$$

Step 7. Draw B i.i.d random variables $\widetilde{\varepsilon}^1, \ldots, \widetilde{\varepsilon}^B$ from the empirical distribution function of the centred residuals from Step 4 and compute $\widetilde{\eta}^r = \hat{\sigma}_g(\chi)\widetilde{\varepsilon}^r$, $r = 1, \ldots, B$. $\widetilde{\eta} = \hat{\sigma}_g(\chi)\widetilde{\varepsilon}$ approximate the error of the model.

Step 8. Compute the set of bootstrap errors:

$$ERRORS.BOOT = \left\{ \boldsymbol{X}^T(\widehat{\boldsymbol{\beta}}_b - \widehat{\boldsymbol{\beta}}_b^{*,r}) + (\widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi)) + \widetilde{\eta}^r \right\}_{r=1}^{B},$$

Step 9. Compute the bootstrap quantile, $q_p^*(\boldsymbol{X}, \chi)$, from the quantile of order p of $ERRORS.BOOT$.

As indicated above, one assumes that the conditional variance, $\nu(\boldsymbol{\chi})$, only depends on the functional explanatory variable $\boldsymbol{\chi}$ and not on the scalar one, $\boldsymbol{X}$. This allows to estimate it using a nonparametric estimator in Step 3. However, if one wants to assume that this conditional variance depends on both covariates, $\boldsymbol{X}$ and $\boldsymbol{\chi}$, one may consider the expression $Var(Y|(\boldsymbol{X}, \boldsymbol{\chi})) = Var(\eta|(\boldsymbol{X}, \boldsymbol{\chi})) = \sigma_\varepsilon^2(\boldsymbol{X}, \boldsymbol{\chi}) = \nu(\boldsymbol{X}, \boldsymbol{\chi})$. It that general case, the estimation of $\nu(\boldsymbol{X}, \boldsymbol{\chi})$ cannot be done by a nonparametric estimator, and alternatives as partial linear or additive models need to be employed.

Related to the bootstrap methodology, an alternative to this algorithm could be to apply wild bootstrap procedure to approximate the quantity

$$\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_h \text{ and } m(\chi) - \widehat{m}_h(\chi)$$

in (6.2) by its bootstrap version

$$\widehat{\boldsymbol{\beta}}_h - \widehat{\boldsymbol{\beta}}_h^* \text{ and } \widehat{m}_b(\chi) - \widehat{m}_{hb}^*(\chi),$$

as it was done to build the bootstrap confidence intervals in Chapter 5. However, the error of the model has to be approximated by $\widetilde{\eta}^r$, as it was obtained in Step 7 above.

Finally, as in the previous algorithms, one can get

$$Y^{*,r}/\{\boldsymbol{X}, \chi\} = \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_h + \widehat{m}_h(\chi) + \boldsymbol{X}^T(\widehat{\boldsymbol{\beta}}_b - \widehat{\boldsymbol{\beta}}_b^{*,r}) + \widehat{m}_b(\chi) - \widehat{m}_{hb}^{*,r}(\chi) + \widetilde{\eta}^r,$$

where $r = 1, \ldots, B$ and thus, the prediction density of $Y/\{\boldsymbol{X}, \chi\}$ can be obtained from the set of bootstrap responses $\{Y^{*,r}/\{\boldsymbol{X}, \chi\}\}_{r=1}^B$ using, for example, the kernel density estimator.

6.3 Application to electricity data

An application to the electrical dataset, separately for the electricity demand and price, is carried out in this section. Within this application, both FNP and SFPL regression models are considered to build prediction intervals. As in the applications reported in Chapters 4 and 5, because of its generality, heteroscedastic models will be contemplated.

Prediction intervals will be calculated following the applications in Chapters 4 and 5. That is, to consider the weekdays from the second quarter of the year 2012, obtaining prediction intervals in two situations: predicting the 24 hours of the same day (June 29, 2012) or considering the same fixed hour along 21 consecutive days. Those prediction intervals will be compared with the confidence intervals obtained in the previous Sections 4.5 and 5.5.

The FNP and SFPL regression models used to predict hourly demand or price are applied in the same way as Sections 4.5 and 5.5, respectively, using as functional covariate the previous daily curve of electricity demand or price. Within the SFPL model, the scalar covariates include information about the temperature through HDD and CDD variables, when dealing with demand, or daily demand and wind power production, when dealing with price.

Together with the prediction intervals, also the prediction density will be computed. In this case, a comparison between some hours in the same day will be presented, showing the different performance of the electricity demand and price.

Along the applications, the parameters involved in the bootstrap procedures will be selected following the ideas in Sections 4.5 and 5.5, and thus, the smoothing parameters b and g are selected using a cross-validation method and h is chosen equal to b. 1000 bootstrap replicates were drawn, quadratic kernel was used and the class of projection-based semi-metrics was considered. The confidence level is $1 - \alpha = 0.95$.

6.3.1 Case study: electricity demand

Prediction intervals were computed for the electricity demand, following the FNP and SFPL regression models. The results are shown separately for each one of the functional regression models in the following pages.

FNP model

The application to the electricity demand of the algorithm developed to build prediction intervals for the FNP regression model under heteroscedasticity is presented in the next paragraphs.

First, the case in which one predicts the 24 hour of the same day (June 29, 2012) is considered. The prediction intervals, together with the observed demand are represented in the right panel of Figure 6.1. Left panel in this Figure shows the confidence intervals obtained in the previous Subsection 4.5.1 using the same FNP model, in chronological order.

One can distinguish in both graphs the shape of the daily demand curves and see how it performs along the 24 hours of the day. Comparing the confidence and prediction intervals one can see that, as expected, the prediction intervals are longer than the confidence intervals and they cover the observed demand for each hour of the day. It is worth to point out that confidence intervals are not really designed to cover the observed demand, as they deal with the expectation, and that is the reason why is necessary to extend the algorithms in Chapter 4.

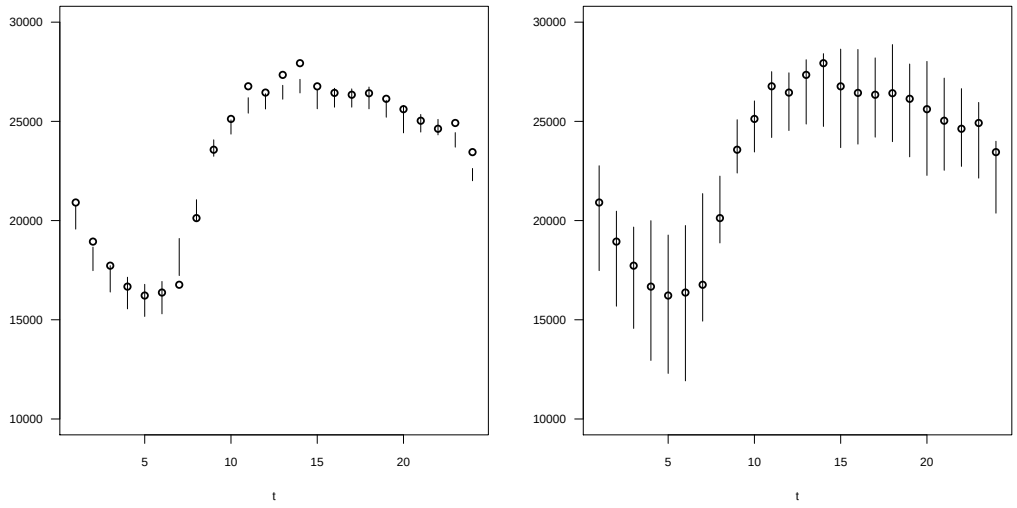


Figure 6.1: Bootstrap confidence intervals (left panel) and prediction intervals (right panel) for the electricity demand using the FNP model and considering a fixed day (June 29, 2012). The circles are the observed demand.

Then, the second case in which one predicts the same fixed hour (20:00) along some consecutive days (weekdays in June, 2012) is dealt. The prediction intervals are plotted in the right panel of Figure 6.2. As in the previous Figure 6.1, the left panel represents the corresponding confidence intervals and observed demand in chronological order.

One can see that demand values remain stable along the days for the same fixed hour. Comparing with Figure 6.1, there are not big differences between the level of demand at 20:00 hour along some days. This is expectable, since the demand barely fluctuate along the same group of days during that period of time (one month). Again, the prediction intervals are longer than the confidence intervals.

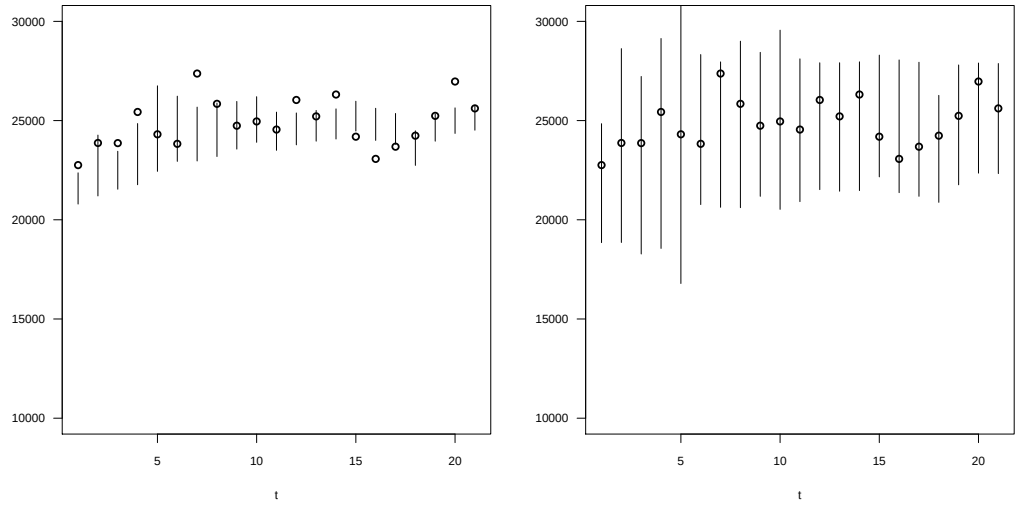


Figure 6.2: Bootstrap confidence intervals (left panel) and prediction intervals (right panel) for the electricity demand using the FNP model and considering a fixed hour, 20:00, along June, 2012. The circles are the observed demand.

In addition to the intervals, also the bootstrap prediction density can be computed. This is done for some fixed hours in the day June 29, 2012, and is represented in Figure 6.3.

Comparing the prediction densities plotted in Figure 6.3, one can distinguish the different behaviour of the demand during the night (hours 1:00 and 5:00) and during the day (hours 10:00, 15:00 and 20:00). In the first case, the demand covers low values, specially at 5:00. For the second case, during the day, there are not many differences between the density curves, as they

cover more or less the same range of demand. This is coincident with the descriptive analysis in Section 1.4.1, in which the shape of the daily demand curves was analysed.

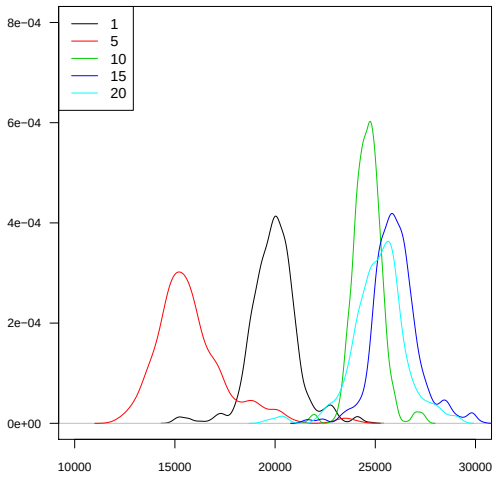


Figure 6.3: Prediction density for the electricity demand using the FNP model, along June 29, 2012 and comparing some hours along the day (1:00 in black line, 5:00 in red, 10:00 in green, 15:00 in dark blue and 20:00 in light blue).

SFPL model

The analysis reported above for the electricity demand using the FNP model can be replayed now considering the SFPL regression model, in which information about temperature is added to the model as scalar covariates (see Subsection 5.5.1 for details).

Prediction intervals are obtained first for the 24 hours of June 29, 2012, in Figure 6.4 and then for the weekdays in June, setting the hour 20:00, in Figure 6.5. In both cases they are compared with the corresponding confidence intervals obtained in Section 5.5.1

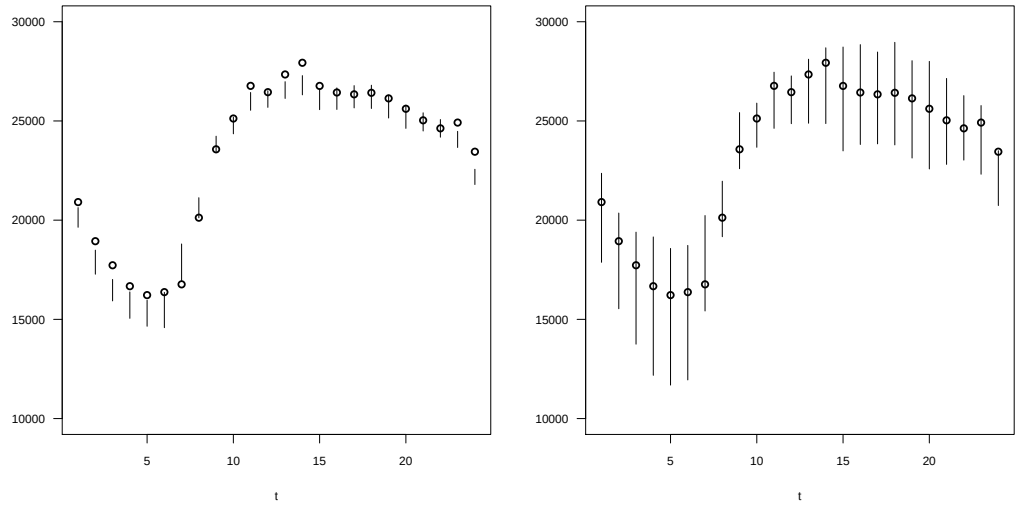


Figure 6.4: Bootstrap confidence intervals (left panel) and prediction intervals (right panel) for the electricity demand using the SFPL model and considering a fixed day (June 29, 2012). The circles are the observed demand.

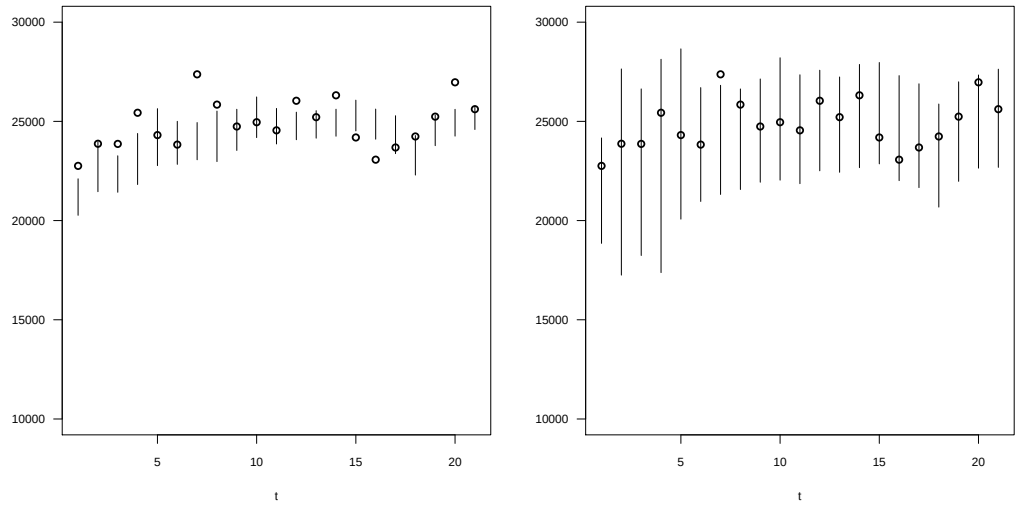


Figure 6.5: Bootstrap confidence intervals (left panel) and prediction intervals (right panel) for the electricity demand using the SFPL model and considering a fixed hour, 20:00, along June, 2012. The circles are the observed demand.

In the first case, looking at Figure 6.4, both confidence and prediction intervals are very similar as those obtained for the FNP model in Figure 6.1. Again, the prediction intervals are larger than the corresponding confidence intervals and they already cover the observed demand at each hour of the day. However, in Figure 6.5 one can see that the prediction intervals are slightly shorter for the SFPL than the ones for the FNP model in Figure 6.2. In any case, the results are very similar for both functional regression models, FNP and SFPL.

Finally, the prediction density is analysed. Figure 6.6 represents the prediction density for the day June 29, 2012, comparing some hours in this day. Similar comments as those made for FNP model, regarding Figure 6.3 can be used here, as there are not many changes in the graphs comparing both functional regression models. Again, the prediction density differs from the night (1:00 and 5:00) and during the day (10:00, 15:00 and 20:00).

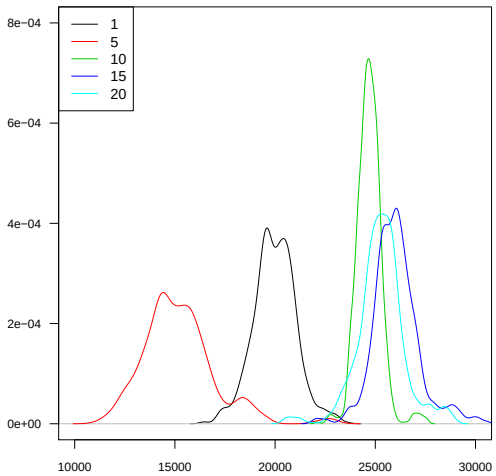


Figure 6.6: Prediction density for the electricity demand using the SFPL model, along June 29, 2012 and comparing some hours along the day (1:00 in black line, 5:00 in red, 10:00 in green, 15:00 in dark blue and 20:00 in light blue).

6.3.2 Case study: electricity price

In this subsection, the prediction intervals were computed for the electricity price, following the FNP and SFPL regression models. The results are also shown separately for each one of the regression models in the following pages, proceeding with the same outline used previously for electricity demand.

FNP model

FNP regression model is applied to build the prediction intervals for electricity price, which are compared with the confidence intervals obtained previously in Section 4.5.2. First, the prediction intervals are represented in Figure 6.7 for the 24 hours of the day June 29, 2012. The fluctuations along the 24 hours of the day are now smooth, as price remains more stable than the electricity demand.

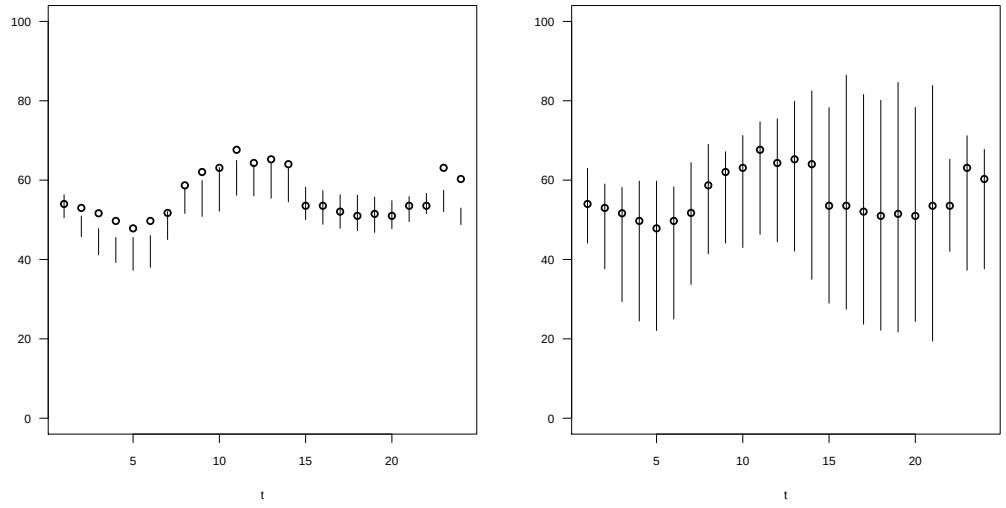


Figure 6.7: Bootstrap confidence intervals (left panel) and prediction intervals (right panel) for the electricity price using the FNP model and considering a fixed day (June 29, 2012). The circles are the observed price.

Comparing the prediction intervals in this figure with those obtained in the demand case, in Figure 6.1, one can observe higher variability in the intervals amplitude. Since the prediction intervals for electricity demand were stable in terms of length, now for the price there are more variations along the hours of the day. For instance, during the first hours of the day, the prediction intervals are smaller than during the rest of the day.

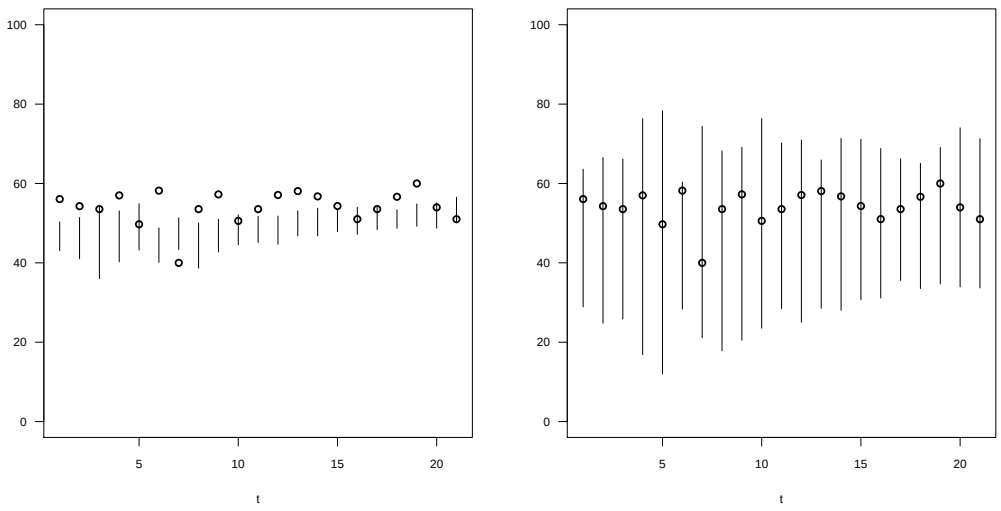


Figure 6.8: Bootstrap confidence intervals (left panel) and prediction intervals (right panel) for the electricity price using the FNP model and considering a fixed hour, 20:00, along June, 2012. The circles are the observed price.

Figure 6.8 represents the confidence and prediction intervals, considering in this case the hour 20:00 along the weekdays of June, 2012. There are not many differences in the price levels between the days. One can see that the prediction intervals are quite stable and they are larger than the corresponding confidence intervals. As expected, the prediction intervals cover the observed hourly price.

Finally, also the prediction density was obtained for the electricity price using the FNP model. Figure 6.9 represents the density obtained for June 29, 2012, comparing some hours of the day. One can see that the curves are closer than in the demand. However, the performance of the price during the night is still different, specially at 5:00, as it reaches lower values than during the day.

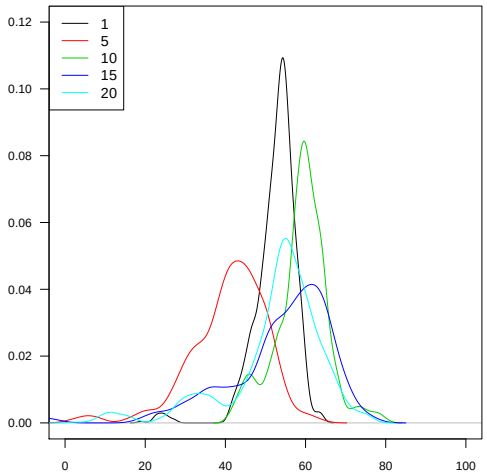


Figure 6.9: Prediction density for the electricity price using the FNP model, along June 29, 2012 and comparing some hours along the day (1:00 in black line, 5:00 in red, 10:00 in green, 15:00 in dark blue and 20:00 in light blue).

SFPL model

Prediction intervals were computed also for the electricity price, following the SFPL regression model. This application is analogous to the previous case, in which the FNP model was used, but adding as scalar covariates the demand and wind power production (see Section 5.5.2 for details).

Analogous results as in the previous sections are shown here. First, considering the 24 hours of June 29, 2012, in Figure 6.10, and then for the hour 20:00 along the weekdays in June, in Figure 6.11.

The difference between the prediction intervals for the two functional regression models is now more remarkable than in the case of the demand. Comparing Figures 6.10 and 6.11 for the SFPL model with Figures 6.7 and 6.8 for the FNP model, respectively, one can see that the intervals for the SFPL model are smaller than the ones for the FNP model. In any case, the prediction intervals are longer than the corresponding confidence intervals and they cover the observed hourly price almost always.

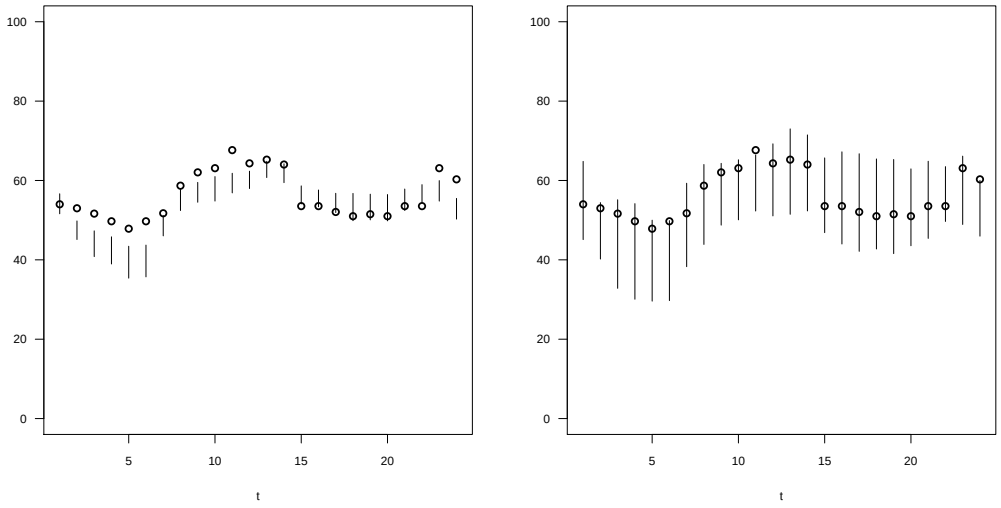


Figure 6.10: Bootstrap confidence intervals (left panel) and prediction intervals (right panel) for the electricity price using the SFPL model and considering a fixed day (June 29, 2012). The circles are the observed price.

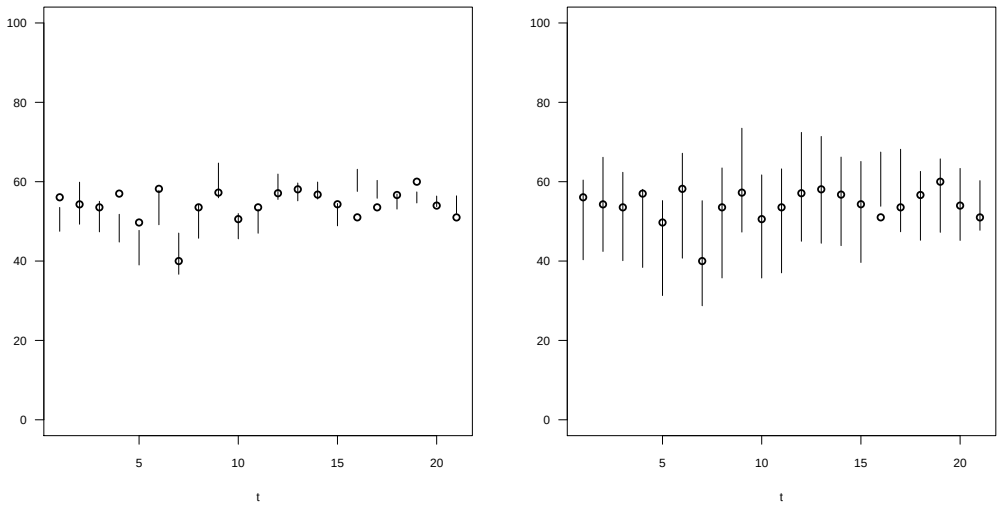


Figure 6.11: Bootstrap confidence intervals (left panel) and prediction intervals (right panel) for the electricity price using the SFPL model and considering a fixed hour, 20:00, along June, 2012. The circles are the observed price.

Following the same outline as in the previous cases, also the prediction densities were computed for the electricity price with SFPL model. In Figure 6.12 one can see the prediction density estimated for June 29, 2012 comparing some hours of the day. Again, one can distinguish the different behaviour between the hour 5:00 and the rest of the day. In this case, the prediction densities during the other considered hours are almost overlapped and this may indicate that the performance of the predictions for those hours is similar for the electricity price using SFPL model.

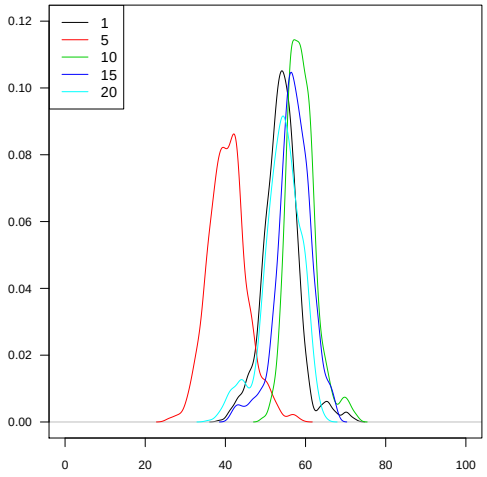


Figure 6.12: Prediction density for the electricity price using the SFPL model, along June 29, 2012 and comparing some hours along the day (1:00 in black line, 5:00 in red, 10:00 in green, 15:00 in dark blue and 20:00 in light blue).

6.4 Conclusions

This chapter has extended the bootstrap procedures developed in chapters 4 and 5, originally applied to construct confidence intervals, to build prediction intervals. This last chapter in the memory also complements the predictions developed in Chapter 3, following the recommendation in Weron (2006) to give a step forward in the problem of prediction in electricity markets giving, not only pointwise predictions, but also prediction intervals and computing prediction density.

The algorithms developed in this chapter are designed to deal with homoscedastic and heteroscedastic models. For the latter case, it is not enough to apply wild bootstrap procedures as for the confidence intervals, because this wild bootstrap can be only applied to approximate the error of the regression estimator, but not the error of the model. Thus, the algorithms developed to use with heteroscedastic models follow the ideas in Fan and Yao (1998) in order to estimate the conditional variance.

Prediction intervals for both electricity demand and price were obtained through the two functional regression models: FNP and also SFPL regression models, and they were also compared with the confidence intervals from the previous Sections 4.5 and 5.5. This applications extends and complements applications in Chapters 4 and 5, in which bootstrap confidence intervals were obtained for demand and price, obtaining now the prediction intervals for the same dataset and also prediction density.

Conclusions

This memory has presented a detailed statistical analysis of electricity demand and price from the Spanish Electricity Market, using functional techniques. This analysis started with an introduction to the electricity market and a descriptive study of the dataset used later.

As any other statistical study, it begins with a detection of the outliers in the dataset, which can disturb the techniques applied in it. Since the study works with functional time series and, up to our knowledge, the available tools in the literature for functional data are not design to deal with this kind of structure, up to three different methods specifically addressed to detect functional outliers in functional time series were developed. Derived from this methodology, which is explained in Chapter 2, two papers were published. The first one, entitled "Detection of outliers in functional time series" by Raña, P., Aneiros, G. and Vilar, J.M. (2015), can be found in *Environmetrics* and it includes the Depth-based trimming method for Functional Time Series described in Section 2.3. The second one, entitled "Using robust FPCA to identify outliers in functional time series, with applications to the electricity market" by Vilar, J.M., Raña, P. and Aneiros, G., including the Projections-based and Prediction errors-based methods presented in Section 2.4, will be found in *SORT*.

A comparative study between different prediction methods in this electrical data field was carried out in Chapter 3. This comparison includes classical tools as the Naïve method or time series analysis through ARIMA models and also proposals based on robust FPCA. The proposal is to apply in this field functional regression models as the FNP or the SFPL model with both scalar or functional response and adding external covariates. This proposal extends Vilar et al. (2012) and has been also included in a paper entitled "Short-term forecast of daily curves of electricity demand and price" by Aneiros, G., Vilar, J. and Raña, P. (2016), which is already published in

International Journal of Electrical Power and Energy Systems.

Bootstrap procedures were established to deal with two functional regression models, the FNP model in Chapter 4 and the SFPL model in 5. In both cases the validity of the bootstrap procedures was proved theoretically and they were applied to build confidence intervals. This application was also extended to build prediction intervals in Chapter 6.

Concerning the bootstrap in FNP model, there is a preceding in Ferraty, F., Van Keilegom, I. and Vieu, P. (2010) in which they proposed and proved the validity of the bootstrap methods for independent data. Our contribution in this topic consists in extend this result to work under dependence conditions on the sample. The paper entitled "Bootstrap confidence intervals in FNP regression under dependence" by Raña, P., Aneiros, G., Vilar, J. and Vieu, P. (2016), which can be found in *Electronic Journal of Statistics*, summarizes both the theoretical development and the simulations and applications presented in Chapter 4.

Nevertheless, when dealing with SFPL model there is no preceding in the literature, up to our knowledge, which achieves the bootstrap procedures presented in Chapter 5 for this regression model. Moreover, it is even difficult to find studies about bootstrap procedures in the classical partial linear regression models. One can find in Liang et al. (2000) and You and Chen (2006) proposals regarding partial linear regression for fixed design and independent data. Thus, probably the main contribution of this memory is the proposal of the bootstrap procedures applied to SFPL model under dependence together with the theoretical proof of its validity. A paper which summarizes the contents from this chapter, both theoretical results and applications is under preparation.

Aforementioned bootstrap procedures were applied first to build confidence intervals for the regression estimators, but this has been extended to build also prediction intervals. When dealing with forecasting, as in Chapter 3 it is important to complement the pointwise predictions with other tools as the prediction intervals. This has been also pointed out by Weron (2006), indicating that most of the studies dealing with prediction in electricity markets are focus on pointwise prediction and that this could not be enough, encouraging researches to go deeper within this problem. Then, this memory contributes with the prediction intervals and prediction densities built for the FNP and SFPL regression models, included in Chapter 6, which conclude this study. A paper which summarizes the contents from this chapter,

both bootstrap algorithms and applications is under preparation.

As in every statistical process, this is not the final point for this project. There are many open problems to be dealt in future researches. Some of them were already pointed out in the conclusions for each chapter, but the most remarkable ones are summarized in the following points:

- Outlier detection methods developed in Chapter 2 can be modified, for instance, to use other new functional depths which improve the results. The computational models to visualize functional data are evolving and so, they can also contribute to the problem of outlier detection.

- Related to Chapter 3, it could be interesting to add new models to the comparison in order to improve the prediction errors. In that sense, the additive models for FDA could be a nice alternative due to its flexibility.

- The theory related to the bootstrap procedures developed in Chapters 4 and 5 to build confidence intervals, which was analysed restricting to models with scalar response, could be extended to work with functional response. Even if we are aware that this is a challenging problem from the theoretical point of view, it could be addressed following the developments in Ferraty el al. (2012).

- Reaching that question of the bootstrap procedures for regression models with functional response, they can be applied to both confidence and prediction tubes.

- Chapter 6 includes procedures to build prediction intervals for heteroscedastic models, by estimating the conditional variance. Other procedures could be also considered within these procedures to approximate the error of the regression model.

Appendix A

Auxiliary results

Lemma 17 *(Lemma 1, Ferraty et al., 2007) Under assumptions (4.4), (4.5), (4.6), (4.9) and (4.10) we have that*

$$\frac{\mathbb{E}\left(\widehat{g}_h(\chi)\right)}{\mathbb{E}\left(\widehat{f}_h(\chi)\right)} - m(\chi) = h\varphi'_\chi(0)J_\chi + o(h).$$

Proof. Because dependence does not influence on the expectation of $\widehat{g}_h(\chi)$ nor $\widehat{f}_h(\chi)$, the proof of this lemma is the same as that obtained for Lemma 1 in Ferraty et al. (2007) for the independent case. □

Lemma 18 *(Lemma 2, Ferraty et al., 2007) Under assumptions (4.6), (4.9) and (4.10) we have that*

$$J_\chi = \frac{K(1) - \int_0^1 (sK(s))'\tau_{h\chi}(s)ds}{K(1) - \int_0^1 K'(s)\tau_{h\chi}(s)ds} \longrightarrow \frac{M_{0\chi}}{M_{1\chi}} \text{ as } n \longrightarrow \infty.$$

Proof. Because J is not random, the proof of this lemma is the same as that obtained for Lemma 2 in Ferraty et al. (2007).

Lemma 19 *(Lemma 4, Ferraty et al., 2007) Under assumptions of Lemma 17 we have that*

$$\mathbb{E}\left(\widehat{f}_h(\chi)\right) \longrightarrow M_{1\chi} \text{ and } \mathbb{E}\left(\widehat{g}_h(\chi)\right) \longrightarrow m(\chi)M_{1\chi} \text{ as } n \longrightarrow \infty.$$

Remark 20 *Actually, lemmas 17 and 19 above were established in Ferraty et al. (2007) under independence conditions. Noting that dependence does not influence on the expectation of $\widehat{g}_h(\chi)$ nor $\widehat{f}_h(\chi)$, one has that they remain valid in our setting of dependent samples. Of course, the validity of Lemma 2 in Ferraty et al. (2007) (Lemma 18 above) also remains because J_χ is not random.*

Proof. Because dependence does not influence on the expectation of $\widehat{g}_h(\chi)$ nor $\widehat{f}_h(\chi)$, the proof of this lemma is the same as that obtained for Lemma 4 in Ferraty et al. (2007) for the independent case. □

Lemma 21 *(Lemma 3, Aneiros and Vieu, 2008) Let V_k be a zero-mean, stationary, α-mixing and real process, such that for some $r > 4, \max_{1\leq k\leq n} E|V_k|^r \leq \mathcal{C} < \infty$. Assume that $a_{i,k}, i, k = 1, \ldots, n$ is a sequence of positive numbers such that $\max_{1\leq i,k\leq n} |a_{i,k}| = O(a_n)$. If, in addition, $\sum_{n=1}^{\infty} n^{\frac{5+4\gamma}{4(1-\gamma)}}\alpha(n) < \infty (0.5 < \gamma < 1)$, then:*

$$\max_{1\leq i\leq n} |\sum_{k=1}^{n} a_{i,k}V_k| = O(a_n n^{1/2+1/r}\log n) a.s.$$

Remark 22 *Last lemma remains unchanged when $a_{i,k}, i, k = 1, \ldots, n$ verifies the conditions almost sure. If the mixing coefficients verify $n^{\frac{5+4\gamma}{4(1-\gamma)}}\alpha(n) \to 0$ as $n \to \infty$ instead of $\sum_{n=1}^{\infty} n^{\frac{5+4\gamma}{4(1-\gamma)}}\alpha(n) < \infty (0.5 < \gamma < 1)$, we obtain the same result but with convergence in probability, that is: $\max_{1\leq i\leq n} |\sum_{k=1}^{n} a_{i,k}V_k| = O_P(a_n n^{1/2+1/r}\log n)$*

Lemma 23 *(Lemma 4, Aneiros and Vieu, 2008) Under assumptions (5.2)-(5.5) if, in addition, $\boldsymbol{\chi}_i$ are identically distributed and come from some $\alpha-$mixing process whose mixing coefficients $\alpha(n)$ verify $\alpha(n) \leq cn^{-a^*}$ for some $a^* > 1$, and $s_{n,1}^{\frac{a^*+1}{2}} = o(n^{\omega})$ for some $\omega > 2$ and $\left(n^{2-\frac{2\omega}{a^*+1}}F(h)^2\right)^{-1}\log n = O(1)$, then:*

$$\max_{i,j} |w_b(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)| = O(\frac{1}{nF(b)})$$

Lemma 24 *(Lemma 5, Aneiros and Vieu, 2008) Under assumptions (5.2)–(5.6) and (5.11) (m not included in (5.4)), if in addition $h \to 0$, $\log n/(nF(h)) \to 0$ and $nF(h)^{\frac{\epsilon a(r-2)}{r}-1} = O(1)$ as $n \to \infty$ (where $a > 1$ and $r > 4$, and $0 \leq \epsilon \leq 1$ was defined in Assumption (5.6)), and*

1. *$\{(\boldsymbol{X}_i, \boldsymbol{\chi}_i)\}_{i=1}^n$ come from some stationary α-mixing process whose mixing coefficients are $\alpha(n) \leq cn^{-a}$,*
2. *$\max_{1\leq i\leq n}(\mathbb{E}|X_{i1}|^r + \ldots + \mathbb{E}|X_{ip}|^r) \leq C < \infty$, and*
3. *$(\sup_{t\in C}(s_{n,1}(t) + s_{n,3}(t)))^{\frac{r(a+1)}{2(a+r)}} = o(n^{\theta})$, for some $\theta > 2$,*

then we have that

$$n^{-1}\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} \to \boldsymbol{B} \ a.s.$$

Theorem 25 *(Theorem 1, Aneiros and Vieu, 2008). Under assumptions (5.2)-(5.14),*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}) \longrightarrow^D N(0, \mathbf{A}) \textit{where } \mathbf{A} = \mathbf{B}^{-1}\mathbf{C}\mathbf{B}^{-1}. \tag{A.1}$$

If in addition the sample is strictly stationary, then

$$\limsup_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{\frac{1}{2}} |\widehat{\boldsymbol{\beta}}_{hj} - \boldsymbol{\beta}_j| = (a_{jj})^{\frac{1}{2}} \ a.s., \tag{A.2}$$

where $a_{jj} = (\mathbf{A})_{jj}$

Theorem 26 *Polya's theorem (see, e.g., Serfling, 1980, p. 18).*
If $F_n \longrightarrow F$ *and* F *is continuous, then* $\lim_{n\leftarrow\infty} \sup_t |F_n(t) - F(t)| = 0$.

Lemma 27 *(Lemma 2.5, Delsol, 2009) Under assumptions (4.4), (4.5), (4.6), (4.7), (4.8), (4.9), (4.10) and (4.12) we have that*

$$Var\left(\widehat{f}_h(\chi)\right) = \frac{M_{2\chi}}{nF_\chi(h)}(1 + o(1)),$$

$$Var\left(\widehat{g}_h(\chi)\right) = (\sigma_\varepsilon^2 + m^2(\chi))\frac{M_{2\chi}}{nF_\chi(h)}(1 + o(1))$$

and

$$Cov\left(\widehat{g}_h(\chi), \widehat{f}_h(\chi)\right) = m(\chi)\frac{M_{2\chi}}{nF_\chi(h)}(1 + o(1)).$$

Theorem 28 *(Theorem 2.7, Delsol, 2009). Under assumptions (4.4)-(4.13), we have:*

$$\frac{M_1}{\sqrt{M_2\sigma_\varepsilon^2}}\sqrt{nF(h)}(\widehat{m}_h(\chi) - m(\chi) - B_n) \to N(0, 1)$$

(Note that $\widehat{m}_h(\chi)$ *is defined in (4.3).*

Appendix B

Resumen en castellano

La memoria de tesis presentada se centra, principalmente, en el estudio de la predicción de demanda y precio de la electricidad, dentro del Mercado Eléctrico Español, desde el punto de vista de la estadística. En concreto, se utilizan técnicas de Análisis de Datos Funcionales para este problema.

La energía es un producto no almacenable, por lo que es de vital importancia anticipar las decisiones y evitar medidas erróneas de cara a maximizar los beneficios y reducir los costes para los agentes involucrados en el mercado. Los datos eléctricos, tanto la demanda como el precio, tienen algunas peculiaridades que dificultan su análisis. Se han hecho numerosas propuestas enfocadas a trabajar con este tipo de datos, muchas de ellas desde el ámbito de la ingeniería, mientras que en esta memoria se aborda este problema desde la estadística.

La memoria se organiza como sigue: el primer capítulo introduce el funcionamiento del Mercado Eléctrico Español, del que proceden los datos de demanda y precio de la electricidad, y los datos funcionales. A continuación se realiza un análisis descriptivo de la base de datos eléctricos que se usará a lo largo de toda la memoria. El Capítulo 2 se centra en el problema de detección de atípicos en series de tiempo funcionales, proponiendo nuevos procedimientos que se analizan mediante estudios de simulación y aplicación a los datos eléctricos. El tercer capítulo contiene un estudio comparativo de métodos de predicción para los datos eléctricos, proponiendo el uso de modelos de regresión funcional y predicciones combinadas. En los capítulos 4 y 5 se proponen procedimientos bootstrap para el cálculo de intervalos de predicción en el modelo de regresión funcional no paramétrico (FNP) y en el modelo semi-funcional parcialmente lineal (SFPL), respectivamente, en el contexto de datos dependientes. En ambos casos se prueba teóricamente la

validez de los procedimientos propuestos y se aplican tanto a datos simulados como a los datos de demanda y precio de la electricidad. Finalmente, el Capítulo 6 extiende los procedimientos bootstrap indicados al cálculo de intervalos de predicción en ambos modelos de regresión funcional, aplicándolos a los datos eléctricos.

Capítulo 1: Introducción a los datos eléctricos

El primer capítulo de la memoria sirve de introducción al contexto en el que se desenvuelve la tesis: los datos eléctricos. En primer lugar se explica el funcionamiento y la composición del Mercado Eléctrico Español, que consta de dos operadores: el operador del mercado (OMIE: Operador del Mercado Ibérico de la Electricidad) y el operador del sistema (REE: Red Eléctrica de España). Los datos de demanda y precio que se utilizan en esta memoria son datos horarios que proceden del operador del mercado y, dentro de él, del mercado diario de la electricidad. Una vez introducido el ámbito de aplicación de la tesis se introduce brevemente el campo de la estadística en el que se centra en análisis: los datos funcionales.

Se lleva a cabo un análisis descriptivo de los datos de demanda y precio de la electricidad a lo largo del año 2012. Sus principales características son la estacionalidad diaria y semanal, su diferente comportamiento en los fines de semana con respecto a los diarios, la presencia de atípicos, la influencia con efecto no lineal de variables meteorológicas y, en el caso particular del precio, el registro de precio cero en algunos momentos a lo largo del año.

Se analiza, utilizando numerosos gráficos, el comportamiento de la demanda y el precio según el día de la semana, según el mes del año, según el trimestre, por días festivos, etc. A la vista de este análisis se concluye que, tanto la demanda como el precio, se comportan de distinta forma en diarios, sábados y domingos. Esta clasificación por tipos de días se tendrá en cuenta a lo largo de toda la memoria a la hora de aplicar cualquier procedimiento estadístico sobre estos datos. Se analizan también algunos datos auxiliares que formarán parte de los métodos de predicción: la temperatura y la producción de energía eólica.

Capítulo 2: Detección de atípicos en series de tiempo funcionales

La detección de atípicos es uno de los primeros puntos a tener en cuenta dentro de cualquier procedimiento estadístico, para evitar que perturben los análisis que se puedan realizar sobre una base de datos. Dentro del contexto

de datos funcionales se pueden distinguir dos tipos principales de atípicos: los atípicos de magnitud y los atípicos de forma. Febrero et al. (2008) fueron los primeros autores en proponer un procedimiento específicamente diseñado para la detección de atípicos funcionales.

Los atípicos de magnitud, cuando uno trabaja con series de tiempo funcionales, pueden permanecer ocultos por la propia dinámica de la serie de tiempo. Debido a esto, es necesario tener en cuenta la dependencia de los datos en este tipo de estructuras, para poder detectar también este tipo de atípicos "escondidos".

Se proponen tres procedimientos para detectar atípicos en series de tiempo funcionales. El primero de ellos se basa en el método presentado por Febrero et al. (2008) que utiliza procedimientos bootstrap y que trabaja con profundidades funcionales. Nuestra propuesta adapta este procedimiento para trabajar con series de tiempo funcionales, entre otras modificaciones, mediante el uso del bootstrap para datos dependientes. Esta primera propuesta se puede encontrar en Raña, Aneiros y Vilar (2015), junto con un estudio de simulación y aplicación a datos reales.

A continuación se proponen otros dos procedimientos que se basan, en esta ocasión, en la descomposición en componentes principales funcionales robusta de Hyndman y Ullah (2007) y que se pueden encontrar en Vilar, Raña and Aneiros (2016).

Finalmente, se aplican los procedimientos de detección de atípicos a los datos de demanda y precio de la electricidad, comparando los métodos adaptados a series de tiempo funcionales con otros que no tienen en cuenta la dependencia de los datos.

Capítulo 3: Predicción de demanda y precio de la electricidad

En este capítulo se realiza un estudio comparativo de diferentes métodos de predicción aplicados a la demanda y precio de la electricidad en España. En primer lugar se consideran algunos métodos clásicos de predicción en este ámbito, como el método Naive y el uso de modelos ARIMA para series de tiempo univariantes y se considera el modelo de predicción de series de tiempo funcionales basado en componentes principales Hyndman y Ullah (2007). Se propone el uso de modelos de regresión funcional en este ámbito, extendiendo el trabajo desarrollado en Vilar et al. (2012) incorporando modelos con respuesta funcional, añadiendo covariables externas y obteniendo también

predicciones combinadas.

En concreto, se trabaja con el modelo de regresión FNP y con el modelo SFPL. En ambos casos se considera variable respuesta escalar y variable explicativa funcional no paramétrica, añadiendo en el segundo de los modelos variables explicativas escalares con efecto lineal sobre la respuesta (información de la temperatura para predecir la demanda eléctrica e información de la propia demanda y de la producción de energía eólica para predecir el precio).

Se proponen finalmente dos modelos de predicción combinados que se basan en promedios de los métodos de predicción considerados, lo que permite compensar las predicciones por defecto y por exceso, logrando resultados más ajustados. La primera de las combinaciones propuestas se basa, simplemente, en el promedio por tipo de día de los métodos con respuesta funcional considerados en el estudio. La segunda combinación, algo más sofisticada pero igualmente sencilla, se basa en promediar, dentro de cada grupo de días, los dos métodos de predicción con respuesta funcional que obtienen mejores errores de predicción.

Los errores de predicción obtenidos tanto en el caso de la demanda como del precio de la electricidad indican un buen funcionamiento de los modelos de regresión funcional, sobre todo en el caso del modelo SFPL que incorpora información externa. En cualquier caso, los métodos combinados suponen una mejora con respecto a cualquiera de los otros métodos considerados a nivel individual. El análisis desarrollado en este capítulo se puede encontrar en Aneiros, Vilar y Raña (2016).

Capítulo 4: Intervalos de confianza en el modelo de regresión funcional no paramétrico

Se proponen procedimientos bootstrap para el cálculo de intervalos de confianza en el modelo de regresión FNP, en el que se considera respuesta escalar y variable explicativa funcional, dentro del contexto de datos dependientes. Este capítulo supone la extensión del trabajo presentado por Ferraty, Van Keilegom y Vieu (2010), que propone los mismos procedimientos bootstrap en el modelo FNP con datos independientes, al contexto de datos dependientes.

Los procedimientos propuestos siguen, por lo tanto, las ideas planteadas en el trabajo de Ferraty, Van Keilegom y Vieu (2010), utilizando los resultados de Delsol (2009) que proporcionan la distribución asintótica del estimador

de la regresión para el mismo modelo de regresión FNP considerado y bajo el mismo escenario de dependencia, trabajando con series de tiempo funcionales.

Se plantean los dos algoritmos bootstrap propuestos, basados en el remuestreo de los residuos: naive bootstrap para modelos homocedásticos y wild bootstrap en el caso de heterocedasticidad. A continuación se prueba teóricamente la validez de los procedimientos bootstrap planteados. Para ello se establecen dos teoremas, uno para cada tipo de boostrap, planteando previamente una serie de hipótesis necesarias. Dichas hipótesis han sido planteadas previamente en los dos trabajos en los que se basa este capítulo: Ferraty, Van Keilegom y Vieu (2010) y Delsol (2009). Se desarrollan a continuación sus demostraciones en detalle.

Se estudia el comportamiento de los procedimientos planteados para el cálculo de intervalos de confianza mediante un estudio de simulación y se aplica al conjunto de datos eléctricos, considerando la demanda y precio de la electricidad.

En primer lugar, se desarrolla un estudio de simulación en el que se utilizan y comparan dos métodos de simulación de series de tiempo funcionales, un primer modelo que genera curvas suaves como variable explicativa funcional para el modelo de regresión FNP y un segundo modelo con curvas más rugosas e irregulares. En ambos casos se construyen intervalos de confianza para la media condicionada de tres tipos: teórico, bootstrap y asintótico.

El procedimiento para el cálculo de los intervalos de confianza teóricos utiliza réplicas de Monte Carlo, asumiendo que la función de regresión del modelo considerado es conocida. Los intervalos de confianza bootstrap se basan en aproximar la distribución del estimador de la función de regresión mediante su correspondiente estimador boostrap, utilizando el procedimiento wild bootstrap desarrollado de forma teórica previamente. Por último, dado que el estudio de Delsol (2009) proporciona constantes explícitas para la distribución asintótica del estimador de la regresión, se pueden calcular también intervalos de confianza asintóticos.

Se obtienen los tres tipos de intervalos de confianza, dentro del estudio de simulación, para una muestra test y se calcula el promedio de las coberturas empíricas en cada caso. Los resultados indican, en general, un buen funcionamiento de los procedimientos. En primer lugar, la cobertura de los intervalos de confianza teóricos se aproxima al nivel de confianza, lo que in-

dica que el estudio está bien calibrado. En segundo lugar, los intervalos de confianza bootstrap se aproximan a su vez a los intervalos teóricos, quedando siempre por delante de los intervalos de confianza asintóticos que, aun así, obtienen resultados muy razonables. En tercer lugar, se observa una mejora en las coberturas a medida que el tamaño muestral considerado para realizar las estimaciones del modelo aumenta. Por último, como es de esperar, los intervalos de confianza obtenidos funcionan mejor cuando se considera el modelo que simula curvas suaves, mientras que las coberturas son algo más bajas cuando se consideran curvas más irregulares. En ambos modelos se obtienen coberturas puntualmente bajas que se corresponden, al analizar esos casos en detalle, con curvas situadas en los extremos en las que es más difícil conseguir buenas estimaciones.

Finalmente, el capítulo concluye con el cálculo de intervalos de confianza bootstrap para datos reales, considerando la demanda y precio de la electricidad. En ambos casos se muestran resultados para dos opciones: predecir las 24 horas de un día fijo o bien predecir una misma hora fija a lo largo de una serie de días consecutivos. El estudio llevado a cabo en este capítulo, tanto la parte teórica como el estudio de simulación y las aplicaciones a datos eléctricos se puede encontrar en Raña, Aneiros, Vilar y Vieu (2016).

Capítulo 5: Intervalos de confianza en el modelo de regresión semi-funcional parcialmente lineal

Este capítulo extiende los procedimientos bootstrap planteados en el Capítulo 4 para el modelo de regresión FNP, al caso del modelo SFPL. La principal novedad radica en que, para este modelo parcialmente lineal, es necesario estimar en un primer momento la componente lineal del modelo y, a continuación, la componente no paramétrica.

A diferencia del modelo FNP, en este caso hasta donde sabemos, no existe en la literatura estadística ningún estudio que aplique este tipo de procedimientos bootstrap sobre modelos parcialmente lineales, no solo en el contexto de datos funcionales y datos dependientes, sino incluso en un contexto más general. Los trabajos de Liang et al. (2000) y de You y Chen (2006) incluyen aproximaciones bootstrap para la componente lineal de un modelo de regresión parcialmente lineal bajo diseño fijo. Por lo tanto, la contribución principal de este capítulo se centra en el planteamiento y estudio de procedimientos bootstrap para el modelo SFPL bajo dependencia (y, como caso particular, para datos independientes) en el que se consideran las

dos componentes: lineal y no paramétrica.

Siguiendo los algoritmos planteados en el Capítulo 4, se proponen también para el modelo SFPL dos procedimientos bootstrap basados en el remuestreo de los residuos: naive y wild bootstrap. En ambos casos, siguiendo el proceso de estimación del modelo, se construye en primer lugar el estimador bootstrap de la componente lineal y, a continuación, el de la componente no paramétrica.

Se realizan los desarrollos teóricos necesarios para probar la validez de los procedimientos bootstrap planteados. Para ello, se toma como base el estudio realizado en el Capítulo 4 para el modelo FNP, junto con el trabajo de Aneiros y Vieu (2008) en el que se estudia la distribución asintótica de los estimadores de este mismo modelo de regresión considerando el mismo escenario de datos dependientes en el que se trabaja con series de tiempo funcionales. En base a las hipótesis planteadas en ambos trabajos, se establecen los cuatro teoremas que proporcionan de forma teórica la validez de los procedimientos bootstrap: dos para cada componente del modelo (lineal y no paramétrica) y dos para cada tipo de bootstrap (naive y wild bootstrap). Se incluyen las demostraciones de dichos teoremas en detalle.

Se aplican los procedimientos desarrollados para el cálculo de intervalos de confianza para este modelo SFPL utilizando tanto datos simulados como los datos de demanda y precio de la electricidad. En primer lugar se realiza el estudio de simulación. Los modelos de simulación de datos utilizados en el Capítulo 4 se usan de nuevo en este estudio, teniendo en cuenta que se añade la componente lineal del modelo mediante una nueva covariable escalar con efecto lineal.

Se replican los algoritmos para el cálculo de intervalos de confianza, teniendo en cuenta que en esta ocasión no es posible obtener intervalos de confianza asintóticos ya que la distribución asintótica exacta de los estimadores no es conocida. Además, por la configuración del propio modelo SFPL, se pueden obtener intervalos de confianza para cada una de las componentes del modelo por separado (para la componente lineal y no paramétrica) y de forma conjunta (para toda la función de regresión).

De nuevo, al analizar los resultados obtenidos, se ve que la cobertura de los intervalos teóricos se aproxima al nivel de confianza, indicando que el proceso está bien calibrado. Los intervalos de confianza bootstrap se aproximan también a los teóricos. En este caso, las coberturas son más bajas, en general,

que las obtenidas en el Capítulo 4 para el modelo FNP, debido a la mayor dificultad para obtener las estimaciones. Se conserva además la tendencia de tener coberturas más bajas cuando los datos simulados son más irregulares.

Por último, se obtienen los intervalos de confianza bootstrap para los datos de demanda eléctrica (incorporando como covariables escalares información de la temperatura) y para el precio (incorporando información de la propia demanda y de la producción de energía eólica). Se encuentra en preparación un artículo que contiene el análisis desarrollado en este capítulo, tanto de la parte teórica como del estudio de simulación y aplicaciones a datos eléctricos.

Capítulo 6: Intervalos de predicción en regresión funcional

Este último capítulo de la tesis extiende los procedimientos bootstrap planteados para el cálculo de intervalos de confianza de los capítulos 4 y 5 al cálculo de intervalos de predicción.

Los intervalos de predicción complementan a las predicciones puntuales, como las obtenidas en el Capítulo 3, ya que reflejan la variabilidad en las predicciones al considerar tanto el error debido a la estimación de la regresión (considerada en los intervalos de confianza) y también el error debido al modelo. Se obtiene también la densidad de predicción en base al mismo procedimiento utilizado para los intervalos de predicción.

Dentro del estudio de mercados eléctricos, como indica Weron (2006), se pueden encontrar numerosas referencias sobre predicción puntual de demanda y precio de la electricidad utilizando un amplio rango de metodologías. Sin embargo, no es fácil encontrar referencias en las que se calculen intervalos de predicción y, menos aún, que trabajen con la densidad de predicción. Por ello, en el este capítulo de va un paso más allá en el estudio de demanda y precio de la electricidad al proporcionar en conjunto tanto predicciones puntuales como intervalos y densidad de predicción, dentro del contexto de la regresión funcional.

Se consideran, por lo tanto, procedimientos bootstrap, basados en los resultados teóricos estudiados previamente, para el cálculo de intervalos de predicción para los dos modelos de regresión considerados: FNP y SFPL, tanto bajo homocedasticidad como heterocedasticidad. Para este último caso es necesario hacer algunas consideraciones adicionales, ya que no se puede extender de forma directa el wild bootstrap visto en los capítulos anteriores

(que se aplica para aproximar el error de las estimaciones) para aproximar el error del modelo. Se incluye, dentro de los algoritmos, la estimación de la varianza condicional del modelo siguiendo las ideas de Fan y Yao (1998).

Se construyen intervalos de predicción bootstrap, además de la densidad de predicción, para la demanda y precio de la electricidad, extendiendo las aplicaciones previas de intervalos de confianza vistas en los Capítulos 4 y 5. El análisis de densidad de predicción permite distinguir diferencias en el comportamiento de la demanda y del precio entre las distintas horas del día. Se encuentra en preparación un artículo que contiene el análisis desarrollado en este capítulo, tanto los algoritmos bootstrap como las aplicaciones a datos eléctricos.

Conclusiones

En resumen, a lo largo de esta memoria se aborda el estudio de la demanda y precio de la electricidad mediante el uso de técnicas de datos funcionales. En un primer momento se realiza un análisis descriptivo de los datos eléctricos, que componen una serie de tiempo funcional, conociendo en detalle su funcionamiento. Se proponen métodos de detección de atípicos diseñados para trabajar con series de tiempo funcionales. Se realiza un estudio comparativo de diferentes métodos de predicción aplicados a demanda y precio de la electricidad, proponiendo el uso de modelos de regresión funcional.

Se proponen procedimientos bootstrap para el cálculo de intervalos de confianza para los modelos de regresión FNP y SFPL, considerando respuesta escalar. En ambos casos se realizan los desarrollos teóricos necesarios para probar la validez de dichos procedimientos bootstrap, que se aplican en la práctica mediante un estudio de simulación y también para los datos eléctricos. Por último, se calculan intervalos de predicción, además de densidad de predicción, para demanda y precio de la electricidad, basándose en los procedimientos bootstrap estudiados previamente.

Bibliography

[1] Aneiros-Pérez, G., Cardot, H., Estévez-Perez, G. and Vieu, P. (2004), Maximum ozone concentration forecasting by functional non-parametric approaches, *Environmetrics*, 15: 675–685.

[2] Aneiros, G. and Vieu, P. (2008), Nonparametric time series prediction: A semi-functional partial linear modeling, *Journal of Multivariate Analysis*, 99: 834–857.

[3] Aneiros, G., Vilar, J.M., Cao, R. and Muñoz-San-Roque, A. (2013), Functional prediction for the residual demand in electricity spot markets, *IEEE Transactions on Power Systems*, 28: 4201–4208.

[4] Aneiros, G., Vilar, J. and Raña, P. (2016), Short-term forecast of daily curves of electricity demand and price, *International Journal of Electrical Power and Energy Systems*, 80: 96–108.

[5] Antoch, J., Prchal, L., De Rosa, M.R. and Sarda, P. (2010), Electricity consumption prediction with functional linear regression using spline estimators. *Journal of Applied Statistics*, 37: 2027–2041.

[6] Arribas-Gil, A. and Romo, J. (2014), Shape outlier detection and visualization for functional data: the outliergram, *Biostatistics*, 15: 603–619.

[7] Aue, A., Norinho, D.D. and Hörmann, S. (2015), On the prediction of stationary functional time series, *Journal of the American Statistical Association* 110(509): 378–392.

[8] Baíllo, A., Cuesta-Albertos, J.A. and Cuevas, A. (2011), Supervised classification for a family of Gaussian functional models, *Scandinavian Journal of Statistics*, 38: 480–498.

[9] Besse, P.C., Cardot, H. and Stephenson, D. (2000), Autoregressive forecasting of some functional climatic variations, *Scandinavian Journal of Statistics*, 27: 673–688.

[10] Boente, G. and Fraiman, R. (2000), Kernel-based functional principal components, *Statistics and Probability Letters*, 48: 335–345.

[11] Brockwell P.J. and Davis R.A. (1996), *Introduction to time series and forecasting (2nd ed.)* New York: Springer-Verlag.

[12] Conejo, J.R., Contreras, J., Espínola, R. and Plazas, M.A. (2005), Forecasting electricity prices for a day-ahead pool-based electric energy market, *International Journal of Forecasting*, 21: 435–462.

[13] Cancelo, J.R. and Espasa, A. (1996), Modelling and forecasting daily series of electricity demand, *Investigaciones Económicas*, XX(3): 359–376.

[14] Cancelo, J.R., Espasa, A. and Grafe, R. (2008), Forecasting the electricity load from one day to one week ahead for the Spanish system operator, *International Journal of Forecasting*, 24: 588–602.

[15] Cao, R. (1991), Rate of convergence for the wild bootstrap in nonparametric regression, *Annals of Statistics*, 19: 2226–2231.

[16] Cardot, H., Ferraty, F. and Sarda, P. (1999), Functional linear model, *Statistics and Probability Letters*, 45: 11–22.

[17] Chang, P., Fan, C. and Lin, J. (2011), Monthly electricity demand forecasting based on a weighted evolving fuzzy neural network approach, *International Journal of Electrical Power & Energy Systems*, 33: 17–27.

[18] Cho, H., Goude, Y., Brossat, X. and Yao, Q. (2013), Modeling and forecasting daily electricity load curves: a hybrid approach, *Journal of the American Statistical Association*, 108(501): 7–21.

[19] Contreras, J., Espínola, R., Nogales, F.J. and Conejo, A.J. (2003), ARIMA models to predict next-day electricity prices, *IEEE Transactions on power systems*, 18(3): 1014–1020.

[20] Cruz, A., Muñoz, A., Zamora, J.L. and Espínola, R. (2011), The effect of wind generation and weekday on Spanish electricity spot price forecasting, *Electric power systems research*, 81: 1924–1935.

[21] Cryer, J.D. and Chan, K.S. (2008), *Time Series Analysis*, New York: Springer.

[22] Cuevas, A. (2014), A partial overview of the theory of statistics with functional data, *Journal of Statistical Planning and Inference*, 147: 1-23.

[23] Cuevas, A., Febrero, M. and Fraiman, R. (2001), Cluster analysis: a further approach based on density estimation, *Computational Statistics & Data Analysis*, 36: 441–456.

[24] Cuevas, A., Febrero, M. and Fraiman, R. (2006), On the use of the bootstrap for estimating functions with functional data, *Computational Statistics & Data Analysis*, 51: 1063–1074.

[25] Cuevas, A., Febrero, M. and Fraiman, R. (2007), Robust estimation and classification for functional data via projection-based depth notions, *Computational Statistics*, 22: 481–496.

[26] Cuevas, A. and Fraiman, R. (1997), A plug–in approach to support estimation, *Annals of Statistics* 25: 2300–2312.

[27] Delsol, L. (2009), Advances on asymptotic normality in non-parametric functional time series analysis, *Statistics*, 43(1): 13–33.

[28] Delsol, L., Ferraty, F. and Vieu, P. (2011), Structural test in regression on functional variables, *Journal of Multivariate Analysis*, 102: 422–447.

[29] Fan, J. and Yao, Q. (1998), Efficient estimation of conditional variance functions in stochastic regression, *Biometrika*, 85(3): 645–660.

[30] Febrero, M., Galeano, P. and González-Manteiga, W. (2007), Functional analysis of NOx levels: location and scale estimation and outlier detection, *Computational Statistics*, 22: 411–427.

[31] Febrero, M., Galeano, P and González-Manteiga, W. (2008), Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels, *Environmetrics*, 19: 331–345.

[32] Febrero-Bande, M. and Oviedo de la Fuente, M. (2012), Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software*, 51(4), 1–28. URL http://www.jstatsoft.org/v51/i04/.

[33] Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2011), Kernel regression with functional response, *Electronic Journal of Statistics*, 5: 159–171.

[34] Ferraty, F., Mas, A. and Vieu, P. (2007), Nonparametric regression on functional data: inference and practical aspects, *Australian and New Zealand Journal of Statistics*, 49: 267–286.

[35] Ferraty, F., Van Keilegom, I. and Vieu, P. (2010), On the Validity of the Bootstrap in Non-Parametric Functional Regression, *Scandinavian Journal of Statistics*, 37: 286–306.

[36] Ferraty, F., van Keilegom, I. and Vieu, P. (2012), Regression when both response and predictor are functions, *Journal of Multivariate Analysis*, 109: 10–28.

[37] Ferraty, F. and Vieu, P. (2002), The functional nonparametric model and application to spectrometric data, *Computational Statistics*, 17: 545–564.

[38] Ferraty, F. and Vieu P. (2004), Nonparametric models for functional data, with application in regression, time–series prediction and curve discrimination, *Nonparameric Statistics*, 16: 111–125.

[39] Ferraty, F. and Vieu, P. (2006), *Nonparametric functional data analysis.* New York: Springer Series in Statistics. Springer-Verlag.

[40] Fraiman, R. and Svarc, M. (2013), Resistant estimates for high dimensional and functional data based on random projections, *Computational Statistics and Data Analysis*, 58: 326–338.

[41] Freedman, D.A. (1981), Bootstrapping regression models, *Annals of Statistics*, 9: 1218–1228.

[42] García-Martos, C., Rodríguez, J. and Sánchez, M.J. (2007), Mixed models for short-run forecasting of electricity prices: application for the Spanish market, *IEEE Transactions on power systems*, 22(2): 544–552.

[43] García-Portugués, E., González-Manteiga, W. and Febrero-Bande, M. (2014), A goodness-of-fit test for the functional linear model with scalar response, *Journal of Computational and Graphical Statistics*, 23: 761–778.

[44] Geidel, C. and Zareipour, H. (2013), Price forecasting in the Spanish day-ahead electricity market using preconditioned wind power information, *Proceedings 12th International Conference on Machine Learning and Applications. IEEE.* 2: 203–210.

[45] Gervini, D. (2012), Outlier detection and trimmed estimation for general functional data, *Statistica Sinica*, 22: 1639–1660.

[46] Goia, A., May, C. and Fusai, G. (2010), Functional clustering and linear regression for peak load forecasting, *International Journal of Forecasting*, 26: 700–711.

[47] González-Manteiga, W. and Martínez-Calvo, A. (2011), Bootstrap in functional linear regression, *Journal of Statistical Planning and Inference*, 141: 453–461.

[48] Hall, P. (1992), On bootstrap confidence intervals in nonparametric regression, *Annals of Statistics*, 20: 695–711.

[49] Hall, P., Müller, H.G. and Wang, J.L. (2006), Properties of principal component methods for functional and longitudinal data analysis, *Annals of Statistics*, 34: 1493–1517.

[50] Härdle, W. and Marron, J.S. (1991), Bootstrap simultaneous error bars for nonparametric regression, *Annals of Statistics*, 19: 778–796.

[51] Horváth, L. and Kokoszka, P. (2012), *Inference for Functional Data with Applications*, New York: Springer.

[52] Hsing, T. and Eubank, R. (2015), *Theoretical foundations of functional data analysis, with an introduction to linear operators*, Wiley and Sons: Chichester.

[53] Hubert, M., Rousseeuw, P.J. and Verboven, S. (2002), A fast method of robust principal components with applications to chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 60: 101–111.

[54] Hyde, O. and Hodnett, P.F. (2015), Modeling the effect of weather in short-term load forecasting, *Mathematical Engineering in Industry*, 6: 155–169.

[55] Hyndman, R.J. (1996), Computing and graphing highest density regions, *The American Statistician* 50: 120–126.

[56] Hyndman R. and Athanasopoulos G. (2013), *Forecasting: principles and practice.* Online at: http://otexts.org/fpp/.

[57] Hyndman, R.J. and Booth, H. (2008), Stochastic population forecasts using functional data models for mortality, fertility and migration, *International Journal of Forecasting*, 24: 323–342.

[58] Hyndman, R.J. and Shang, H.L. (2010), Rainbow plots, bagplots, and boxplots for functional data, *Journal of Computational and Graphical Statistics* 19: 29–45.

[59] Hyndman, R.J. and Ullah, M.S. (2007), Robust forecasting of mortality and fertility rates: A functional data approach, *Computational Statistics and Data Analysis* 51: 4942–4956.

[60] Kim, C.I., Yu, I.K and Song, Y.H. (2002), Prediction of system marginal price of electricity using wavelet transform analysis, *Energy Conversion and Management*, 43: 1839–1851.

[61] Künsch, H.R. (1989), The jacknife and the bootstrap for general stationary observations, *Annals of Statistics* 17: 1217–1241.

[62] Li, Y. and Hsing, T. (2007), On rates of convergence in functional linear regression, *Journal of Multivariate Analysis*, 98: 1782–1804.

[63] Liang, H., Hardle, W. and Sommerfeld, V. (2000), Bootstrap approximation in a partially linear regression model, *Journal of Statistical Planning and Inference*, 91: 413–426.

[64] Liebl, D. (2013), Modeling and forecasting electricity spot prices: a functional data perspective, *Annals of Applied Statistics*, 7: 1562–1592.

[65] López-Pintado, S. and Romo, J. (2009), On the concept of depth for functional data, *Journal of the American Statistical Association* 104: 718–734.

[66] Mammen, E. (1993), Bootstrap and wild bootstrap for high dimensional linear models, *Annals of Statistics*, 21: 255–285.

[67] Masry, E. (2005), Nonparametric regression estimation for dependent functional data: asymptotic normality, *Stochastic Processes and their Applications*, 115: 155–177.

[68] Misiorek, A., Trueck, S. and Weron, R. (2006), Point and interval forecasting on spot electricity prices: linear vs. non-linear time series models, *Studies in Nonlinear Dynamics & Econometrics*, 10(3), Article 2.

[69] Nogales, F.J., Contreras, J., Conejo, A.J. and Espínola, R. (2002), Forecasting next-day electricity prices by time series models, *IEEE Transactions on power systems*, 17(2): 342–348.

[70] Ocaña, F.A., Aguilera, A.M. and Escabias, M. (2007), Computational considerations in functional principal component analysis, *Computational Statistics*, 22: 449–465.

[71] Paparoditis, E. and Sapatinas, T. (2013), Short-term load forecasting: the similar shape functional time-series predictor, *IEEE Transactions on power systems*, 28(4): 3818–3825.

[72] Pardo, A., Meneu, V. and Valor, E. (2002), Temperature and seasonality influences on Spanish electricity load, *Energy Economics*, 24: 55–70.

[73] Politis, D.N. and Romano, J.P. (1994), The stationary bootstrap, *Journal of the American Statistical Association* 89: 1303–1313.

[74] Rachdi, M. and Vieu, P. (2007), Nonparametric regression for functional data: Automatic smoothing parameter selection, *Journal of Statistical Planning and Inference* 137: 2784–2801.

[75] Raña, P., Aneiros, G. and Vilar, J.M. (2015), Detection of outliers in functional time series, *Environmetrics*, 26: 178–191.

[76] Raña, P., Aneiros, G., Vilar, J. and Vieu, P. (2016) Bootstrap confidence intervals in functional nonparametric regression under dependence. *Electronic Journal of Statistics*, 10(2): 1973–1999.

[77] Raña, P., Aneiros, G., Vilar, J. and Vieu, P. Bootstrap confidence intervals in semi-functional partial linear regression. *(Preprint).*

[78] Sawant, P., Billor, N. and Shin, H. (2012), Functional outlier detection with robust functional principal component analysis, *Computational Statistics* 27: 83–102.

[79] Shang, H.L. (2014), Bayesian bandwidth estimation for a functional nonparametric regression model with mixed types of regressors and unknown error density, *Journal of Nonparametric Statistics*, 26: 599–615.

[80] Singhal, D. and Swarup, K.S. (2011), Electricity price forecasting using artificial neural networks, *International Journal of Electrical Power & Energy Systems*, 33: 550–555.

[81] Sguera, C., Galeano, P. and Lillo, R. (2014), Spatial depth-based classification for functional data, *Test*, 23: 725–750.

[82] Shumway R.H. and Stoffer D.S. (2006), *Time series analysis and its applications (2nd ed.)* Springer.

[83] Suganthi, L. and Samuel, A.A. (2012), Energy models for demand forecasting - A review, *Renewable & Sustainable Energy Reviews*, 16: 1223–1240.

[84] Sun, Y. and Genton, M.G. (2011), Functional Boxplot, *Journal of Computational and Graphical Statistics* 20: 316–334.

[85] Sun, Y. and Genton, G. (2012), Adjusted functional boxplots for spatio-temporal data visualization and outlier detection, *Environmetrics* 23: 54–64.

[86] Taylor, J.W. (2010), Triple seasonal methods for short term electricity demand forecasting, *European Journal of Operational Research*, 204: 139–152.

[87] Taylor, J.W. and Buizza, R. (2003), Using weather ensamble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19: 57–70.

[88] Taylor, J.W., de Menezes, L. and McSharry, P.E. (2006), A comparison of univariate methods for forecasting electricity demand up to a day ahead, *International Journal of Forecasting*, 22(1): 1–16.

[89] Taylor, J.W. and Majithia, S. (2000), Using combined forecasts with ganging weights for electricity demand profiling, *Journal of the Operational Research Society*, 51: 72–82.

[90] Timmermann, A.G. (2006), *Forecasts combinations.* In Handbook of Economic Forecasting: Elsevier.

[91] Tukey, J.W. (1970), *Exploratory Data Analysis (limited preliminary edition)*, Vol. 1. Addison-Wesley: Reading, MA.

[92] Tukey, J.W. (1977), *Exploratory Data Analysis.* Addison-Wesley: Reading, MA.

[93] Vilar, J.M., Cao, R. and Aneiros, G. (2012), Forecasting next-day electricity demand and price using nonparametric functional methods, *International Journal of Electrical Power & Energy Systems*, 39: 48–55.

[94] Vilar, J.M., Raña, P. and Aneiros, G. (2016), Using robust FPCA to identify outliers in functional time series, with applications to the electricity market. (To appear in Statistics and Operations Research Transactions).

[95] Wallis, K.F. (2011), Combining forecasts: forty years later, *Applied Financial Economics*, 21: 33–41.

[96] Wand, M.P. and Jones, M.C. (1995), *Kernel smoothing*. London: Monographs on Statistics and Applied Probability. Chapman & Hall.

[97] Weron, R. (2006), *Modeling and forecasting electricity loads and prices. A statistical approach*. Wiley.

[98] Weron, R. (2014), Electricity price forecasting: a review of the state-of-the-art with a look into the future, *International Journal of Forecasting*, 30: 1030–1081.

[99] Weron, R and Misiorek, A. (2008), Forecasting spot electricity prices: a comparison of parametric and semiparametric time series models, *International Journal of Forecasting*, 24: 744–763.

[100] You, J. and Chen, G. (2006), Wild bootstrap estimation in partially linear models with heteroscedasticity, *Statistics & Probability Letters*, 76: 340–348.

[101] Yu, G., Zou, C. and Wang, Z. (2012), Outlier detection in functional observations with applications to profile monitoring, *Technometrics*, 54: 308–318.

[102] Zhao, J.H., Dong, Z.Y., Xu, Z. and Wong, K.P. (2008), A statistical approach for interval forecasting of the electricity price, *IEEE Transactions on Power Systems*, 23(2): 267–276.