UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXIV

Conformal Prediction Bands for Functional Data

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Simone Vantini

Co-supervisore: Dott. Matteo Fontana

Dottorando: Jacopo Diquigiovanni

12th January 2022

Abstract

Functional Data Analysis represents a field of growing interest in statistics. Biomedicine, demography, public health, finance and environmental science are only a few examples of fields that can greatly benefit from innovative functional data analysis techniques. Within the broad range of challenges involving functional data, uncertainty quantification in prediction represents a topic of great importance both from a methodological and an application point of view. The thesis focuses on the development of methods to create prediction sets, namely subsets of the sample space that include the new random functional object we aim to predict with a certain nominal confidence level. The thesis deals with three scenarios of increasing complexity, each time providing the presentation of the methodologies used, ad-hoc simulation studies and a case study characterized by a strong application interest.

Sommario

L'analisi dei dati funzionali rappresenta un settore di sempre maggiore interesse in statistica. Biomedicina, demografia, salute pubblica, finanza e scienze ambientali sono solo alcuni dei campi che possono largamente beneficiare dallo sviluppo di tecniche innovative per l'analisi dei dati funzionali. All'interno della vasta gamma di sfide che coinvolgono i dati funzionali, la quantificazione dell'incertezza nel contesto predittivo rappresenta un argomento assai rilevante sia dal punto di vista metodologico che da quello applicativo. L'obiettivo della tesi è lo sviluppo di metodi per la creazione di insiemi predittivi, ossia sottoinsiemi dello spazio campionario capaci di contenere l'oggetto funzionale aleatorio che si vuole prevedere con un dato livello di confidenza nominale. La tesi tratta tre scenari di complessità crescente, presentando di volta in volta la metodologia utilizzata, proponendo studi di simulazione specifici e affrontando un caso studio dal forte interesse applicativo.

Ringraziamenti

Tra le tante persone che mi sono state accanto in questi anni mi sembra doveroso ringraziarne alcune.

A Simone, per essere stato una guida scientifica sempre disponibile con me. Mi hai dato fiducia fin da subito nonostante la scarsa conoscenza reciproca e questo ha significato molto per me.

A Matteo, sei stato un co-supervisore attento e i tuoi suggerimenti e consigli mi hanno aiutato molto.

Ai miei compagni di Padova e Milano, per tutto il tempo (troppo poco) passato insieme.

Ai miei genitori e a mio fratello Davide, mi avete dato tutto e per voi le parole non basteranno mai. I confronti non vanno mai bene, ma sono stato davvero fortunato.

Ad Anna, per aver tifato, lottato, rincuorato, gioito dalla prima all'ultima curva, con il sole che spacca e la pioggia battente. E' inebriante raggiungere questo risultato con te accanto, e lo è ancora di più sapendo di essere amato mai per ciò che raggiungo od ottengo, ma unicamente per ciò che sono. E' una sensazione che auguro a chiunque.

Contents

List of Figures

List of Tables

Introduction

Overview

Functional Data Analysis is the broad field of statistics whose purpose is to study sets of curves (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). Despite the intrinsic issues of the framework (e.g. the fact that a probability density function generally does not exist for random functions Delaigle *et al.*, 2010), Functional Data Analysis has focused on a wide range of topics, such as classification, linear regression, functional depth and nonparametric techniques (Goia and Vieu, 2016; Aneiros *et al.*, 2019).

This thesis deals with uncertainty quantification in the prediction of functional data. Within this broad field of research, the interest is in the generation of prediction sets, namely subsets of the sample space that include a new random functional object with a certain nominal confidence level. Intuitively, the main goal is to obtain either exact - i.e. ensuring a coverage equal to the nominal confidence level - or at least valid - i.e. ensuring a coverage no less than the nominal confidence level - prediction sets.

To do that, we build on top of Conformal Prediction (CP, Vovk *et al.*, 2005; Shafer and Vovk, 2008), a novel method of forecasting firstly developed in the Machine Learning community as a way to define prediction intervals for Support Vector Machines (Gammerman *et al.*, 1998). The interested reader can find a recent review in Fontana *et al.* (2022). In the non-functional setting, Conformal Prediction is able to generate distribution-free, valid/exact prediction intervals, while it has also been used as a data exploration tool for Functional Data (Lei *et al.*, 2015) via the use of a truncated basis approach.

In addition to being valid/exact, the prediction sets we will focus on are characterized by a specific shape. In the classical multivariate non-functional statistical setting, elliptic regions have been and are still considered as the standard shapes for prediction sets. Differently, in the functional context many authors (López-Pintado and Romo, 2009; Lei *et al.*, 2015) note how the focus should be on a particular type of prediction set, commonly known as *prediction band*. In order to introduce it, we will consider the

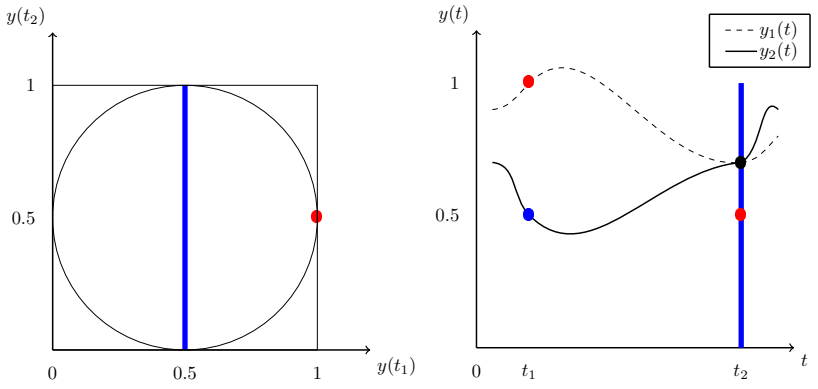

FIGURE 1: Example of importance of obtaining prediction bands.

simplest case of univariate functional data, but the same arguments can be generalized to hold in the multivariate functional framework, that will be addressed starting from Chapter 3. Formally, a band is defined as

$$\{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in B(t), \quad \forall t \in \mathcal{T}\},$$

with $y : \mathcal{T} \to \mathbb{R}$, $\mathcal{Y}(\mathcal{T})$ the space in which the random function Y takes value and $B(t) \subseteq \mathbb{R}$ interval for each $t \in \mathcal{T}$ (López-Pintado and Romo, 2009; Degras, 2017). The focus on this type of sets, that can be defined as the Cartesian product of the (infinitely many) intervals $\{B(t) : t \in \mathcal{T}\}$, comes from the fact that – differently from a generic region of $\mathcal{Y}(\mathcal{T})$ – such a shape can be easily visualized on a plot (i.e., it is a band, in parallel coordinates, as noted by López-Pintado and Romo, 2009) and thus interpreted with respect to the domain $\mathcal{T}$. In order to clarify this concept, let us consider the following example. Let $\mathcal{C}^1$ be a prediction band and let us consider the simple case in which $B(t)$ is the interval $[0, 1]$ for each $t \in \mathcal{T}$: in doing so, from a geometrical point of view $\mathcal{C}^1$ is an infinite-dimensional hypercube. Specifically, let us focus on two points of the domain, t_1 and t_2 respectively, and with a slight abuse of notation let us indicate with $\mathcal{C}^1(t_1, t_2) := \{(y(t_1), y(t_2)) : (y(t_1), y(t_2)) \in [0, 1] \times [0, 1]\}$ the "restriction" of prediction band related to $\{t_1, t_2\}$. In addition, let $\mathcal{C}^2$ be for example a different hypothetical prediction set having the shape of an infinite-dimensional hyper-sphere such that for instance $\mathcal{C}^2(t_1, t_2) := \{(y(t_1), y(t_2)) : (y(t_1) - 0.5)^2 + (y(t_2) - 0.5)^2 \leq 0.5^2\}$, i.e. $\mathcal{C}^2(t_1, t_2)$ is the closed disk of center (0.5,0.5) and radius 0.5. Both $\mathcal{C}^1(t_1, t_2)$ and $\mathcal{C}^2(t_1, t_2)$ are plotted on the left side of Figure 1. Drawing conclusions only on the basis of the behavior of the plotted functions in t_1 and t_2 and ignoring it in all the other points of the domain, the right side of Figure 1 shows a function that does not belong to $\mathcal{C}^2$ (the dashed curve y_1) and a function that belongs to such set (the solid curve

y_2): indeed, conditional on the fact that $y_1(t_1) = 1$, the dashed curve y_1 must satisfy $y_1(t_2) = 0.5$ to be included in $\mathcal{C}^2$, as shown by the red dot on the left of Figure 1. Conversely, conditional on the fact that $y_2(t_1) = 0.5$, $y_2(t_2)$ can assume whatever value between 0 and 1 to be included in $\mathcal{C}^2$, as shown by the blue solid vertical line on the left of Figure 1. The fact that the point where y_1 and y_2 intersect (the black dot on the right of Figure 1) determines whether to include or not a function in $\mathcal{C}^2$ on the basis of the value assumed by that function in t_1 represents an undeniable limit to the visualization of prediction sets, especially considering that this phenomenon involves all $t \in \mathcal{T}$. Fortunately, this problem is completely avoided by prediction sets as $\mathcal{C}^1$, and more generally by every prediction band: indeed, differently from prediction sets characterized by other shapes, prediction bands always coincide with (and are not only a subset of) their envelope.

In view of this, the thesis focuses on the development of methods that necessarily output valid/exact prediction bands - instead of more general prediction sets - in three scenarios of increasing complexity: in case of univariate i.i.d. functional data, in a regression framework with multivariate functional response variable and in a multivariate functional time series framework.

Main contributions of the thesis

The contributions of this thesis can be summarized as follows:

- in Chapter 1 functional prediction sets are formally defined and the Semi-Off-Line Inductive Conformal framework is introduced. Specifically, we contribute in two ways to the Conformal Prediction literature: via enriching the results about the validity of Split Conformal prediction sets by making the exact probability reached by them explicit (Theorem 1.1) and we provide what is to the best of our knowledge the first formal proof of the exactness of Smoothed Split Conformal prediction sets (Appendix A.2).

- in Chapter 2 we propose a new family of nonconformity measures inducing Conformal predictors able to create closed-form finite-sample valid or exact prediction bands for i.i.d. univariate functional data under very minimal distributional assumptions. The procedure is also fast, scalable, does not rely on functional dimension reduction techniques and allows the user to select different nonconformity measures depending on the problem at hand always obtaining valid/exact bands. Within this family of measures, we propose also a specific measure that guarantees an asymptotic result in terms of efficiency.

- in Chapter 3 we extend the results obtained in the previous chapter to multivariate functional data and to a regression framework. Under the mild assumption of exchangeable regression pairs, the procedure outputs closed-form finite-sample either valid or exact multivariate simultaneous prediction bands for multivariate functional response variable. The fact that the prediction bands modulate their width based on the local behavior and magnitude of the functional data and that they can be built around any regression estimator yields a very widely usable method, which is fairly straightforward to implement. The method is used to study the usage of a bike-sharing system in the Italian city of Milan in order to identify the periods of time in which the imbalance between the picked up bikes and the dropped off bikes could become critical based on some external covariates.

- in Chapter 4 we propose a scalable procedure that outputs closed-form simultaneous prediction bands for multivariate functional response variable in a time series setting. The time dependence does not allow to obtain the aforementioned finite-sample validity/exactness, but the method is still able to guarantee performance bounds in terms of coverage and asymptotic exactness, both under some conditions. After evaluating its performance on synthetic data, the method is applied to build multivariate prediction bands for daily demand and offer curves in the Italian gas market. The prediction framework thus obtained allows traders to directly evaluate the impact of their own offers/bids on the market, providing an intriguing tool for the business practice.

In order to allow the interested reader to separately understand the content of the three topics covered by the thesis, we introduce the specific problem and the related bibliography - as well as the notation used - for each of the Chapters 2, 3, 4. For this reason, the partial overlapping of contents between the chapters is aimed at creating self-contained sections.

Chapter 1

Conformal Prediction

Let us consider a framework in which we are interested in quantifying the uncertainty in predicting the new (possibly multivariate) functional response variable $\boldsymbol{Y}_{n+1}$ on the basis of the new set of covariates x_{n+1}, the observed regression data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ - which are drawn from independent and identically distributed regression pairs $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \sim P$, $\boldsymbol{Z}_i = (X_i, \boldsymbol{Y}_i)$, $i \in \{1, \ldots, n\}$ - and a given regression estimator, and we want to achieve this goal even when there are only a few regression data (i.e., n is small) and/or the regression estimator is poor. While in the traditional regression framework in which the response variable is scalar this typically means to find an interval, in the functional context we are considering it means to determine a subset of the sample space to which the new response variable will hopefully belong. The setting considered here is a general regression framework, but all definitions and results still hold in the case of i.i.d. functional data.

The tool we use to develop our prediction sets is *Conformal Prediction*, a nonparametric approach proposed in the multivariate (non-functional) literature for the first time by Gammerman *et al.* (1998) and thoroughly described in Vovk *et al.* (2005), that builds either valid or exact prediction sets under no assumptions other than exchangeable data (see also Fontana *et al.*, 2022, for a presentation of the topic more oriented to a statistical audience). Moreover, the CP framework ensures that valid/exact prediction sets are obtained regardless of the sample size n (i.e., not only asymptotically), a fact that allows Conformal Prediction to be used in an extremely wide range of different scenarios. Specifically, we will focus on the Semi-Off-Line Inductive Conformal framework, also known simply as Split Conformal (Papadopoulos *et al.*, 2002), a computationally and methodologically convenient alternative to the original Transductive Conformal method. Split Conformal approach is characterized by two sub-frameworks: Non-Smoothed Split Conformal framework and Smoothed Split Conformal framework.

Since the term 'Split Conformal' itself is used to indicate 'Non-Smoothed Split Conformal', later in the discussion we will use the following two terms to indicate the two sub-frameworks: Split Conformal, Smoothed Split Conformal. The two procedures are defined in the next two sections.

1.1 Split Conformal

Consistently with the notation of Lei *et al.* (2018), a valid prediction set for $\boldsymbol{Z}_{n+1} = (X_{n+1}, \boldsymbol{Y}_{n+1})$ — which is independent from and identically distributed to $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ — is the set $\mathcal{C}_{n,1-\alpha}$ based on $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ such that

$$\mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}\left(X_{n+1}\right)\right) \geq 1 - \alpha \tag{1.1}$$

for any significance level $\alpha \in (0, 1)$, with $\mathcal{C}_{n,1-\alpha}(x) = \{\boldsymbol{y} \in \mathcal{Y}(\mathcal{T}) : (x, \boldsymbol{y}) \in \mathcal{C}_{n,1-\alpha}\}$ and $\mathcal{Y}(\mathcal{T})$ the space in which $\boldsymbol{Y}_i$ takes values (which will be formally defined in each chapter on a case-by-case basis). It is possible to notice that the left side of Inequality (1.1) refers to the unconditional coverage reached by the prediction set, i.e., the probability is taken over the i.i.d. draws $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n+1}$. In order to avoid ambiguity, later in the discussion the term *coverage* (or *unconditional coverage*) will be used to refer to $\mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}\left(X_{n+1}\right)\right)$, the term *conditional coverage*[1] will be used to refer to $\mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}\left(X_{n+1}\right) | \mathcal{C}_{n,1-\alpha}\left(X_{n+1}\right)\right)$ and the terms *empirical coverage* and *empirical conditional coverage* will be used to refer to the estimate - from simulated data - of the coverage and conditional coverage respectively.

In order to present the Split Conformal approach, let us consider the following procedure: given data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, let $\{1, \ldots, n\}$ be randomly divided into two sets $\mathcal{I}_1, \mathcal{I}_2$ and let us define the training set as $\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}$ and the calibration set as $\{\boldsymbol{z}_d : d \in \mathcal{I}_2\}$, with $|\mathcal{I}_1| = m$, $|\mathcal{I}_2| = l$ and $m, l \in \mathbb{N}_{>0}$ such that $n = m + l$. Let us also define *non-conformity measure* as any measurable function $A(\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}, \boldsymbol{z})$ taking values in $\bar{\mathbb{R}}$, which is the set of the affinely extended real numbers. The Split Conformal prediction set for $\boldsymbol{Y}_{n+1}$ is defined as

$$\mathcal{C}_{n,1-\alpha}\left(x_{n+1}\right) := \{\boldsymbol{y} \in \mathcal{Y}(\mathcal{T}) : \delta_{\boldsymbol{y}} > \alpha\},$$

[1] It is important to notice that the definition of conditional coverage we provided is non-standard since typically one conditions on the new set of covariates X_{n+1} rather than on the prediction set.

with

$$\delta_{\boldsymbol{y}} := \frac{|\{d \in \mathcal{I}_2 \cup \{n+1\} : R_d \geq R_{n+1}\}|}{l+1},$$

and nonconformity scores $R_d := A(\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}, \boldsymbol{z}_d)$ for $d \in \mathcal{I}_2$, $R_{n+1} := A(\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}, (x_{n+1}, \boldsymbol{y}))$. Intuitively, the nonconformity score R_d (R_{n+1} respectively) scores how different $\boldsymbol{z}_d$ ($(x_{n+1}, \boldsymbol{y})$ respectively) is from the training set, and so $\delta_{\boldsymbol{y}}$ indicates the conformity of $(x_{n+1}, \boldsymbol{y})$ to the training set compared to the conformity of the elements of the calibration set to the same training set (i.e., it is a valid p-value for testing the null hypothesis that $\boldsymbol{Y}_{n+1} = \boldsymbol{y}$, Vovk *et al.*, 2005). For example, common functional depths represent outstanding candidates in the choice of the nonconformity measure.

The essential result (due to Vovk *et al.*, 2005) traditionally evoked when dealing with the Conformal approach concerns the validity of split prediction sets: indeed, under the exchangeability assumption $\delta_{\boldsymbol{y}}$ is uniformly distributed over $\{1/(l+1), 2/(l+1), \ldots, 1\}$ and then Inequality (1.1) holds. Theorem 1.1 proves and enriches such known result by making the exact probability reached by split prediction sets explicit, by only assuming that nonconformity scores $\{R_d : d \in \mathcal{I}_2\}$ have a continuous joint distribution (an assumption that we will made hereafter). The proof is given in Appendix A.1.

Theorem 1.1. *Let $\mathcal{C}_{n,1-\alpha}(X_{n+1})$ be a Split Conformal prediction set. If $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n+1}$ are i.i.d. and $\{R_d : d \in \mathcal{I}_2\}$ have a continuous joint distribution, then*

$$\mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}(X_{n+1})\right) = 1 - \frac{\lfloor (l+1)\alpha \rfloor}{l+1}.$$

Specifically, $\mathcal{C}_{n,1-\alpha}(X_{n+1})$ always satisfies

$$1 - \alpha \leq \mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}(X_{n+1})\right) < 1 - \alpha + \frac{1}{l+1}. \tag{1.2}$$

A natural consequence of the first part of Theorem 1.1 is that when $\lfloor (l+1)\alpha \rfloor = (l+1)\alpha$ the procedure automatically outputs exact prediction sets: in practice, since in most cases both α and l are given by the application in hand, such property should be simply considered as an useful by-product that may occur in some circumstances. More generally, Theorem 1.1 states that Conformal approach ensures an easy-to-compute precise coverage for split prediction sets, and not only their validity. Furthermore, the second part of Theorem 1.1 suggests that the coverage provided by Split Conformal prediction sets is no less than $1-\alpha$ and over-coverage is basically avoided when sample size is large. In particular, Inequality 1.2 represents a minimal modification of Theorem 2 of Lei *et al.* (2018): the only difference - besides notation - is the change of ''$\leq$" with "$<$" in the upper bound of the inequality.

1.2 Smoothed Split Conformal

Moving from the notation introduced in Section 1.1, an exact prediction set for $\boldsymbol{Z}_{n+1}$ is the set such that

$$\mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}\left(X_{n+1}\right)\right) = 1 - \alpha. \tag{1.3}$$

Let us consider a single realization of a uniform random variable in $[0, 1]$, called τ_{n+1}, which is independent from all the other random objects. The Smoothed Split Conformal approach defines the prediction set for $\boldsymbol{Y}_{n+1}$ as:

$$\mathcal{C}_{n,1-\alpha,\tau_{n+1}}\left(x_{n+1}\right) := \left\{\boldsymbol{y} \in \mathcal{Y}(\mathcal{T}) : \delta_{\boldsymbol{y},\tau_{n+1}} > \alpha\right\},$$

with

$$\delta_{\boldsymbol{y},\tau_{n+1}} := \frac{\left|\{d \in \mathcal{I}_2 : R_d > R_{n+1}\}\right| + \tau_{n+1}\left|\{d \in \mathcal{I}_2 \cup \{n+1\} : R_d = R_{n+1}\}\right|}{l+1}.$$

By introducing the element of randomization τ_{n+1}, Smoothed Split Conformal prediction sets are, by construction, finite-sample exact for any α, l: to the best of our knowledge, in the literature there is no formal proof of this well-established result (due to Vovk *et al.*, 2005), and so a proof is given in Appendix A.2. It is important to notice that the division of data into the training and calibration sets induces an element of randomness into the procedure also in the Split Conformal scenario. A possible approach to limit the effect of this evidence consists of combining prediction sets obtained from different splits, but the results provided by Lei *et al.* (2018) suggest to perform a single split. As a consequence, in this thesis the aforementioned single-split process is considered.

Chapter 2

Prediction bands for univariate i.i.d. functional data

2.1 Introduction

One of the main roles of statistics in our new, data-rich world is to provide scientists, business people and policy makers with tools able to deal with an increasing amount of data, of increasing complexity. Automated sensor arrays and measuring systems now provide huge quantities of high-frequency and high-dimensional data about all sorts of social or physical phenomena.

Among the most popular toolboxes that have the capacity to deal with this kind of complex data one can find Functional Data Analysis (FDA, Ramsay and Silverman, 2005). FDA is an ebullient field of statistics which aim is to develop theory and methods to deal with data sets made of functions defined over a domain, either uni- or multi-dimensional, and usually characterized by some degree of smoothness. Since in this chapter we focus on univariate i.i.d. functional data, in the following we will indicate with $\mathcal{Y}(\mathcal{T})$ the family of functions $y : \mathcal{T} \to \mathbb{R}$ belonging to $L^\infty(\mathcal{T})$ with $\mathcal{T}$ closed and bounded subset of $\mathbb{R}^\eta$, $\eta \in \mathbb{N}_{>0}$, and with $y_1, \ldots, y_n$ possible realizations of n i.i.d. random functions $Y_1, \ldots, Y_n \sim P$ taking values in $\mathcal{Y}(\mathcal{T})$. Without loss of generality, hereafter we will consider $\eta = 1$ since it is the most common practical case.

Despite being born in relatively recent times (Ramsay, 1982), a plethora of standard multivariate tools have ported to the functional realm: among others Functional Principal Component Analysis (Ramsay and Silverman, 2005, Chapter 10), Functional Linear Regression (Ramsay and Silverman, 2005, Chapter 12) and Functional Boxplots (Sun and Genton, 2011). A problem that, perhaps surprisingly, has not been covered in a

satisfactory way in the FDA literature is the issue of uncertainty quantification in prediction and forecasting. In a more formal way, the interest is in the creation of prediction sets, namely subsets of $\mathcal{Y}(\mathcal{T})$ that include a new function Y_{n+1} (i.i.d to $Y_1, \ldots, Y_n$) with a certain nominal confidence level $1 - \alpha$. In particular, the aim is to obtain either exact - i.e. ensuring a coverage equal to the nominal confidence level - or at least valid - i.e. ensuring a coverage no less than the nominal confidence level - prediction sets. Recent works in FDA provide novel insights into this very meaningful applied and theoretical issue. These attempts can be broadly classified in two classes: a first one, composed of works based mainly on parametric bootstrapping techniques (e.g., Degras, 2011; Cao *et al.*, 2012), and a second one, where a dimensionality reduction technique is applied to render the naturally infinite-dimensional problem more tractable by projecting it on a finite dimensional functional basis (e.g., Hyndman and Shahid Ullah, 2007; Antoniadis *et al.*, 2016). These approaches carry some shortcomings: the first group of techniques is computationally intensive, thus requiring long calculation times, while the second ones rely on the approximations introduced by basis projection. Both of them, in any case, either rely on not easily provable distributional assumptions and/or on asymptotic results.

In this chapter, we build on top of the literature about set prediction for functional data and the approach presented in Chapter 1 (i.e. Conformal Prediction), by introducing several theoretical and methodological innovations.

1. In Section 2.2.1 we propose a nonconformity measure inducing a conformal predictor able to create closed-form finite-sample either valid or exact prediction bands of constant amplitude, under minimal distributional assumptions. The procedure is fast, scalable and does not rely on widespread functional dimension reduction techniques.

2. In Section 2.2.2 we propose a family of nonconformity measures (to which the nonconformity measure introduced in Section 2.2.1 belongs) indexed by modulation function $s_{\mathcal{I}_1}$ that allows for prediction bands with non-constant width, but able to keep all the aforementioned appealing properties. As a consequence, prediction bands induced by the nonconformity measures belonging to this family can be compared on the basis of features other than validity, such as efficiency (i.e. the size).

3. In Section 2.2.2 we focus on a specific nonconformity measure belonging to this family which leads to valid prediction bands asymptotically no less efficient than those obtained by not modulating (Theorem 2.4, Theorem 2.5).

Finally, in Section 2.3 we propose a simulation study to compare our method with four alternatives, and in Section 2.4 we apply our approach to the Berkeley Growth Study data set (Tuddenham and Snyder, 1954). Section 2.5 provides an overview of the main results.

2.2 Methods

2.2.1 The Nonconformity Measure

Although some authors proposed different approaches to find prediction bands under the Gaussian assumption (Yao *et al.*, 2005) and through finite dimensional projection (Lei *et al.*, 2015), to the best of our knowledge no method to create valid prediction bands by only assuming i.i.d. functional data and by avoiding dimension reduction is available in the literature.

In light of this, we propose a fast and scalable Conformal predictor that outputs closed-form finite-sample valid (or even exact) prediction bands under only the i.i.d. assumption. Indeed, the Conformal framework ensures, by construction, that the prediction sets obtained are always valid/exact (see Chapter 1), but other features such as shape and size depend on the specific nonconformity measure used: as a consequence, the core of the Conformal approach is represented by the choice of such measure.

In particular, the nonconformity measure we propose automatically allows to obtain prediction bands and is based on the essential supremum:

$$A(\{y_h : h \in \mathcal{I}_1\}, y) = \operatorname*{ess\,sup}_{t \in \mathcal{T}} |y(t) - g_{\mathcal{I}_1}(t)|, \tag{2.1}$$

with $g_{\mathcal{I}_1} : \mathcal{T} \to \mathbb{R}$ a function belonging to $L^\infty(\mathcal{T})$ based on $\{y_h : h \in \mathcal{I}_1\}$ and acting as a point predictor of the new observation. Although valid prediction bands are obtained regardless the specific $g_{\mathcal{I}_1}$ involved, a careful choice of this function helps to obtain small prediction bands, a desirable property from an application point of view which will be investigated in Section 2.2.2 (Lei *et al.*, 2018). In view of this, $g_{\mathcal{I}_1}$ is typically a point predictor summarizing information provided by $\{y_h : h \in \mathcal{I}_1\}$, e.g. the sample functional mean. However, since the purpose of the chapter is to construct either valid or exact prediction bands starting from any point predictor in order to obtain a widely usable procedure, later in the discussion we will always consider $g_{\mathcal{I}_1}$ as given - and properly chosen by the expert according to the specific framework considered. Focusing on the Split Conformal scenario (the minor changes needed for the Smoothed Split case are

introduced in Appendix B.1) and by using $\mathcal{C}_{n,1-\alpha}$ instead of $\mathcal{C}_{n,1-\alpha}(x_{n+1})$ to indicate a given prediction set due to the absence of covariates, first of all it is possible to notice that if $\alpha \in (0, 1/(l+1))$ then $\mathcal{C}_{n,1-\alpha} = \mathcal{Y}(\mathcal{T})$ since δ_y can not be less than $1/(l+1)$: for this reason, later in the discussion we will always consider $\alpha \in [1/(l+1), 1)$, unless otherwise stated. If $\alpha \in [1/(l+1), 1)$, the definition of $\mathcal{C}_{n,1-\alpha}$ and δ_y implies that $y \in \mathcal{C}_{n,1-\alpha} \iff R_{n+1} \leq k$, with k the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d : d \in \mathcal{I}_2\}$. Then

$$\begin{aligned} &\operatorname*{ess\,sup}_{t \in \mathcal{T}} |y(t) - g_{\mathcal{I}_1}(t)| \leq k \iff \\ &|y(t) - g_{\mathcal{I}_1}(t)| \leq k \quad \forall t \in \mathcal{T} \iff \\ &y(t) \in [g_{\mathcal{I}_1}(t) - k, g_{\mathcal{I}_1}(t) + k] \quad \forall t \in \mathcal{T}. \end{aligned}$$

Therefore, the Split Conformal prediction set induced by the nonconformity measure (2.1) is

$$\begin{aligned} \mathcal{C}_{n,1-\alpha} := \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in &[g_{\mathcal{I}_1}(t) - k, g_{\mathcal{I}_1}(t) + k] \\ &\forall t \in \mathcal{T}\}. \end{aligned} \tag{2.2}$$

Besides having the shape of a a band, the introduced prediction set can be found in closed form, an appealing property that incredibly speeds up computation time. In addition, the Conformal framework and the simplicity of the nonconformity measure ensure highly scalable prediction bands as, on top of the cost needed to build the point predictor $g_{\mathcal{I}_1}$, the time required to find k increases linearly with l. Then, if a particularly sophisticated predictor is chosen for $g_{\mathcal{I}_1}$, one is justified in expecting the total computation cost to be dominated by the calculation of such point predictor. Moreover, as usual in the prediction framework the band is built around a "central" object ($g_{\mathcal{I}_1}$ in this case), a fact that further suggests to define this function as a data-driven point predictor. Finally, the prediction bands defined in (2.2) are simultaneous by construction, i.e. bands ensuring the desired coverage globally (in addition to the pointwise validity). Similarly to the multivariate (non-functional) setting, a simple concatenation of pointwise prediction intervals based on the pointwise nonconformity score $|y(t) - g_{\mathcal{I}_1}(t)|$ for all $t \in \mathcal{T}$ would lead to a prediction band: that is a subset of the simultaneous prediction band (2.2) (the proof is given in Appendix B.1); with guaranteed pointwise coverage for all $t \in \mathcal{T}$; but whose simultaneous coverage over the domain $\mathcal{T}$ can be dramatically lower than the desired one. Moreover, in application scenarios where data are characterized by specific features (e.g., positivity, monotonicity etc...), the approach presented in this

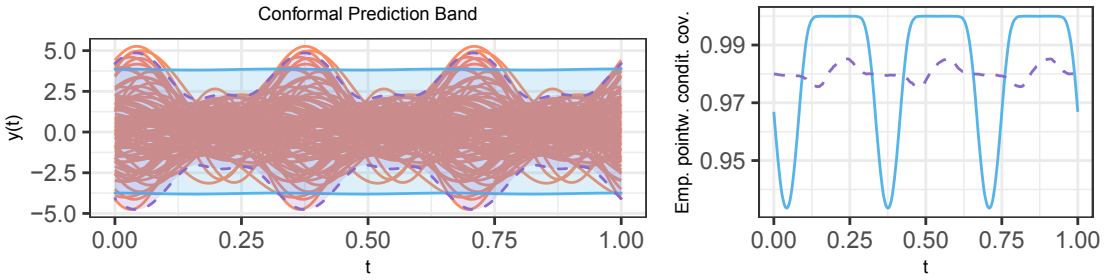

FIGURE 2.1: The left panel shows the Split Conformal prediction band computed as in (2.2) (solid light blue band) and that computed as in (2.4) by considering the standard deviation function as $s_{\mathcal{I}_1}$ (dashed purple band). For visualization, a random subsample of $y_1, \ldots, y_{198}$ is plotted. The right panel shows the empirical pointwise conditional coverage reached by the first band (solid light blue line) and by the second one (dashed purple line). $\alpha = 0.1$.

section allows to remove portions of the observed prediction bands that violate such known characteristics, without affecting the coverage and at the same time leading to "smaller" prediction bands (as shown by Chernozhukov *et al.*, 2019). An example of this band trimming procedure is given in Section 2.4. This possibility is a desirable implication which derives from using a fully nonparametric approach to prediction, since this takes away the burden of an explicit and possibly non-trivial modeling of the existing constraints.

2.2.2 Improving Efficiency: the Choice of the Modulation Function

It can be easily noted that the width of (2.2) over $\mathcal{T}$ is constant and equal to $2k$ but, intuitively, prediction bands that do not adapt their width according to the local variability of functional data, even though theoretically sound, may be of limited interest in real applications. Let us consider the following running example: let $y_1, \ldots, y_{198}$ be independent realizations of the random function $Y(t) := X_1 + X_2 \cos(6\pi t) + X_3 \sin(6\pi t)$, with $t \in [0, 1]$ and (X_1, X_2, X_3) being a Gaussian random vector such that $\mathrm{E}[X_i] = 0$, $\mathrm{Var}[X_i] = 1$, $\mathrm{Cov}[X_i, X_j] = 0.6$ for $i, j = 1, 2, 3$, $i \neq j$. The solid light blue band in the left panel of Figure 2.1 shows the prediction band obtained by the procedure presented in Section 2.2.1 considering $\alpha = 0.1$, $m = n/2$ and $g_{\mathcal{I}_1}$ sample functional mean of the training set: given the different variability of functional data over $\mathcal{T}$, in the low-variance parts of the domain the prediction band is dramatically large containing all the pointwise evaluations of the functional data (see, for example, $t = 0.5$ and nearby points).

A possible solution to this drawback consists of defining the following nonconformity measure and nonconformity scores:

$$A(\{y_h : h \in \mathcal{I}_1\}, y) = \operatorname*{ess\,sup}_{t \in \mathcal{T}} \left| \frac{y(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right|, \tag{2.3}$$

$$R^s_d := \operatorname*{ess\,sup}_{t \in \mathcal{T}} \left| \frac{y_d(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right|,$$
$$R^s_{n+1} := \operatorname*{ess\,sup}_{t \in \mathcal{T}} \left| \frac{y(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right|,$$

with $d \in \mathcal{I}_2$ and $s_{\mathcal{I}_1} := s(\{y_h : h \in \mathcal{I}_1\}) : \mathcal{T} \to \mathbb{R}_{>0}$ a function which belongs to $L^\infty(\mathcal{T})$ based on $\{y_h : h \in \mathcal{I}_1\}$. At the interpretative level, the new nonconformity measure (2.3) can be suitably considered as the nonconformity measure (2.1) taking the transformed functions $y^s(t) := y(t)/s_{\mathcal{I}_1}(t)$ and $g^s_{\mathcal{I}_1}(t) = g_{\mathcal{I}_1}(t)/s_{\mathcal{I}_1}(t)\ \forall t \in \mathcal{T}$ as input instead of the original functions $y(t)$, $g_{\mathcal{I}_1}(t)$. It is important to notice that, since $s_{\mathcal{I}_1}(t) > 0\ \forall t \in \mathcal{T}$, the function $s_{\mathcal{I}_1}$ modulates the original data without altering the order of the functions at each point t: for this reason, later in the discussion the term *modulation function* will be used to refer to $s_{\mathcal{I}_1}$.

Therefore, the Split Conformal prediction band induced by the nonconformity measure (2.3), obtained by replicating the computations of Section 2.2.1 (see Appendix B.2 for the proof for both the Split Conformal and Smoothed Split Conformal frameworks), is

$$\begin{aligned} \mathcal{C}^s_{n,1-\alpha} :=& \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in \\ & [g_{\mathcal{I}_1}(t) - k^s s_{\mathcal{I}_1}(t), g_{\mathcal{I}_1}(t) + k^s s_{\mathcal{I}_1}(t)] \forall t \in \mathcal{T}\}, \end{aligned} \tag{2.4}$$

with k^s the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R^s_d : d \in \mathcal{I}_2\}$. In other words, the procedure presented in this section consists of modulating the data, computing the prediction band (2.2) by using the transformed data and back-transforming it in the non-modulated space: in so doing, prediction bands adapt their width according to the specific modulation function chosen and their validity is guaranteed by the Conformal framework. A similar consideration has been highlighted also in the scalar regression setting by Lei *et al.* (2018), who proposed a locally weighted Split Conformal method to vary the width of the prediction sets over the covariates $x \in \mathbb{R}^p$.

In order to understand the modification introduced by the modulation function, let us consider the aforementioned running example and specifically the left panel of Figure 2.1: in this case, the band obtained by considering the standard deviation function

(Ramsay and Silverman, 2005) as $s_{\mathcal{I}_1}$ (dashed purple band) is deeply different from the solid light blue one and it seems to better adapt to the variability of the data over $\mathcal{T}$. Intuitively, one is justified in accepting the bands to become wider in the parts of the domain where data show high variability in order to obtain narrower and more informative prediction bands in those parts characterized by low variability.

Remark 2.1. Replacing function $s_{\mathcal{I}_1}$ with $s_{\mathcal{I}_2}$ does not allow to obtain closed-form valid prediction bands. This is due to the fact that their dependence on the calibration set involves $\{R_d^s : d \in \mathcal{I}_2 \cup \{n+1\}\}$ not being exchangeable, and consequently validity not being guaranteed.

Remark 2.2. Prediction bands induced by the modulation functions $s_{\mathcal{I}_1}$ and $\lambda \cdot s_{\mathcal{I}_1}$, with $\lambda \in \mathbb{R}_{>0}$, are identical. The proof is given in Appendix B.2. As a consequence, an equivalence relation naturally arises and so for each specific equivalence class (made up of modulation functions equal up to a multiplicative factor) we will consider the modulation function whose integral is equal to 1. In view of this, the original nonconformity measure (2.1) can be interpreted as the nonconformity measure induced by the modulation function $s^0(t) := 1/|\mathcal{T}| \ \forall t \in \mathcal{T}$, whose notation does not include the subscript $\mathcal{I}_1$ to underline the lack of dependence of this function on the training set.

Remark 2.3. One of the aim of the introduction of $s_{\mathcal{I}_1}$ is to reduce the variability of the pointwise miscoverage over $\mathcal{T}$. In order to clarify this concept, let us consider the right panel of Figure 2.1. The solid light blue (dashed purple respectively) line shows the empirical pointwise conditional coverage of the solid light blue (dashed purple respectively) prediction band showed in the left panel of the same figure, that was obtained by setting $\alpha = 0.1$. The empirical conditional coverage has been computed considering the number of times that 200,000 - independent from and identically distributed to the original sample - new functions belong to the two prediction bands over $\mathcal{T}$. As expected, the absence of modularization involves the empirical pointwise converage being highly variable over $\mathcal{T}$, whereas the use of the standard deviation function as modulation function leads to an empirical pointwise coverage concentrated around 0.98.

However, in absence of an optimality criterion there are no formal reasons to prefer a specific modulation function over another, as Conformal approach ensures valid prediction sets regardless the choice of $s_{\mathcal{I}_1}$. In this regard, a criterion that naturally arises in the prediction framework to discriminate between modulation functions is maximization of efficiency, i.e. minimization of the size of prediction sets (Vovk *et al.*, 2005) . The reason of this choice is very intuitive: since prediction bands are, by construction, valid, one is justified in seeking small prediction bands because they include subregions of the sample space where the probability mass is concentrated (Lei *et al.*, 2013). In view of

this, first of all it is essential to define what the size of a prediction band is, a nontrivial topic in the functional framework. The definition we will consider is simply the area between the upper and lower bound of the prediction band:

$$\mathcal{Q}(s_{\mathcal{I}_1}) := \int_{\mathcal{T}} 2 \cdot k^s \cdot s_{\mathcal{I}_1}(t)dt = 2 \cdot k^s, \tag{2.5}$$

that is equal to k^s up to a constant since $\int_{\mathcal{T}} s_{\mathcal{I}_1}(t)dt = 1$.

Formally, in the usual finite-dimensional setting the aim would be to find the optimal modulation function that minimizes the risk functional $\mathrm{E}[k^s]$. Unfortunately, in the functional setting even the concept of probability density function is generally not well defined since there is no σ-finite dominating measure (Delaigle *et al.*, 2010), and so that minimization is not feasible for general P. As a consequence, the minimization problem must be simplified: by considering k^s as a non-random quantity depending on observed functions $y_1, \ldots, y_n$ instead of random functions $Y_1, \ldots, Y_n$, the aim becomes the direct minimization of k^s. Although initially it may seem like an oversimplification to some readers, it is important to underline that this approach is made possible by a well-established principle representing the core idea of many algorithms and methods (e.g. machine learning techniques) known as empirical risk minimization principle (Vapnik, 1992).

The proposed adjustment reduces the complexity of the optimization task, but the problem still presents tricky aspects. Indeed, not only the minimization can not be analytically addressed by calculus of variations given the complexity of k^s, but also the optimal modulation function can not be uniquely determined given the specific structure of $R^s_d, d \in \mathcal{I}_2$. In fact, the dependency of $s_{\mathcal{I}_1}$ only on the functions of the training set and of the numerator of R^s_d (i.e. $|y_d(t) - g_{\mathcal{I}_1}(t)|, \quad d \in \mathcal{I}_2$) also on the functions of the calibration set makes the optimization unfeasible for all P and the general problem ill-posed.

In such a non-standard context, the line of reasoning must necessarily be changed. Therefore, in the discussion below we focus on finding a function - called c-function hereafter for the sake of simplicity - satisfying the definition of modulation function but depending also on the calibration set through $\{y_d : d \in \mathcal{I}_2\}$ and such that

1. for $m, l \to +\infty$ it converges to a given function and its training counterpart (i.e. the function - called t-function hereafter - equal to the c-function but whose dependence on $\{y_d : d \in \mathcal{I}_2\}$ is replaced by the dependence on the training set through $\{y_h : h \in \mathcal{I}_1\}$) converges to the same function

2. it leads to prediction bands that are not wider (in the sense of (2.5)) than those obtained by not modulating (i.e. by using s^0)

If these two conditions are met, the use of the t-function as modulation function ensures that valid prediction bands are obtained (due to its dependence only on $\{y_h : h \in \mathcal{I}_1\}$) and that asymptotically the second condition is satisfied. Specifically, that condition represents a desirable and appealing property since, if violated, the modulation process could represent a meaningless complication compared to the original nonconformity measure (2.1).

In order to construct a c-function able to meet these two conditions, it is important to focus on what k^s is: ignoring just for now the contribution of the modulation function, k^s is a quantity derived by the $\lceil (l+1)(1-\alpha) \rceil$th least extreme function between those in the calibration set, in which the concept of "extreme" is naturally induced by the metric used. In light of this, the guidelines we decided to follow in the construction of a meaningful c-function are two. First of all, the behavior of the $l - \lceil (l+1)(1-\alpha) \rceil$ most extreme functions in the calibration set should not be taken into account since they do not affect the value of k^s. Secondly, given the specific nonconformity measure considered, the c-function should modulate data considering the remaining $\lceil (l+1)(1-\alpha) \rceil$ functions on the basis of the most extreme value observed $\forall t \in \mathcal{T}$.

Inspired by these guidelines, we propose the following c-function:

$$\bar{s}^c_{\mathcal{I}_1}(t) := \frac{\max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|}{\int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt} \tag{2.6}$$

with

$$\mathcal{H}_2 := \{d \in \mathcal{I}_2 : \operatorname*{ess\,sup}_{t \in \mathcal{T}} |y_d(t) - g_{\mathcal{I}_1}(t)| \leq k\}$$

and k defined as in Section 2.2.1, i.e. the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d : d \in \mathcal{I}_2\}$. The corresponding t-function is

$$\bar{s}_{\mathcal{I}_1}(t) := \frac{\max_{h \in \mathcal{H}_1} |y_h(t) - g_{\mathcal{I}_1}(t)|}{\int_{\mathcal{T}} \max_{h \in \mathcal{H}_1} |y_h(t) - g_{\mathcal{I}_1}(t)| dt} \tag{2.7}$$

with $\mathcal{H}_1 = \mathcal{I}_1$ if $\lceil (m+1)(1-\alpha) \rceil > m$, otherwise

$$\mathcal{H}_1 := \{h \in \mathcal{I}_1 : \operatorname*{ess\,sup}_{t \in \mathcal{T}} |y_h(t) - g_{\mathcal{I}_1}(t)| \leq \gamma\}$$

with γ the $\lceil (m+1)(1-\alpha) \rceil$th smallest value in the set $\{\operatorname{ess\,sup}_{t \in \mathcal{T}} |y_h(t) - g_{\mathcal{I}_1}(t)| : h \in \mathcal{I}_1\}$.

In order not to overcomplicate the notation, in the definition of $\bar{s}^c_{\mathcal{I}_1}$ and $\bar{s}_{\mathcal{I}_1}$ we quietly assumed that both numerators are different from 0 $\forall t \in \mathcal{T}$ almost surely. If not,

the adjustment described in Appendix B.2 is developed. From an operational point of view, t-function $\bar{s}_{\mathcal{I}_1}(t)$ ignores the most extreme functions (i.e. the functions belonging to $\mathcal{I}_1 \setminus \mathcal{H}_1$) and modulates data on the basis of the remaining non-extreme functions. Specifically, the dependence of γ on α allows to provide carefully chosen modulation process according to the specific level $1-\alpha$ chosen for the prediction set.

The fulfillment of the two aforementioned conditions by the function (2.6) is proved by the following two theorems.

Theorem 2.4. *Let $m/n = \theta$ with $0 < \theta < 1$ and let $\mathrm{Var}[g_{\mathcal{I}_1}(t)] \to 0$ when $m \to +\infty$. Then $\bar{s}^c_{\mathcal{I}_1}$ and $\bar{s}_{\mathcal{I}_1}$ converge to the same function when $n \to +\infty$.*

Theorem 2.5. *$\mathcal{Q}(s^0) \geq \mathcal{Q}(\bar{s}^c_{\mathcal{I}_1})$. Specifically, $\mathcal{Q}(s^0) = \mathcal{Q}(\bar{s}^c_{\mathcal{I}_1})$ if and only if $\max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|$ is constant almost everywhere.*

Both proofs are given in Appendix B.2. It is important to notice that Theorem 2.4 requires very mild conditions, an evidence that allows it to hold in many general contexts.

In light of this, the function (2.7) represents an outstanding candidate in the choice of the modulation function since the Conformal setting and the nonconformity measure (2.3) guarantee valid prediction bands - as well as all the other desirable properties highlighted in Section 2.2.1 - and at the same time to asymptotically obtain prediction bands no less efficient than those induced by s^0.

The fact that $\bar{s}^c_{\mathcal{I}_1}(t)$ leads to prediction bands that are not wider than those obtained by not modulating is not the only relevant result that is possible to obtain. The following Theorem shows that prediction bands induced by $\bar{s}^c_{\mathcal{I}_1}$ are also smaller than those induced by the functions belonging to a specific group. This theorem provides a further theoretical justification for preferring function (2.7) to other possible modulation functions.

Theorem 2.6. *Let us define $\mathcal{CH}_2 := \mathcal{I}_2 \setminus \mathcal{H}_2$ and let t^*_d be the value such that*

$$|y_d(t^*_d) - g_{\mathcal{I}_1}(t^*_d)| = \operatorname*{ess\,sup}_{t\in\mathcal{T}} |y_d(t) - g_{\mathcal{I}_1}(t)| \quad \forall d \in \mathcal{I}_2. \tag{2.8}$$

If t^*_d is not unique, it is randomly chosen from the values that satisfy (2.8).

Let $s^{\zeta}_{\mathcal{I}_1}$ be a modulation function such that:

1. *$s^{\zeta}_{\mathcal{I}_1} \neq \bar{s}^c_{\mathcal{I}_1}$ in the sense of Lebesgue, i.e. $\exists\, \mathcal{T}^* \subseteq \mathcal{T}$ such that $s^{\zeta}_{\mathcal{I}_1}(t) \neq \bar{s}^c_{\mathcal{I}_1}(t)\ \forall t \in \mathcal{T}^*$ and $\mu(\mathcal{T}^*) > 0$, with μ the Lebesgue measure*

2. *$s^{\zeta}_{\mathcal{I}_1}(t^*_i) \leq \bar{s}^c_{\mathcal{I}_1}(t^*_i)\ \forall i \in \mathcal{CH}_2$*

If $|\mathcal{H}_2| = \lceil (l+1)(1-\alpha) \rceil$, *then* $\mathcal{Q}(s^{\zeta}_{\mathcal{I}_1}) > \mathcal{Q}(\bar{s}^{c}_{\mathcal{I}_1})$.

The proof is given in Appendix B.2, along with the demonstration that Theorem 2.5 is not a direct consequence of Theorem 2.6 since s^0 may not fulfill $s^0(t_i^*) \leq \bar{s}^c_{\mathcal{I}_1}(t_i^*)$ $\forall i \in \mathcal{CH}_2$. Also in this case, the field of application of Theorem 2.6 is particularly wide since the condition about the cardinality of $|\mathcal{H}_2|$ is always met under the assumption concerning the continuous joint distribution of $\{R_d : d \in \mathcal{I}_2\}$ made in Chapter 1. The definitions of functions (2.6), (2.7) and Theorems 2.4, 2.5 and 2.6 can be easily generalized to hold also in the Smoothed Conformal framework. Technical details are provided in Appendix B.2. Before moving on to the Simulation Study, we would like to point out that the criterion used in this chapter to find an optimal modulation function is different in nature from the typical results provided in the Conformal Prediction framework (which for example are Oracle results, Lei *et al.*, 2018). This choice is due to the fact that the Conformal Prediction asymptotic optimality properties typically require specifying models for the true data generating process and/or analyzing the asymptotic properties of estimators, which is something out of scope given the intrinsic issues of the functional context and the full generality that characterizes the chapter.

2.3 Simulation Study

2.3.1 Study Design

In this section, we summarize the results of a two-stage simulation study comparing our approach with four alternative methods from the literature that will be detailed in the following: Naive, Band Depth, Modified Band Depth, and Bootstrap. In Section 2.3.2 the empirical coverage is evaluated for each approach in three different scenarios, whereas in Section 2.3.3 the prediction bands obtained by the methods that guarantee a proper coverage are compared in terms of efficiency. The hierarchical structure of the simulation study reflects the "nested" nature of the two features we are considering, i.e. coverage and size: indeed, the size of a prediction set should be investigated only after verifying that the method which outputted that specific prediction set guarantees the desired coverage, which represents the primary aspect when assessing prediction sets.

Specifically, the three scenarios allow to compare the methods in three different frameworks: when data show a constant variability over the domain (Scenario 1), when data show a different variability over the domain (Scenario 2) and when data are characterized by outliers (Scenario 3). Formally, the three scenarios are:

- Scenario 1. $\forall i = 1, \ldots, n$

$$y_i(t) = x_{i1} + x_{i2}\cos(6\pi(\,t + u_i)) + x_{i3}\sin(6\pi\,(t + u_i))$$

with $\mathcal{T} = [0, 1]$, $(x_{11}, x_{12}, x_{13})^T, \ldots, (x_{n1}, x_{n2}, x_{n3})^T$ i.i.d. realizations of

$$X \sim N_3\left(\mathbf{0}, \begin{bmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 1 \end{bmatrix}\right)$$

and $u_1, \ldots, u_n$ i.i.d. realizations of

$$U \sim \text{Unif}\left[-\frac{1}{6}, \frac{1}{6}\right].$$

- Scenario 2. $\forall i = 1, \ldots, n$

$$y_i(t) = \sum_{j=1}^{13} c_{ij} B_j^{\omega}(t)$$

with $\mathcal{T} = [0, 1]$, $B_j^{\omega}(t)$ the b-spline basis system of order 4 with interior knots $\omega = (0.1, 0.2, \ldots, 0.9)$ and $(c_{1,1}, \ldots, c_{1,13})^T, \ldots, (c_{n,1}, \ldots, c_{n,13})^T$ i.i.d. realizations of $C = (C_1, \ldots, C_{13}) \sim N_{13}(\mathbf{0}, \Sigma)$ such that $\text{Var}[C_i] = 0.03^2$ $\forall i \neq 7$, $\text{Var}[C_7] = 0.003^2$ and $\text{Cov}[C_i, C_j] = 0$ for $i, j = 1, \ldots, 13$, $i \neq j$.

- Scenario 3. The scenario is the previous one after contamination with outliers. Formally, $(c_{1,1}, \ldots, c_{1,13})^T, \ldots, (c_{n,1}, \ldots, c_{n,13})^T$ are i.i.d. realizations of a vector random variable whose probability density function is a Gaussian mixture density with weights $(1 - \beta, \beta)$, shared mean vector $\mathbf{0}$, the covariance matrix defined as in Scenario 2 for the first group and such that $\text{Var}[C_7] = 0.3^2$ instead of $\text{Var}[C_7] = 0.003^2$ for the second group.

A graphical representation of a replication for each scenario with $n = 18$ is provided in Figure 2.2. The Conformal approach is evaluated in the Split Conformal framework and considering three different modulation functions: s^0, the normalized standard deviation function $s^{\sigma}_{\mathcal{I}_1}$ as natural representative of functions that capture data variability, and $\bar{s}_{\mathcal{I}_1}$. Since the focus of the work is not on the construction of sophisticated point predictors $g_{\mathcal{I}_1}$ but rather on the construction of valid prediction bands around any point predictor $g_{\mathcal{I}_1}$, we hereby simply set $g_{\mathcal{I}_1}(t) = \bar{y}_{\mathcal{I}_1}(t)$.

The performance of our approach is compared to four alternative methods. These are: *Naive* method, which outputs prediction bands defined as $\{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in [q_{\frac{\alpha}{2}}(t), q_{1-\frac{\alpha}{2}}(t)]\ \forall t \in \mathcal{T}\}$ with $q_{\alpha}(t)$ empirical quantile of order α for $(y_1(t), \ldots, y_n(t))$. Such approach represents a very naive solution to the prediction task we are considering

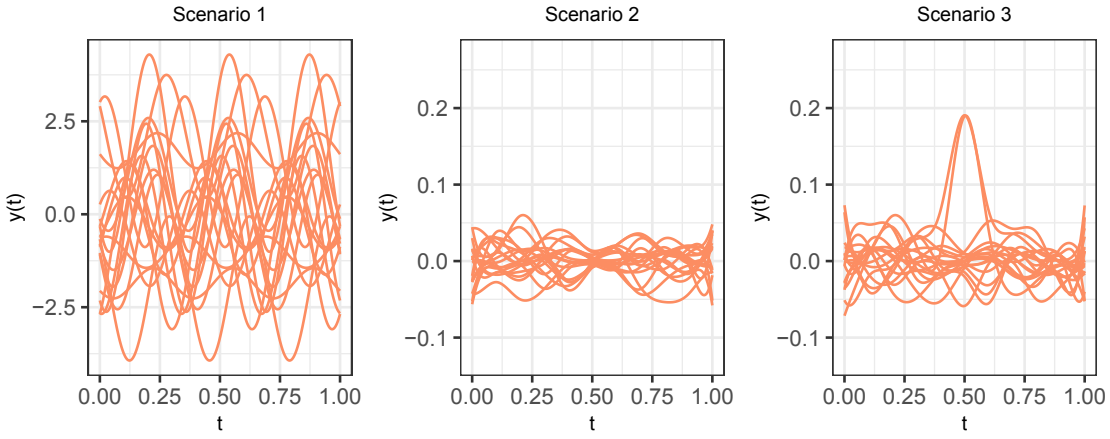

FIGURE 2.2: Graphical representation of the scenarios. The sample size is $n = 18$.

and we expect it to suffer greatly from undercoverage; *BD* and *MBD* methods, which output the sample $(1 - \alpha)$ central region induced by the band depth (BD) and the modified band depth (MBD) respectively (Sun and Genton, 2011); *Boot.* method, which outputs the band based on 2500 bootstrap samples, as proposed by Degras (2011). We consider $\alpha = 0.1$, $\beta = 0.06$ and three different sample sizes: $n = 18$, $n = 198$, $n = 1998$. In order not to overcomplicate the simulation study, the ratio $\rho = l/n$ is kept fixed and equal to 0.5 as commonly suggested in the Conformal literature. A deeper investigation about the possible effect of the ratio $\rho = l/n$ on efficiency - even though possibly interesting - is out of the scope of this work. The atypical values of n in the simulations have been simply chosen to have a miscoverage exactly equal to α (indeed in these cases $\lfloor (l+1)\alpha \rfloor / (l+1) = \alpha$) and consequently making the simulation results easier to read. Similar results would have been attained with rounded values of n (e.g. $n = 20$, $n = 200$, $n = 2000$) by evaluating the empirical miscoverage considering the theoretical one: $\lfloor (l+1)\alpha \rfloor / (l+1)$ (see Theorem 1.1). The simulations are achieved by using the R Programming Language (R Core Team, 2020) and the computation of the band depth and the modified band depth by *roahd* package (Tarabelloni *et al.*, 2018). Finally, every combination of scenario and sample size is evaluated considering $N = 500$ replications.

2.3.2 Coverage

Table 2.1 shows the sample mean and the standard deviation of the empirical conditional coverage provided by the prediction bands generated by each method for each combination of sample size and scenario. Specifically, the empirical conditional coverage of a given prediction band (i.e. the empirical coverage obtained conditioning on the prediction band obtained by the observed data) is computed as the fraction of times

		Conformal Method			**Alternative Methods**			
		s^0	$s^{\sigma}_{\mathcal{I}_1}$	$\bar{s}_{\mathcal{I}_1}$	Naive	MBD	BD	Boot.
$n = 18$	Scenario 1	0.902	0.900	0.900	0.409	0.504	0.547	0.875
		(0.088)	(0.085)	(0.087)	(0.092)	(0.109)	(0.111)	(0.064)
	Scenario 2	0.901	0.910	0.909	0.048	0.123	0.145	0.922
		(0.089)	(0.081)	(0.083)	(0.021)	(0.044)	(0.051)	(0.042)
	Scenario 3	0.904	0.904	0.907	0.049	0.124	0.148	0.932
		(0.084)	(0.089)	(0.085)	(0.023)	(0.049)	(0.055)	(0.061)
$n = 198$	Scenario 1	0.901	0.902	0.901	0.625	0.861	0.900	0.865
		(0.029)	(0.030)	(0.031)	(0.031)	(0.028)	(0.028)	(0.019)
	Scenario 2	0.901	0.899	0.900	0.189	0.733	0.788	0.897
		(0.029)	(0.031)	(0.029)	(0.019)	(0.036)	(0.032)	(0.015)
	Scenario 3	0.897	0.900	0.899	0.197	0.742	0.798	0.892
		(0.031)	(0.030)	(0.031)	(0.020)	(0.034)	(0.030)	(0.020)
$n = 1998$	Scenario 1	0.900	0.899	0.900	0.666	0.942	0.918	0.866
		(0.010)	(0.010)	(0.010)	(0.011)	(0.006)	(0.008)	(0.008)
	Scenario 2	0.900	0.900	0.899	0.233	0.958	0.971	0.899
		(0.009)	(0.010)	(0.010)	(0.007)	(0.006)	(0.005)	(0.008)
	Scenario 3	0.900	0.899	0.900	0.240	0.959	0.973	0.884
		(0.010)	(0.010)	(0.010)	(0.008)	(0.006)	(0.005)	(0.007)

TABLE 2.1: For each combination of sample size and scenario, the first line shows the sample mean of the empirical conditional coverage, the second line the sample standard deviation in brackets. A combination of mean and st. deviation is gray-colored if the corresponding 99% confidence t-interval for the (unconditional) coverage includes value $1 - \alpha$.

that 10,000 new functions - independent from and identically distributed to the original sample - belong to such prediction band. The purpose of this scheme is twofold: first of all, by averaging the $N = 500$ empirical conditional coverages obtained for each combination of scenario and sample size it is possible to obtain the empirical coverage, which is an estimate of the (unconditional) coverage. Secondly, this scheme allows to evaluate the variability of the conditional coverage when the observed sample varies, a particularly useful indication in real applications. In order to facilitate the visualization of the results and to allow inferential conclusions, a specific combination of sample mean and standard deviation is gray-colored in Table 2.1 if the corresponding 99% confidence t-interval for the (unconditional) coverage includes 0.90, i.e. the value $1 - \alpha$.

The simulation study fully confirms the theoretical property concerning the validity of Split Conformal prediction sets with 53 out of the 54 99%-confidence intervals associated to conformal bands including the nominal value $1 - \alpha$. The evidence provided is particularly appealing since the desired coverage is guaranteed also when a very small sample size ($n = 18$) is considered, a framework in which such property is traditionally hard to obtain. Vice versa, in almost all cases the alternative methods do not ensure the

		s^0		$s^{\sigma}_{\mathcal{I}_1}$		$\bar{s}_{\mathcal{I}_1}$	
		$Mean$	$st.dev$	$Mean$	$st.dev$	$Mean$	$st.dev$
$n = 18$	Scenario 1	8.113	(2.044)	10.088	(3.618)	11.638	(4.309)
	Scenario 2	0.142	(0.025)	0.165	(0.041)	0.185	(0.049)
	Scenario 3	0.246	(0.192)	0.448	(0.550)	0.505	(0.633)
$n = 198$	Scenario 1	7.175	(0.560)	7.295	(0.608)	7.556	(0.647)
	Scenario 2	0.127	(0.006)	0.109	(0.005)	0.120	(0.006)
	Scenario 3	0.139	(0.013)	0.139	(0.013)	0.137	(0.020)
$n = 1998$	Scenario 1	7.059	(0.179)	7.065	(0.176)	7.128	(0.184)
	Scenario 2	0.125	(0.002)	0.106	(0.001)	0.117	(0.002)
	Scenario 3	0.136	(0.003)	0.137	(0.004)	0.131	(0.003)

TABLE 2.2: Size of the prediction bands. For each row, the lowest value of the sample mean is gray-colored.

desired coverage with some estimates dramatically far from $1 - \alpha$, especially for small sample sizes (i.e., $n = 18$). In view of this, in Section 2.3.3 only the efficiency of the Conformal methods is evaluated and compared.

2.3.3 Efficiency

Table 2.2 shows the sample mean and the standard deviation of the size defined as in Equation 2.5 of the prediction bands computed in the previous section for each combination of modulation function, sample size and scenario. First of all, it is noticeable that when $n = 18$ the absence of modulation (i.e. s^0) seems to provide smaller prediction bands than those induced by $s^{\sigma}_{\mathcal{I}_1}$ and $\bar{s}_{\mathcal{I}_1}$, conceivably because the extremely low number of functions belonging to the training set ($m = 9$) leads to an unstable and possibly misleading modulation function supporting the statistical intuition that for small sample sizes simpler modulation functions should be preferred.

More deeply, focusing now on each scenario separately and considering the remaining sample sizes, Scenario 1 represents a framework in which a constant width prediction band is the ideal candidate since the horizontal shift due to the random variable U induces constant variance along the domain. As a consequence, the pointwise evaluations $Y(t)$ are equally distributed $\forall t \in \mathcal{T}$ and so one is justified in expecting $s^{\sigma}_{\mathcal{I}_1}$ and $\bar{s}_{\mathcal{I}_1}$ to be of no practical use. The results confirm this conjecture, but the differences between the three modulation functions seems to decrease as the sample size grows (see, for example, the difference between s^0 and $\bar{s}_{\mathcal{I}_1}$ when n increases from 198 to 1998).

Scenario 2 represents a completely different setting, in which a modulation process is appropriate since the curves highlight a reduction of variability in the central part of the domain. As expected, s^0 induces larger predictions bands (on average) than those

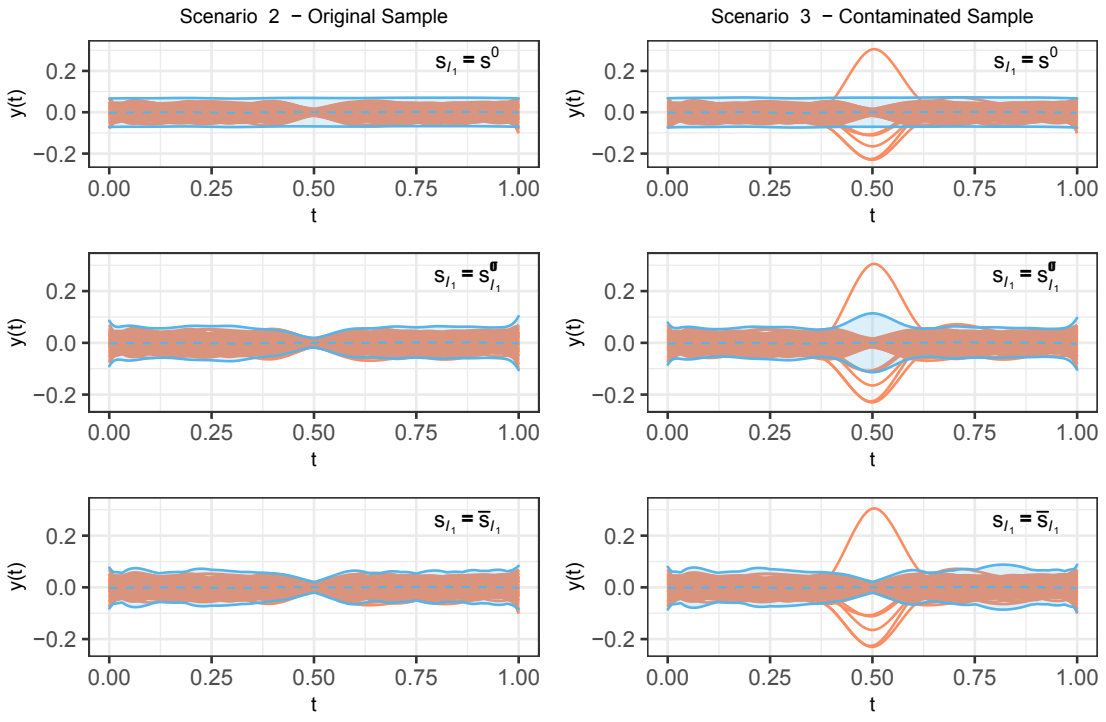

FIGURE 2.3: The prediction bands obtained considering a combination of modulation functions (s^0 at the top, $s^{\sigma}_{\mathcal{I}_1}$ in the middle, $\bar{s}_{\mathcal{I}_1}$ at the bottom) and sample (the original one on the left, the contaminated one on the right). In all cases, the dashed line represents $g_{\mathcal{I}_1}$.

obtained by $s^{\sigma}_{\mathcal{I}_1}$ and $\bar{s}_{\mathcal{I}_1}$ and it forces the band to be unnecessary large around $t = 0.5$. On the other hand, the other two modulation functions (especially $s^{\sigma}_{\mathcal{I}_1}$) provide a better performance since they allow the band width to be adapted according to the behavior of data over $\mathcal{T}$.

Scenario 3 is obtained by contaminating Scenario 2 with outliers. Table 2.2 suggests that $\bar{s}_{\mathcal{I}_1}$ outperforms both s^0 and - unlike Scenario 2 - also $s^{\sigma}_{\mathcal{I}_1}$. In order to clarify this evidence, let us consider a sample $y_1, \ldots, y_{198}$ generated as in Scenario 2 that, after being created, is exposed to a contamination process in which each function $y_i, i = 1, \ldots, 198$, becomes an outlier as described in Scenario 3 with probability $\beta = 0.06$. Figure 2.3 shows examples of prediction bands induced by the three modulation functions (s^0 at the top, $s^{\sigma}_{\mathcal{I}_1}$ in the middle, $\bar{s}_{\mathcal{I}_1}$ at the bottom) obtained by considering the original sample (on the left) and the contaminated one (on the right). Moving from Scenario 2 to Scenario 3 and focusing on $s^{\sigma}_{\mathcal{I}_1}$, it is possible to notice that the increased variability in the central part of the domain due to the contamination process involves an increase in the band width around $t = 0.5$. This behavior, although not surprising, is counterproductive since the purpose of the method is to create prediction bands with coverage at the level $1 - \alpha = 0.9$ and in this specific case $\sim 94\%$ of the functions tends to be highly

concentrated around $g_{\mathcal{I}_1}$ in the central part of the domain, and not overdispersed. By contrast, $\bar{s}_{\mathcal{I}_1}$ by construction removes the most extreme (in terms of measure (2.1)) functions and properly modulates data on the basis of the non-extreme functions keeping the band shape unchanged. From a methodological point of view, this is due to the dependency of $\bar{s}_{\mathcal{I}_1}$ on α which allows only a portion of the training set - chosen according to the specific level $1-\alpha$ - to be taken into account and the trend of the "misleading" functions to be completely ignored. Overall, the evidence provided by this example - together with the results provided by Table 2.2 - suggests that s^0 is not affected by the contamination process (pro) but does not modulate (con), $s^{\sigma}_{\mathcal{I}_1}$ modulates (pro) but overreacts to the contamination process (con), whereas $\bar{s}_{\mathcal{I}_1}$ is able to simultaneously modulate (pro) and manage the contamination process (pro).

In short, the three scenarios seem to highlight that s^0 is an outstanding candidate when the sample size is very small, whereas a modulation process is useful in the very common case in which the variability over $\mathcal{T}$ varies and the sample size is either moderate or large. Specifically, $\bar{s}_{\mathcal{I}_1}$ provides encouraging results in some complex scenarios as it focuses on the specific behavior of the central (according to the level $1-\alpha$) portion of data.

2.4 Application

In order to show the wide generality of our approach, in this section we apply our Conformal approach to a well known data set in the FDA community (i.e., the Berkeley Growth Study data set Tuddenham and Snyder, 1954) that is characterized by features that cannot be trivially framed in a standard probabilistic parametric model, i.e.: heteroscedasticity along the functional domain, phase misalignment, presence of outlier curves, and positivity constraint. The specific data set contains in detail the heights (in cm) of 54 female and 39 male children measured quarterly from 1 to 2 years, annually from 2 to 8 years and biannually from 8 to 18 years. We focus on the first derivative of the growth curves, which are estimated in a standard fashion by R function *smooth.monotone* of *fda* package (Ramsay *et al.*, 2020) implementing monotonic cubic regression splines (Ramsay and Silverman, 2005, chap. 6). Specifically, the prediction bands here reported refer to the growth velocity curves between 4 and 18 years for girls and boys separately comparing, in the Split Conformal framework, the three modulation functions analyzed in Section 2.3 and with $g_{\mathcal{I}_1}$ being simply for each group the corresponding functional sample mean, $\alpha = 0.5$, $m = 27$ for girls, $m = 20$ for boys.

The prediction bands are shown in Figure 2.4. Note that since the application at hand

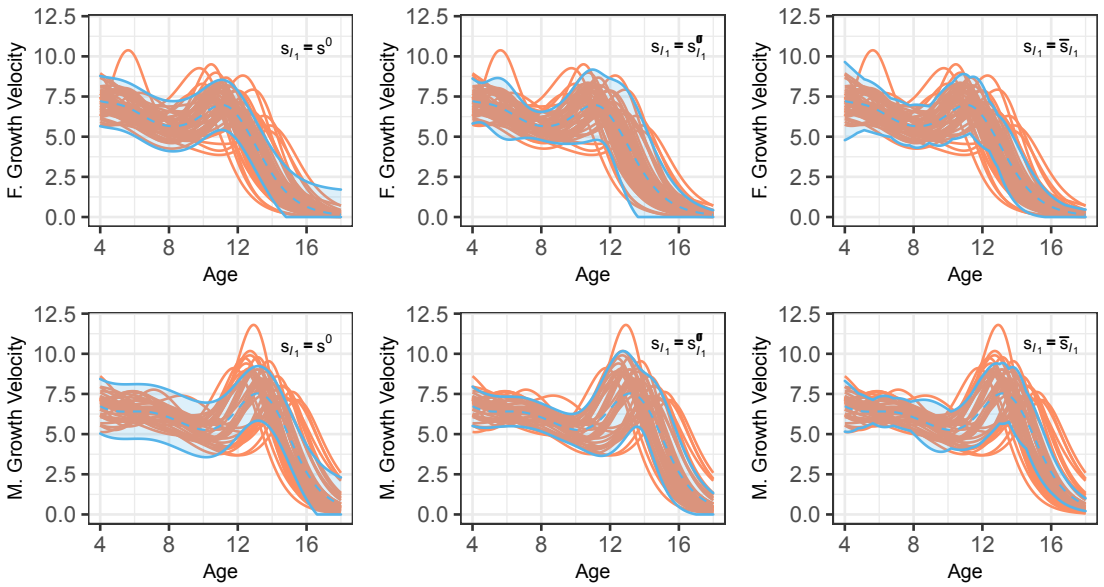

FIGURE 2.4: Berkeley Growth Study data: each panel shows the prediction band obtained considering a different modulation function (s^0 on the left, $s^{\sigma}_{\mathcal{I}_1}$ in the middle, $\bar{s}_{\mathcal{I}_1}$ on the right). In all cases, the dashed line represents $g_{\mathcal{I}_1}$. Predictions for girls at the top and predictions for boys at the bottom.

does not allow the functions to be negative in any subset of the domain, the prediction bands can be (and are indeed) truncated to 0 without decreasing their coverage.

Focusing on Figure 2.4, the graphical representation of the prediction bands highlights the well-known different growth path between girls and boys, in which the latter group typically starts to grow later but achieves higher growth velocities. In terms of the role of modulation functions, their impact on female growth velocity prediction seems to be less than the one on the male bands. From a prediction point of view, girls' curves represent a simpler scenario in which the variance is lower along the domain, while boys' curves represent a more tricky scenario with strong heteroscedasticity of the functions over $\mathcal{T}$ (due to the joint presence of misalignment of data and a very localized high peak around 13 years of age). As expected from these considerations, the prediction bands for a new girl's velocity curve obtained using the different modulation functions are relatively similar, with the prediction band associated to $\bar{s}_{\mathcal{I}_1}$ being aslightly narrower due to the presence of outliers. Instead focusing on boys' curves, the strong heteroscedasticity forces the prediction band induced by s^0 to be uselessly large in some parts of the domain, whereas in general the prediction band induced by $s^{\sigma}_{\mathcal{I}_1}$ seems to be smoother than that induced by $\bar{s}_{\mathcal{I}_1}$, whose "bumps" are caused by the specific modulation function used. Both for boys and girls $\bar{s}_{\mathcal{I}_1}$ outputs the smallest prediction band, as shown in Table 2.3 where the quantity $\mathcal{Q}(\cdot)/|\mathcal{T}|$ is reported. As mentioned at the beginning

	s^0	$s^{\sigma}_{\mathcal{I}_1}$	$\bar{s}_{\mathcal{I}_1}$
Females	2.904	3.244	2.811
Males	3.334	3.107	2.690

TABLE 2.3: Berkeley Growth Study data: average width of the prediction bands.

of the section, the results shown are obtained without aligning the raw data since the aim is to predict the actual growth curve of a new individual. However, it is interesting to notice that the development of the procedure with aligned data may have an impact on the size and shape of the obtained prediction bands (e.g. by creating smaller prediction bands due to the decrease in data variability) as long as the alignment procedure maintains the assumptions underlying the Conformal approach.

2.5 Conclusion

The creation of prediction sets for univariate i.i.d. functional data is still an open problem of paramount importance in statistical methodology research. In order to define and compute them, the great majority of methods currently presented in the literature rely on non-provable distributional assumption, dimension reduction techniques and/or asymptotic arguments. On the contrary, the approach proposed in this chapter represents an innovative proposal in this field: indeed, the Conformal framework ensures that finite-sample either valid or exact prediction sets are obtained under minimal distributional assumptions, whereas the specific family of nonconformity measures introduced guarantees - besides prediction sets that are bands - also a fast, scalable and closed-form solution. Moreover, despite the fact that our approach works regardless the specific choice of $s_{\mathcal{I}_1}$ (which can be chosen, for example, a priori), we proposed a specific data-driven modulation function, namely $\bar{s}_{\mathcal{I}_1}$, which leads to prediction bands asymptotically no less efficient than those obtained by not modulating.

Our procedure is able to achieve encouraging results and could represent a promising starting point for future developments, but at least two aspects, among others, should be carefully investigated. First of all, the division of data into the training and calibration sets induces an intrinsic element of randomness into the method and, although this phenomenon is well known in the Conformal literature, a quantification of the effect of the split process - and also of the values m and l - on the procedure has not yet been properly analyzed. Secondly, the prediction sets proposed in this chapter are purposely shaped as functional bands. This geometrical characterization in most application scenarios can be considered well suited. Nevertheless, one can think at more complicated

scenarios (e.g., functional mixtures) where prediction set made of multiple bands could be considered more suited from an application point of view. This possible extension will be the object of future work.

Chapter 3

Prediction bands for multivariate functional data in a regression framework

3.1 Introduction

Functional Data Analysis (FDA, Ramsay and Silverman, 2005) is a fairly established field of statistics whose goal is to develop theory and methods to treat datasets composed of smooth functions. Since the first seminal paper by Jim O. Ramsay (Ramsay, 1982), the research concerning FDA has focused on a wide range of topics: classification, linear regression, functional depth and nonparametric techniques are only a few examples of up-to-date fields of research, and a detailed presentation of some of the most modern topics concerning FDA can be found, for example, in Goia and Vieu (2016); Aneiros *et al.* (2019).

Among them, various inferential tools have been developed to deal with functional data. Simultaneous confidence bands (for recent contributions see, e.g., Choi and Reimherr, 2018; Telschow and Schwartzman, 2022) and related testing procedures such as envelope tests (see, e.g., Myllymäki *et al.*, 2017; Lopez-Pintado and Qian, 2021) represent key instruments to make inferential conclusions on functional parameters. Closely related to these procedures (as suggested by the work Liebl and Reimherr, 2019) is the creation of prediction sets, namely subsets of the sample space including a new functional observation with a certain nominal confidence level $1-\alpha$. Some very recent works in FDA provide some knowledge into this theoretical (but yet full of applied repercussions) issue. A first group of approaches consists of works principally based on parametric bootstrapping techniques (e.g., Degras, 2011; Cao *et al.*, 2012), and a

second one is characterized by the application of dimensionality reduction techniques to manage the naturally infinite dimensionality (e.g., Hyndman and Shahid Ullah, 2007; Antoniadis *et al.*, 2016). These first two groups carry obvious drawbacks since they are either based on not easily provable distributional assumptions and/or on asymptotic results. In addition, the first class of approaches is computationally demanding, whereas the second one relies on the approximations induced by basis projection. A third group is based on the novel approach to forecasting in the framework of Conformal Prediction presented in Chapter 2. This approach is able to output either exact — i.e., ensuring a coverage equal to $1-\alpha$ — or valid — i.e., ensuring a coverage no less than $1-\alpha$ — prediction bands under minimal distributional assumptions and in an efficient way, thus bypassing the aforementioned methodological shortcomings. However, this is done in the setting of univariate i.i.d. functional data. The objective of the present work is to complement the results obtained in Chapter 2 by extending the method to multivariate functional data and to a regression framework. As will become clearer below, Conformal Prediction is strongly connected to another ebullient field of research in FDA, namely the functional depth, with which it shares the goal of evaluating the centrality/conformity of a given functional observation with respect to a group of observed functions. Actually, functional depths (see, e.g., Nagy, 2016; Gijbels and Nagy, 2017, for some recent contributions on the topic) can be used directly — i.e., without considering the Conformal Prediction framework — to construct prediction sets in a very intuitive and simple way, but in doing so no theoretical guarantee in terms of coverage is obtained. In order to clarify this aspect, a simulation study comparing the approach proposed in this work with the $(1-\alpha)\cdot 100\%$ central region induced by a specific functional depth is provided in Section 3.3.

Formally, we will consider independent and identically distributed regression pairs $\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_n \sim P$, with $\boldsymbol{Z}_i = (X_i, \boldsymbol{Y}_i)$ consisting of a multivariate functional response variable $\boldsymbol{Y}_i$ and a set of (not necessarily scalar) covariates X_i $i \in \{1,\dots,n\}$. Let $\boldsymbol{Y}_i = (Y_{i,1},\dots,Y_{i,p})$ be a multivariate random function such that its j-th component $Y_{i,j}$ ($j \in \{1,\dots,p\}$) is a random function taking values in $L^\infty(\mathcal{T}_j)$, which is the family of bounded functions $y : \mathcal{T}_j \to \mathbb{R}$ with $\mathcal{T}_j$ compact subset of $\mathbb{R}^{d_j}$, $d_j \in \mathbb{N}_{>0}$, with $\mathbb{N}_{>0}$ the set of positive integers. For the sake of brevity, later in the discussion we will indicate the space $L^\infty(\mathcal{T}_1)\times\cdots\times L^\infty(\mathcal{T}_p)$ in which $\boldsymbol{Y}_i$ takes values as $\prod_{j=1}^p L^\infty(\mathcal{T}_j)$. Note that the framework considered is extremely wide since both the domain $\mathcal{T}_j$ and the image of $Y_{i,j}$ (i.e., the subset of $L^\infty(\mathcal{T}_j)$ made up of all the possible realizations of $Y_{i,j}$) are allowed to be very different when j varies. X_i is the set made up of all the covariates related to $Y_{i,1},\dots,Y_{i,p}$ (although each $Y_{i,j}$ need not necessarily depend on the whole

set of covariates X_i). It belongs to a measurable space and can be very general: for example, X_i can be a usual vector of predictors, or it can be a set of functional covariates allowing for a function-on-function regression model, or it can contain both scalar and functional predictors. Let $\mu^j(x_i) = E[Y_{i,j}|X_i = x_i]$ denote the regression function for the j-th component of the i-th observation, and consistently with this notation let us define the scalar value $[\mu^j(x_i)](t) = E[Y_{i,j}(t)|X_i = x_i]$.

The aim of the chapter is to build a procedure able to output valid/exact multivariate functional prediction bands under no assumptions on P and $\mu^1(\cdot), \ldots, \mu^p(\cdot)$ other than i.i.d. regression pairs. A multivariate functional prediction band is a specific kind of prediction set that can be defined, consistently with the definition of univariate functional prediction band provided in the Introduction of the thesis, as

$$\left\{ \boldsymbol{y} = (y_1, \ldots, y_p) \in \prod_{j=1}^{p} L^\infty(\mathcal{T}_j) : y_j(t) \in B_j(t), \quad \forall j \in \{1, \ldots, p\}, \quad \forall t \in \mathcal{T}_j \right\},$$

with $B_j(t)$ an interval $\forall j, t$. Prediction bands are so relevant in the functional set prediction framework — despite their inability to reflect the shape properties of data (Nagy *et al.*, 2017) — due to their conceptual simplicity and because they can be plotted in parallel coordinates (Inselberg, 1985). For the sake of simplicity, later in the chapter the term prediction band will be used to indicate a multivariate functional prediction band, unless otherwise specified.

The chapter is organized as follows: in Section 3.2 we present the method developed; in Section 3.3 we discuss two simulation studies aimed at investigating different aspects of the method; in Section 3.4 we apply our method to a real-world application; in Section 3.5 we provide an overview of the main findings and sketch directions of future research.

3.2 Methods

3.2.1 The Nonconformity Measure

Let us consider the Split Conformal method presented in Chapter 1.1. Moving from the results obtained in Chapter 2 and inspired by a depth of infimal type (Mosler, 2013) based on the univariate projection depth (Zuo, 2003), we propose the following

nonconformity measure and nonconformity scores:

$$A(\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}, \tilde{\boldsymbol{z}}) = \sup_{j\in\{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t\in\mathcal{T}_j} \left| \frac{\tilde{y}_j(t) - [\hat{\mu}^j(\tilde{x})](t)}{s_j(t)} \right| \right), \tag{3.1}$$

$$R_d = \sup_{j\in\{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t\in\mathcal{T}_j} \left| \frac{y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)}{s_j(t)} \right| \right), \quad d \in \mathcal{I}_2, \tag{3.2}$$

$$R_{n+1} = \sup_{j\in\{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t\in\mathcal{T}_j} \left| \frac{y_j(t) - [\hat{\mu}^j(x_{n+1})](t)}{s_j(t)} \right| \right),$$

with $\tilde{\boldsymbol{z}} = (\tilde{x}, \tilde{\boldsymbol{y}})$, $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \dots, \tilde{y}_p)$, y_j the j-th component of $\boldsymbol{y}$, $[\hat{\mu}^j(x_d)](t)$ estimate of $[\mu^j(x_d)](t)$ based on $\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}$, $s = \{s_j\}_{j=1}^p$ set of modulation functions with $s_j : \mathcal{T}_j \to \mathbb{R}_{>0}$ a (strictly positive) function belonging to $L^\infty(\mathcal{T}_j)$ based on $\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}$ called modulation function. In doing so, the simpler notation $[\hat{\mu}^j(x_d)](t)$ (s_j, respectively) is used instead of $[\hat{\mu}^j_{\mathcal{I}_1}(x_d)](t)$ ($s_{j,\mathcal{I}_1}$, respectively) to facilitate readability. It is fundamental to notice that no specific assumptions are made on the estimators $[\hat{\mu}^1(\cdot)](t), \dots, [\hat{\mu}^p(\cdot)](t)$ (considered in this case as random variables instead of observed values) since the Conformal framework only requires the nonconformity scores R_d and R_{n+1} to be computed on the basis of the observations belonging to the training set and on $\boldsymbol{z}_d$ and $(x_{n+1}, \boldsymbol{y})$ respectively. As a consequence, finite-sample, either valid or exact prediction sets are obtained regardless the choice of the regression estimators, allowing Conformal Inference to be satisfactorily performed also when the underlying model is completely misspecified.

By considering the Split Conformal method and the nonconformity measure (3.1), if $\alpha \in (0, 1/(l+1))$ then $\mathcal{C}_{n,1-\alpha}(x_{n+1}) = \prod_{j=1}^p L^\infty(\mathcal{T}_j)$ since $\delta_{\boldsymbol{y}}$ is always greater or equal than $1/(l+1)$. If $\alpha \in [1/(l+1), 1)$ (representing the scenario on which we will focus on hereafter), then

$$\mathcal{C}_{n,1-\alpha}(x_{n+1}) := \Bigg\{ \boldsymbol{y} \in \prod_{j=1}^p L^\infty(\mathcal{T}_j) : y_j(t) \in \big[[\hat{\mu}^j(x_{n+1})](t) - k^s \cdot s_j(t), [\hat{\mu}^j(x_{n+1})](t) + k^s \cdot s_j(t) \big], \\ \forall j \in \{1, \dots, p\}, \forall t \in \mathcal{T}_j \Bigg\},$$

with k^s the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d : d \in \mathcal{I}_2\}$, and the superscript s introduced in order to emphasize the dependence of k^s on s. The computation needed to find analytically $\mathcal{C}_{n,1-\alpha}(x_{n+1})$ is provided in Appendix C.1, together with the definition of $\mathcal{C}_{n,1-\alpha,\tau_{n+1}}(x_{n+1})$, i.e., the Smoothed Split Conformal prediction set induced by the nonconformity measure (3.1).

From a practical point of view, first of all the observed sample $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ is used to

compute k^s and $s_1, \dots, s_p$, and after that the prediction set is built around the regression estimates $[\hat{\mu}^j(x_{n+1})](t)$, $j \in \{1, \dots, p\}$. Despite the fact that no specific constraints on $[\hat{\mu}^j(\cdot)](t)$ are required by the Split Conformal framework, the choice of the regression estimators is fundamental in providing small prediction sets, a key topic that will be investigated in Section 3.2.2: indeed, intuitively one is justified in expecting prediction sets to be smaller when improved regression estimators are chosen since they typically provide smaller nonconformity scores and so a smaller value of k^s (Lei *et al.*, 2018). However, later in the discussion (and specifically in Section 3.3 and Section 3.4) we will always consider the regression estimators as given by the application at hand: in fact, our aim is to construct valid/exact prediction sets in general and arbitrary prediction scenarios and not only in specific, well informed frameworks.

Under the exchangeability assumption of the regression pairs and regardless the choice of s and $[\hat{\mu}^j(\cdot)](t)$, the prediction sets induced by the nonconformity measure (3.1) are:

- either finite-sample valid (Split Conformal method) or finite-sample exact (Smoothed Split Conformal method) for any distribution P;
- in closed form;
- bands;
- scalable. Indeed, conditional on the computational cost required to calculate the regression estimates and the set of modulation functions (a set that can be chosen to be computationally parsimonious), and by keeping the ratio l/n fixed when n grows, the time required to compute k^s (and therefore to output the prediction set) increases linearly with l, and so linearly with n.

Note that nonconformity measure (3.1) ensures multivariate simultaneous bands, i.e., bands guaranteeing the desired coverage globally (i.e., for the multivariate random function $\boldsymbol{Y}_{n+1}$). Proper multivariate simultaneous coverage represents a leap forward with respect to univariate simultaneous coverage (i.e., coverage holding for $Y_{n+1,j}$) and pointwise coverage (i.e., coverage holding for $Y_{n+1,j}(t)$).

Alongside the choice to base the nonconformity measure on the essential supremum, the set s of (strictly positive) modulation functions s_j represents the core of our approach. First of all, one can notice that prediction bands induced by $\{s_j\}_{j=1}^p$ and by $\{\lambda \cdot s_j\}_{j=1}^p$ coincide $\forall \lambda \in \mathbb{R}_{>0}$ (see Appendix C.1 for the proof), and so later in the discussion we will consider, for any equivalence class, the set of modulation functions such that $\sum_{j=1}^p \int_{\mathcal{T}_j} s_j(t)dt = 1$. In the next section, we detail the role of s by highlighting its

impact on the efficiency (i.e., the size) of the prediction bands and we propose a specific set of modulation functions able to guarantee an asymptotic result in terms of efficiency.

3.2.2 The Choice of the Set of Modulation Functions

Intuitively, in addition to the appealing properties presented in Section 3.2.1, a prediction band should modulate its width over $\mathcal{T}_1, \ldots, \mathcal{T}_p$ according to the local variability of the data. Specifically, the aim is to obtain prediction bands able to properly manage the fact that: focusing on the j-th component, the pointwise evaluations of functional data may be characterized by highly different variability when $t \in \mathcal{T}_j$ varies; the p components may be characterized by different magnitude. In order to achieve these two purposes, a careful choice of a data-driven set of modulation functions s is recommended. In order to clarify this concept, let us consider the following example: let $p = 2$ with $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{200}$ independent realizations of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{200}$ such that $Y_{i,1}(t) = \beta_1(t) + \varepsilon_{i,1}(t)$ and $Y_{i,2}(t) = \beta_2(t) + \varepsilon_{i,2}(t)$, $i \in \{1, \ldots, 200\}$, $\mathcal{T}_1 = \mathcal{T}_2 = [0, 1]$, with the systematic components defined simply as $\beta_1(t) = 1$, $\beta_2(t) = 0$ $\forall t \in [0, 1]$ and the independent functional error components $\{\varepsilon_{i,1}\}_{i=1}^{200}$ ($\{\varepsilon_{i,2}\}_{i=1}^{200}$ respectively) obtained by means of a B-spline basis expansion (Fourier basis expansion respectively) with normally distributed random vectors as coefficients. In full generality, we consider $[\hat{\mu}^j(x_{n+1})](t) = \hat{\beta}_j(t)$, $j \in \{1, 2\}$, with $\hat{\beta}_1(t), \hat{\beta}_2(t)$ the estimates, based on $\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}$, obtained by fitting the two concurrent function-on-function linear models (Ramsay and Silverman, 2005). This example represents the simplest, almost trivial regression scenario which allows to — hopefully — easily understand the crucial role of s, but the discussion presented hereafter naturally holds also when decidedly more complex regression functions and regression estimators are taken into account. Fig. 3.1 shows the multivariate prediction band for $\boldsymbol{Y}_{201} = (Y_{201,1}, Y_{201,2})$ obtained by considering two different sets of modulation functions: the two panels at the top of the Fig. 3.1 show the multivariate prediction band obtained by not modulating (i.e., by setting $s_1(t) = s_2(t) = 1/\sum_{j=1}^{2} |\mathcal{T}_j| = 1/2 \propto 1$ $\forall t \in [0, 1]$, with $|\cdot|$ the Lebesgue measure), whereas the two panels at the bottom of the same figure show the prediction band obtained by considering the two standard deviation functions of the functional residuals as modulation functions (after normalization in order to meet the condition $\sum_{j=1}^{2} \int_{\mathcal{T}_j} s_j(t)dt = 1$). In order to distinguish the two sets of modulation functions, later in the discussion we will denote the first set by $s^0 := \{s_j^0\}_{j=1}^p$ and the second one by $s^\sigma := \{s_j^\sigma\}_{j=1}^p$. The prediction sets are obtained by considering the Split Conformal framework and by setting $\alpha = 0.25$, $m = l = 100$. Focusing on the two panels at the top of Figure 3.1, it is possible to notice that the two univariate prediction bands are far from desirable: specifically, the univariate prediction

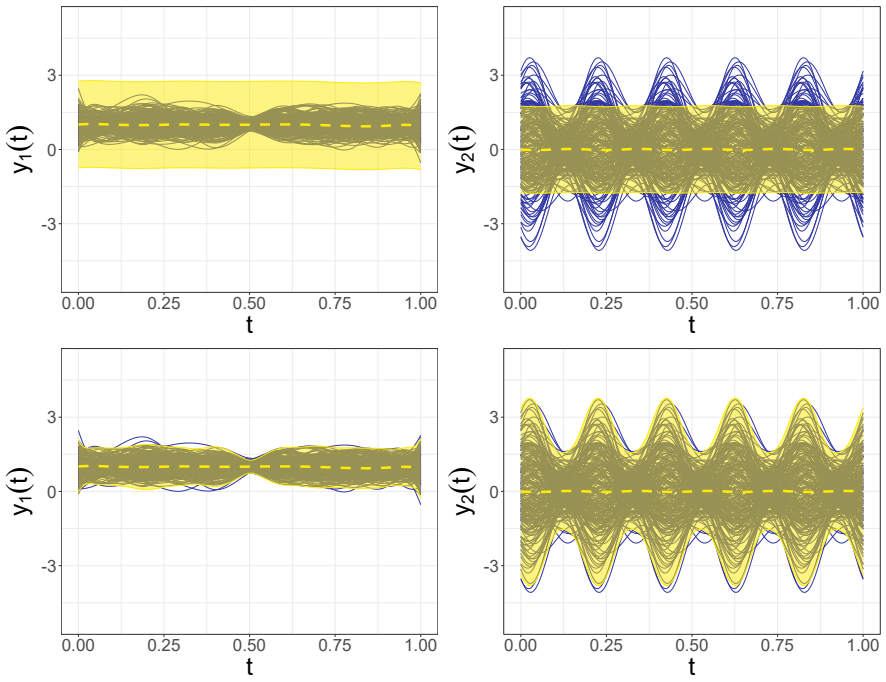

FIGURE 3.1: Split Conformal multivariate prediction band for $\boldsymbol{Y}_{201} = (Y_{201,1}, Y_{201,2})$ obtained by considering $\{s_j^0\}_{j=1}^2$ (at the top) and $\{s_j^\sigma\}_{j=1}^2$ (at the bottom) as set of modulation functions. The dashed yellow lines represent the regression estimates. $\alpha = 0.25$, $n = 200$, $m = l = 100$.

band related to $Y_{201,1}$ is large along all the domain $\mathcal{T}_1$, whereas the one related to $Y_{201,2}$ contains almost all the pointwise evaluations of the functional data in the low-variance parts of $\mathcal{T}_2$ but excludes many pointwise evaluations in the other, high-variance parts of the domain. In this specific case, the absence of a modulation process does not allow to take into account: first of all, the different variability of the data over $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively; secondly, the different magnitude that characterizes the two components. In doing so, one is justified in expecting that a procedure based on $\{s_j^0\}_{j=1}^2$, although able to output a valid prediction band, may be of limited practical use in real applications. Vice versa, the set of modulation functions $\{s_j^\sigma\}_{j=1}^2$ properly adapts the width of the prediction band according to the local variability of functional data, allowing for a meaningful, interpretable and useful prediction band.

Beyond these common-sense considerations, a criterion that is both reasonable and well-established in Conformal Prediction to discriminate between procedures able to guarantee validity is the minimization of the size of the prediction sets outputted (also known, in the Conformal framework, as maximization of efficiency, Balasubramanian *et al.*, 2014): this choice is due to the fact that desirable prediction sets should include subsets of the sample space where the probability mass is highly concentrated (Lei *et al.*, 2013). In the context of the thesis, the aim would be to find the nonconformity measure

$A(\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}, \cdot)$ (and so, practically, the set of modulation functions s) inducing the smallest prediction bands. The first, fundamental step in assessing the size of a prediction band for multivariate functional data is the definition of the concept of 'size', a nontrivial task if compared to the traditional univariate and multivariate statistical settings. By generalizing the definition given in Chapter 2 to the multivariate case, we define the size of a multivariate prediction band as the sum of the p areas between the upper and lower bound of the p univariate prediction bands:

$$\mathcal{Q}(s) := \sum_{j=1}^{p} \int_{\mathcal{T}_j} 2 \cdot k^s \cdot s_j(t)dt = 2 \cdot k^s, \tag{3.3}$$

where the equality is due to the fact that $\sum_{j=1}^{p} \int_{\mathcal{T}_j} s_j(t)dt = 1$. Since $\mathcal{Q}(s)$ is a random variable depending on $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, the task of finding the set of modulation functions minimizing the risk functional $E[\mathcal{Q}(s)]$ is unfeasible in the case of no assumptions on P. A simplification of such a complex task consists of considering the quantity to be minimized $k^s (\propto \mathcal{Q}(s))$ as an observed value depending on $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ instead of on $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ according to the empirical risk minimization principle (Vapnik, 1992). In doing so, the optimization problem is certainly simplified, but its resolution still remains unfeasible due to the specific structure of k^s. Indeed, k^s is a specific empirical quantile of $\{R_d : d \in \mathcal{I}_2\}$, and R_d (see Equation 3.2) depends by construction both on the training set through $\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}$ and on the calibration set through $\boldsymbol{z}_d$. Since by construction the set of modulation functions s depends only on the training set (as its dependence on the calibration set would imply not to obtain closed-form valid prediction bands), no rule minimizing k^s only by combining the elements of the training set (i.e., by varying s) can be found for general $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$.

In view of this, we propose an alternative, unconventional strategy to build a set of modulation functions able to guarantee an asymptotic result in terms of efficiency. Specifically, the purpose is to find a couple of sets of functions $(\bar{s}, \bar{s}^c)$ such that:

- $\bar{s}^c := \{\bar{s}_j^c\}_{j=1}^p$ is a set of functions such that $\bar{s}_j^c$ meets the definition of modulation function, but depends also on the calibration set through $\{\boldsymbol{z}_d : d \in \mathcal{I}_2\}$.
- prediction bands obtained by using $\bar{s}^c$ as set of modulation functions are smaller than or equal to (in terms of Equation 3.3) those induced by the set of modulation functions s^0 for every possible value of n and for every possible observed sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$.
- $\bar{s} = \{\bar{s}_j\}_{j=1}^p$ is a set of modulation functions such that $\bar{s}_j^c$ and $\bar{s}_j$ converge to the same function when $m, l \to +\infty$, $\forall j \in \{1, \ldots, p\}$.

In doing so, prediction bands induced by the set of modulation functions $\bar{s}$ are characterized by all the appealing properties presented in Section 3.2.1 (including validity) and are asymptotically not wider than those induced by s^0 regardless the specific sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. From the operational point of view, a natural candidate for $\bar{s}_j$ is to consider $\bar{s}_j^c$ and to replace its dependence on $\{\boldsymbol{z}_d : d \in \mathcal{I}_2\}$ with the dependence on $\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}$, and consequently to check their convergence to the same function. In order to find $(\bar{s}, \bar{s}^c)$ satisfying the aforementioned conditions, let us consider the structure of k^s: operationally, k^s computes a summary of the multivariate functional residual for every observation in the calibration set, and selects the $\lceil (l+1)(1-\alpha) \rceil$th smallest value among them. In particular: the summary is naturally induced by the specific nonconformity measure used, which searches the greatest value of the absolute value of the modulated multivariate functional residual over the p domains $\mathcal{T}_1, \ldots, \mathcal{T}_p$; k^s is not affected by the $l - \lceil (l+1)(1-\alpha) \rceil$ greatest values of $\{R_d : d \in \mathcal{I}_2\}$. In view of this, a proper candidate for $\bar{s}^c$ should ignore the elements of $\{\boldsymbol{z}_d : d \in \mathcal{I}_2\}$ leading to the $l - \lceil (l+1)(1-\alpha) \rceil$ greatest values of $\{R_d : d \in \mathcal{I}_2\}$ and should modulate data based on the most extreme value observed $\forall t \in \mathcal{T}_j, j \in \{1, \ldots, p\}$.

Therefore, the couple of sets of functions $(\bar{s}, \bar{s}^c)$ we propose — which represents a generalization of the finding of Chapter 2, and which induces nonconformity scores related to the notion of infimal depth introduced by Narisetty and Nair (2016) — is defined below. Formally, the set of functions $\bar{s}^c$ is such that for $j \in \{1, \ldots, p\}, t \in \mathcal{T}_j$:

$$\bar{s}_j^c(t) := \frac{\max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|}{\sum_{j=1}^p \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| dt}$$

with

$$\mathcal{H}_2 := \left\{ d \in \mathcal{I}_2 : \sup_{j \in \{1,\ldots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| \right) \leq k \right\}$$

and $k = k^{s^0} / \sum_{j=1}^p |\mathcal{T}_j|$ the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set

$$\left\{ \sup_{j \in \{1,\ldots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) : d \in \mathcal{I}_2 \right\}.$$

For the sake of simplicity, we assume $\max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| \neq 0 \; \forall j \in \{1, \ldots, p\}, t \in \mathcal{T}_j$. If this condition does not hold for at least one couple (t, j) but the condition $\sum_{j=1}^p \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| dt \neq 0$ still holds, in order to have that $\bar{s}_j^c(t) > 0$, $j \in \{1, \ldots, p\}, t \in \mathcal{T}_j$ it is sufficient to add a small, positive value to $\bar{s}_j^c(t)$ and to normalize accordingly. Vice versa, the case in which $\sum_{j=1}^p \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| dt = 0$ represents a pathological case of no practical interest.

The set of modulation functions $\bar{s}$ is such that for $j \in \{1, \dots, p\}, t \in \mathcal{T}_j$:

$$\bar{s}_j(t) := \frac{\max_{h \in \mathcal{H}_1} |y_{h,j}(t) - [\hat{\mu}^j(x_h)](t)|}{\sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{h \in \mathcal{H}_1} |y_{h,j}(t) - [\hat{\mu}^j(x_h)](t)| dt}$$

with $\mathcal{H}_1 = \mathcal{I}_1$ if $\lceil (m+1)(1-\alpha) \rceil > m$, otherwise

$$\mathcal{H}_1 := \left\{ h \in \mathcal{I}_1 : \sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} |y_{h,j}(t) - [\hat{\mu}^j(x_h)](t)| \right) \leq \gamma \right\}$$

and γ the $\lceil (m+1)(1-\alpha) \rceil$th smallest value in the set

$$\left\{ \sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left| y_{h,j}(t) - [\hat{\mu}^j(x_h)](t) \right| \right) : h \in \mathcal{I}_1 \right\}.$$

If $\exists (t, j)$ such that $\max_{h \in \mathcal{H}_1} |y_{h,j}(t) - [\hat{\mu}^j(x_h)](t)| = 0$, the adjustment used for $\bar{s}_j^c$ is implemented.

Specifically, the fact that the set of modulation functions $\bar{s}$ depends on α (through γ) allows for a procedure able to modulate data according to the specific value $1 - \alpha$, i.e., the desired nominal coverage. In addition, such an unconventional set of modulation functions is particularly useful when functional residuals show a non-standard behavior (e.g., there are outliers). The following two theorems show that $(\bar{s}, \bar{s}^c)$ satisfies the aforementioned conditions.

Theorem 3.1. *Let $m/n = \theta$ with $0 < \theta < 1$ and let $\operatorname{Var}\left[[\hat{\mu}^j(X_i)](t) \right] \to 0 \ \forall i \in \{1, \dots, n\}$, $\forall t \in \mathcal{T}_j$, $\forall j \in \{1, \dots, p\}$ when $m \to +\infty$. Then $\bar{s}_j^c$ and $\bar{s}_j$ converge to the same function, $j \in \{1, \dots, p\}$, when $n \to +\infty$.*

Theorem 3.2. *If at least one of the functions $\{\bar{s}_j^c(t)\}_{j=1}^p$ is not constant almost everywhere over its domain, then $\mathcal{Q}(s^0) > \mathcal{Q}(\bar{s}^c)$. Otherwise, $\mathcal{Q}(s^0) = \mathcal{Q}(\bar{s}^c)$.*

See C.2 for both proofs, together with the generalization of $(\bar{s}, \bar{s}^c)$, Theorem 3.1 and Theorem 3.2 to the Smoothed Split Conformal framework. Due to the very mild conditions required by the two theorems to hold, the set of modulation functions $\bar{s}$ can be used in many general frameworks and provides a new, we believe appealing data-driven alternative to other solutions (e.g., s^σ). In the next section, the set of modulation functions $\bar{s}$ is compared to other sets of modulation functions in different simulated scenarios.

3.3 Simulation Study

In this section we perform two simulation studies aimed at evaluating different practical aspects of the method presented in Section 3.2. In Section 3.3.1, the empirical coverage provided by the prediction bands induced by our method is evaluated in different scenarios, considering different sample sizes and different kinds of model misspecification, and the results are compared with those obtained by considering the corresponding depth-based approach. In Section 3.3.2, the three sets of modulation functions presented in Section 3.2 ($\{s_j^0\}_{j=1}^p$, $\{s_j^\sigma\}_{j=1}^p$, $\{\bar{s}_j\}_{j=1}^p$) are compared in terms of efficiency in order to highlight their strengths and weaknesses.

In both simulation studies, some quantities are kept fixed: $p = 2$, $\mathcal{T}_1 = \mathcal{T}_2 = [0, 1]$, $\alpha = 0.10$. Three possible sample sizes are taken into account: $n = 20, n = 200, n = 2000$. We focus on the Split Conformal method and since the coverage reached by Split Conformal prediction set is $1 - \lfloor (l+1)\alpha \rfloor / (l+1)$ (see Theorem 1.1), the size of the calibration set is set equal to $l = 9, l = 99, l = 999$ respectively in order to obtain $1 - \lfloor (l+1)\alpha \rfloor / (l+1) = 1 - \alpha$ and consequently to facilitate the readability of the results. A possible alternative would be to consider a different value of l (e.g., $n/2$) and to evaluate the empirical coverage taking into account the coverage $1 - \lfloor (l+1)\alpha \rfloor / (l+1)$. Each combination of simulation study, scenario, method, sample size, regression estimators and set of modulation functions is evaluated based on $N = 5000$ replications. Specifically, for each replication, a sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n+1}$ is generated and n randomly chosen elements are assigned to the training and calibration sets, whereas the remaining element is considered as the one we aim to predict (however, for the sake of simplicity, hereafter we will simply define the two sets as $\{\boldsymbol{z}_i\}_{i=1}^n$ and $\boldsymbol{z}_{n+1}$). All simulations are computed using the R Programming Language (R Core Team, 2020).

3.3.1 Simulation Study 1: Coverage

The aim of the simulation study in this section is to compare the empirical coverage (computed as the fraction of the $N = 5000$ replications in which $\boldsymbol{y}_{n+1}$ belongs to $\mathcal{C}_{n,1-\alpha}(x_{n+1})$) reached by the method presented in Section 3.2 (*CP Method*) with that obtained by an alternative method in different scenarios and for different values of n. The alternative method (*D Method*) is a depth-based approach, in which the prediction band is defined as the sample $(1-\alpha)\cdot 100\%$ central region as computed in Sun and Genton (2011). In detail, to allow a meaningful comparison between the two approaches, the band depth used in Sun and Genton (2011) has been replaced by the nonconformity measure in Equation 3.1. Technically, after computing the nonconformity score in

Equation 3.2 $\forall d \in \{1, \ldots, n\}$ by calculating $[\hat{\mu}^j(x_d)](t)$ and $s_j(t)$ without splitting the data into training and calibration sets (i.e., by using the whole information provided by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$), the prediction band is determined by the envelope of the $(1-\alpha)\cdot 100\%$ central region.

The simulation study consists of two scenarios. In the first one, the systematic component generating data is linear and, in addition to the case in which the model is correctly specified, two different kinds of model misspecification are taken into account: misspecification due to omitted relevant variable and misspecification due to inclusion of irrelevant variable (see Rao, 1971). In the second scenario, a third kind of model misspecification is evaluated, i.e., functional form misspecification (see Wooldridge, 1994). The two scenarios are formally defined as follows:

- Scenario 1

$$\begin{aligned} Y_{i,1}(t) =& \beta_0(t) + \beta_1(t) w_i + \varepsilon_{i,1}(t), \quad i \in \{1, \ldots, n+1\}, t \in [0,1] \\ Y_{i,2}(t) =& \beta_0(t) + \beta_2(t) w_i^2 + \varepsilon_{i,2}(t), \quad i \in \{1, \ldots, n+1\}, t \in [0,1] \end{aligned}$$

  with $w_i = i/(n+1)$, $\beta_0(t), \beta_1(t), \beta_2(t)$ generated by means of a B-spline basis expansion of order four, with six basis functions, equally spaced knots, coefficients generated independently by a standard normal random variable and $\varepsilon_{i,1}(t), \varepsilon_{i,2}(t)$ independent functional errors obtained by means of the same B-spline basis expansion with independent standard normal random variables as coefficients. It is important to notice that regression coefficient functions $\beta_0, \beta_1, \beta_2$ are generated only once, i.e., they do not vary between the $N = 5000$ replications.

- Scenario 2

$$\begin{aligned} Y_{i,1}(t) =& \exp(\beta_0(t) + \beta_1(t) w_i + \varepsilon_{i,1}(t)), \quad i \in \{1, \ldots, n+1\}, t \in [0,1] \\ Y_{i,2}(t) =& \exp(\beta_0(t) + \beta_2(t) w_i^2 + \varepsilon_{i,2}(t)), \quad i \in \{1, \ldots, n+1\}, t \in [0,1] \end{aligned}$$

  with $w_i, \beta_0(t), \beta_1(t), \beta_2(t), \varepsilon_{i,1}(t), \varepsilon_{i,2}(t)$ defined as in Scenario 1.

Both scenarios are evaluated considering the following three regression estimates:

- Set of Covariates 1; $[\hat{\mu}^1(x_i)](t) = [\hat{\mu}^2(x_i)](t) = \hat{\beta}_0(t)$,
- Set of Covariates 2; $[\hat{\mu}^1(x_i)](t) = \hat{\beta}_0(t) + \hat{\beta}_1(t) w_i$ and $[\hat{\mu}^2(x_i)](t) = \hat{\beta}_0(t) + \hat{\beta}_2(t) w_i^2$,

Scenario 1 — CP Method			
n	Set of Cov. 1	Set of Cov. 2	Set of Cov. 3
20	0.894[0.886,0.903]	0.896[0.888,0.905]	0.904[0.896,0.912]
200	0.902[0.894,0.910]	0.894[0.885,0.903]	0.901[0.893,0.909]
2000	0.899[0.890,0.907]	0.906[0.898,0.914]	0.902[0.894,0.911]
Scenario 1 — D Method			
n	Set of Cov. 1	Set of Cov. 2	Set of Cov. 3
20	0.136[0.127,0.146]	0.105[0.096,0.113]	0.067[0.060,0.074]
200	0.680[0.667,0.693]	0.671[0.658,0.684]	0.673[0.660,0.686]
2000	0.862[0.853,0.872]	0.858[0.848,0.867]	0.856[0.846,0.866]
Scenario 2 — CP Method			
n	Set of Cov. 1	Set of Cov. 2	Set of Cov. 3
20	0.907[0.899,0.915]	0.899[0.890,0.907]	0.904[0.896,0.913]
200	0.899[0.891,0.907]	0.898[0.890,0.907]	0.901[0.893,0.909]
2000	0.893[0.884,0.901]	0.893[0.884,0.901]	0.899[0.891,0.908]
Scenario 2 — D Method			
n	Set of Cov. 1	Set of Cov. 2	Set of Cov. 3
20	0.150[0.140,0.160]	0.129[0.120,0.138]	0.107[0.098,0.115]
200	0.714[0.701,0.727]	0.691[0.679,0.704]	0.679[0.666,0.692]
2000	0.859[0.849,0.868]	0.868[0.859,0.878]	0.872[0.863,0.881]

TABLE 3.1: Simulation study 1: empirical coverage and related 95% confidence interval $[\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/N}]$ in brackets reached by the method presented in Section 3.2 (*CP Method*) and those obtained by the alternative method (*D Method*) for each combination of scenario, sample size and set of covariates. $\alpha = 0.10$, set of modulation functions $\{s_j^\sigma\}_{j=1}^2$.

- Set of Covariates 3; $[\hat{\mu}^1(x_i)](t) = [\hat{\mu}^2(x_i)](t) = \hat{\beta}_0(t) + \hat{\beta}_1(t)w_i + \hat{\beta}_2(t)w_i^2$,

with $\hat{\beta}_0(t), \hat{\beta}_1(t), \hat{\beta}_2(t)$ the estimates obtained by fitting each time the corresponding function-on-scalar linear model. Focusing on Scenario 1, 'Set of Covariates 1' represents the omitted relevant variable case, 'Set of Covariates 2' represents the case in which the model is correctly specified and 'Set of Covariates 3' represents the case in which an irrelevant variable is included, whereas Scenario 2 is characterized by functional form misspecification.

Table 3.1 shows the empirical coverage $\hat{p}$, as well as the 95% confidence interval $[\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/N}]$, obtained for each combination of scenario, method, sample size and set of covariates considering the set of modulation functions $\{s_j^\sigma\}_{j=1}^2$. As regards our method (CP Method), the results are decidedly satisfactory, since the empirical coverages are really close to $1 - \alpha = 0.90$ and the observed confidence intervals always include the desired coverage regardless the specific combination of scenario, sample size and set of covariates considered. Specifically, the method is able to guarantee the desired coverage also when the sample size is small and the model misspecified. Vice versa, the

prediction bands induced by the depth-based approach (D Method) are characterized by empirical coverages really far from $1-\alpha$, particularly when $n = 20$. In light of this, the Conformal Prediction framework seems to be fundamental to construct prediction bands having the desired coverage.

3.3.2 Simulation Study 2: Efficiency

The aim of the simulation study of this section is to compare the three sets of modulation functions ($\{s_j^0\}_{j=1}^p$, $\{s_j^\sigma\}_{j=1}^p$, $\{\bar{s}_j\}_{j=1}^p$) in terms of efficiency (i.e., size of the prediction bands). Specifically, for each of the $N = 5000$ replications the size of the observed prediction band $\mathcal{C}_{n,1-\alpha}(x_{n+1})$ is defined as the average value $\mathcal{Q}(\cdot)/2$ (see Equation 3.3). Three different scenarios are taken into account: focusing just for now on the error terms and ignoring the systematic components, in the first scenario the error terms are characterized by a constant variability over the domains, in the second scenario the variability differs whereas in the third scenario the presence of outliers further complicates their specification. Formally, the three scenarios are:

- Scenario 1. The two systematic components are defined as in the first scenario of Section 3.3.1, while the independent functional errors $\varepsilon_{i,1}(t), \varepsilon_{i,2}(t)$ are defined as follows:

$$\begin{aligned}\varepsilon_{i,j}(t) = & B_{i+(n+1)(j-1),1} + \\ & B_{i+(n+1)(j-1),2} \cos\left(10\pi\left(t + U_{i+(n+1)(j-1)}\right)\right) + \\ & B_{i+(n+1)(j-1),3} \sin\left(10\pi\left(t + U_{i+(n+1)(j-1)}\right)\right)\end{aligned}$$

  $i \in \{1, \ldots, n+1\}, j \in \{1, 2\}, t \in [0, 1]$, with i.i.d. random vectors $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_{2(n+1)} \sim N_3(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ having the entries on the main diagonal equal to 1 and the entries outside the main diagonal equal to 0.7, i.i.d. random variables $U_1, \ldots, U_{2(n+1)} \sim U[-0.5, 0.5]$.

- Scenario 2. The two systematic components are defined as in the first scenario of Section 3.3.1, while the independent functional errors $\varepsilon_{i,1}(t), \varepsilon_{i,2}(t)$ are obtained by means of a B-spline basis expansion of order four, with 13 basis functions, equally spaced knots and normal random vectors as vectors of coefficients. Specifically, the $2 \cdot (n+1)$ (observed) vectors of coefficients are independent realizations of $\boldsymbol{C} = (C_1, \ldots, C_{13}) \sim N_{13}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ diagonal matrix such that $\mathrm{Var}[C_a] = 0.001$, $a \neq 7$, $\mathrm{Var}[C_7] = 9 \cdot 10^{-6}$.

- Scenario 3

$$Y_{i,1}(t) = \beta_0(t) + \eta_{i,1}(t), \quad i \in \{1, \dots, n+1\}, t \in [0,1],$$
$$Y_{i,2}(t) = \beta_0(t) + \eta_{i,2}(t), \quad i \in \{1, \dots, n+1\}, t \in [0,1]$$

with $\beta_0(t) = 0 \ \forall t \in [0,1]$,

$$\eta_{i,j}(t) = \beta_1(t) w_{i,j} + \varepsilon_{i,j}(t), \quad i \in \{1, \dots, n+1\}, j \in \{1,2\}, t \in [0,1]$$

with $\beta_1(t)$ obtained by means of a B-spline basis expansion of order four, with 13 basis, equally spaced knots and all coefficients equal to 0 but the seventh equal to 0.5, $\varepsilon_{i,j}(t)$ defined as in Scenario 2, and if $n = 20$ then $w_{i,j} = 0 \ \forall \{i,j\} \neq \{1,1\}$, $w_{1,1} = 1$, whereas if $n \in \{200, 2000\}$ then

$$w_{i,j} = \begin{cases} 1 & \text{if } i \in \left\{ j + 40 \cdot \zeta : \zeta \in \{0, 1, \dots, \frac{n}{40} - 1\} \right\} \\ 0 & \text{otherwise} \end{cases}.$$

Despite the complex notation, the introduction of $w_{i,j}$ is aimed at obtaining that $\sim 5\%$ of the multivariate functions $\boldsymbol{y}_1, \dots, \boldsymbol{y}_{n+1}$ (i.e., 1 out of 21 when $n = 20$, 10 out of 201 when $n = 200$, 100 out of 2001 when $n = 2000$) is characterized, in one of the two components, by the anomalous behavior induced by $\beta_1(t)$. We propose such an unconventional structure for the error terms to simulate, for example, a regression framework in which relevant variables are not available.

All three scenarios are evaluated considering only one set of covariates each, namely the case in which the corresponding model is correctly specified. Fig. 3.2 shows, for each scenario, a realization of the error terms $\{\varepsilon_{i,1}\}_{i=1}^{n+1}$ ($\{\eta_{i,1}\}_{i=1}^{n+1}$ for Scenario 3) when $n = 20$.

After verifying that all scenarios ensure the desired coverage (see Table 3.2, in which 25 out of the 27 confidence intervals at level 95% include the nominal value $1-\alpha$), Table 3.3 shows the median size (first and third quartile in brackets) of the $N = 5000$ prediction bands obtained for each combination of scenario, sample size and set of modulation functions. All three scenarios share the evidence that the prediction bands induced by $\{s_j^0\}_{j=1}^2$ are typically smaller than those induced by the other two sets of modulation functions when the sample size is very small ($n = 20$). This is due to the fact that regression estimates obtained with a small training set size likely provide an unreliable (and potentially misleading) set of modulation functions, leading to a preference for a

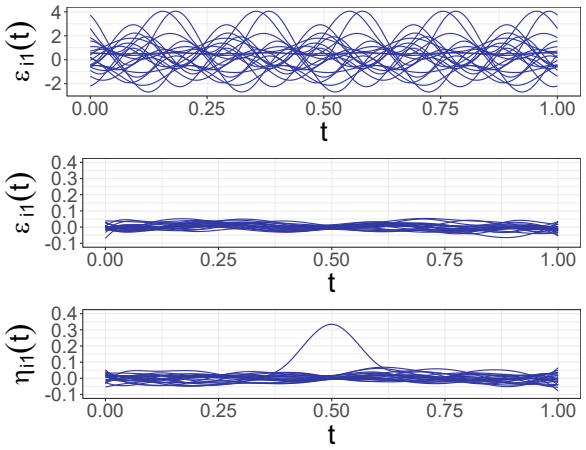

FIGURE 3.2: Example of realization of the error term related to $\{Y_{i,1}\}_{i=1}^{n+1}$. First scenario at the top, second scenario in the middle, third scenario at the bottom. $n = 20$.

	Scenario 1		
n	$\{s_j^0\}_{j=1}^2$	$\{s_j^\sigma\}_{j=1}^2$	$\{\bar{s}_j\}_{j=1}^2$
20	**0.889**[0.880,0.898]	0.897[0.889,0.905]	0.896[0.888,0.905]
200	0.900[0.891,0.908]	0.896[0.888,0.905]	0.900[0.891,0.908]
2000	0.900[0.891,0.908]	0.903[0.894,0.911]	0.894[0.886,0.903]
	Scenario 2		
n	$\{s_j^0\}_{j=1}^2$	$\{s_j^\sigma\}_{j=1}^2$	$\{\bar{s}_j\}_{j=1}^2$
20	0.893[0.885,0.902]	0.896[0.888,0.905]	0.896[0.888,0.905]
200	0.897[0.889,0.906]	0.900[0.892,0.908]	0.901[0.893,0.910]
2000	0.899[0.890,0.907]	0.895[0.887,0.904]	0.906[0.898,0.914]
	Scenario 3		
n	$\{s_j^0\}_{j=1}^2$	$\{s_j^\sigma\}_{j=1}^2$	$\{\bar{s}_j\}_{j=1}^2$
20	0.898[0.889,0.906]	0.902[0.894,0.910]	0.902[0.894,0.910]
200	0.901[0.893,0.909]	0.899[0.891,0.908]	**0.890**[0.881,0.898]
2000	0.898[0.890,0.907]	0.901[0.893,0.909]	0.895[0.887,0.904]

TABLE 3.2: Simulation study 2: empirical coverage and related 95% confidence interval in brackets for each combination of scenario, sample size and set of modulation functions. $\alpha = 0.10$. The values in bold highlight that the corresponding confidence intervals do not include $1 - \alpha = 0.9$.

set of modulation functions not depending on $\mathcal{I}_1$. As proof of that, it is not surprising that the two data-driven sets of modulation functions $\{s_j^\sigma\}_{j=1}^2, \{\bar{s}_j\}_{j=1}^2$ deliver the worst performance in the most complex Scenario, i.e., Scenario 3. Focusing on the other two sample sizes, in Scenario 1 the choice of not modulating seems appropriate due to the equal magnitude of the two components and the constant variability over $\mathcal{T}_1$, $\mathcal{T}_2$, but, as expected, the difference between the three alternative sets of modulation functions

	Scenario 1		
n	$\{s_j^0\}_{j=1}^2$	$\{s_j^\sigma\}_{j=1}^2$	$\{\bar{s}_j\}_{j=1}^2$
20	9.599[8.348,11.116]	12.205[10.121,14.853]	14.241[11.730,17.798]
200	8.658[8.289,9.077]	8.835[8.442,9.246]	9.315[8.892,9.784]
2000	8.568[8.449,8.692]	8.587[8.469,8.712]	8.681[8.561,8.806]
	Scenario 2		
n	$\{s_j^0\}_{j=1}^2$	$\{s_j^\sigma\}_{j=1}^2$	$\{\bar{s}_j\}_{j=1}^2$
20	0.168[0.152,0.188]	0.190[0.167,0.221]	0.213[0.186,0.249]
200	0.148[0.144,0.153]	0.126[0.123,0.130]	0.139[0.135,0.144]
2000	0.146[0.145,0.148]	0.122[0.121,0.123]	0.134[0.133,0.136]
	Scenario 3		
n	$\{s_j^0\}_{j=1}^2$	$\{s_j^\sigma\}_{j=1}^2$	$\{\bar{s}_j\}_{j=1}^2$
20	0.201[0.157,0.667]	0.294[0.212,1.767]	0.407[0.277,1.869]
200	0.162[0.155,0.170]	0.167[0.161,0.172]	0.151[0.145,0.158]
2000	0.160[0.157,0.162]	0.161[0.160,0.163]	0.145[0.143,0.147]

TABLE 3.3: Simulation Study 2: median size (first and third quartile in brackets) for each combination of scenario, sample size and set of modulation functions. $\alpha = 0.10$.

decreases when n grows. Differently from Scenario 1, Scenario 2 is characterized by multivariate residuals showing a lower variability in the central portion of $\mathcal{T}_1$ and $\mathcal{T}_2$: as a consequence, $\{s_j^0\}_{j=1}^2$ provides large prediction bands since it is not able to adapt the width of the band according to the local variability of the residuals, whereas $\{s_j^\sigma\}_{j=1}^2$ is particularly effective since it induces a modulation process based on the two standard deviation functions. Finally, $\{\bar{s}_j\}_{j=1}^2$ represents the best solution in Scenario 3 given its ability to focus on the 'least extreme' $\sim (1-\alpha)\cdot 100\%$ of data: indeed, differently from $\{s_j^0\}_{j=1}^2$ it is able to reduce the width of the band in the central part of the domains, and differently from $\{s_j^\sigma\}_{j=1}^2$ it does not uselessly enlarge the band in the same subinterval of $\mathcal{T}_1, \mathcal{T}_2$. Consequently, the simulation study seems to confirm the statistical intuition given in Section 3.2.2 that the newly launched set of modulation functions $\{\bar{s}_j\}_{j=1}^p$ represents an interesting solution when functional residuals show a non-standard behavior and a modulation process driven by the value $1-\alpha$ is needed.

3.4 Case Study: Analysis of Bike Mobility in the City of Milan

In order to illustrate the application potential of the method presented in this chapter, in this section we focus on a case study concerning urban mobility, and specifically the usage of a bike-sharing system in the Italian city of Milan. Moving from the raw

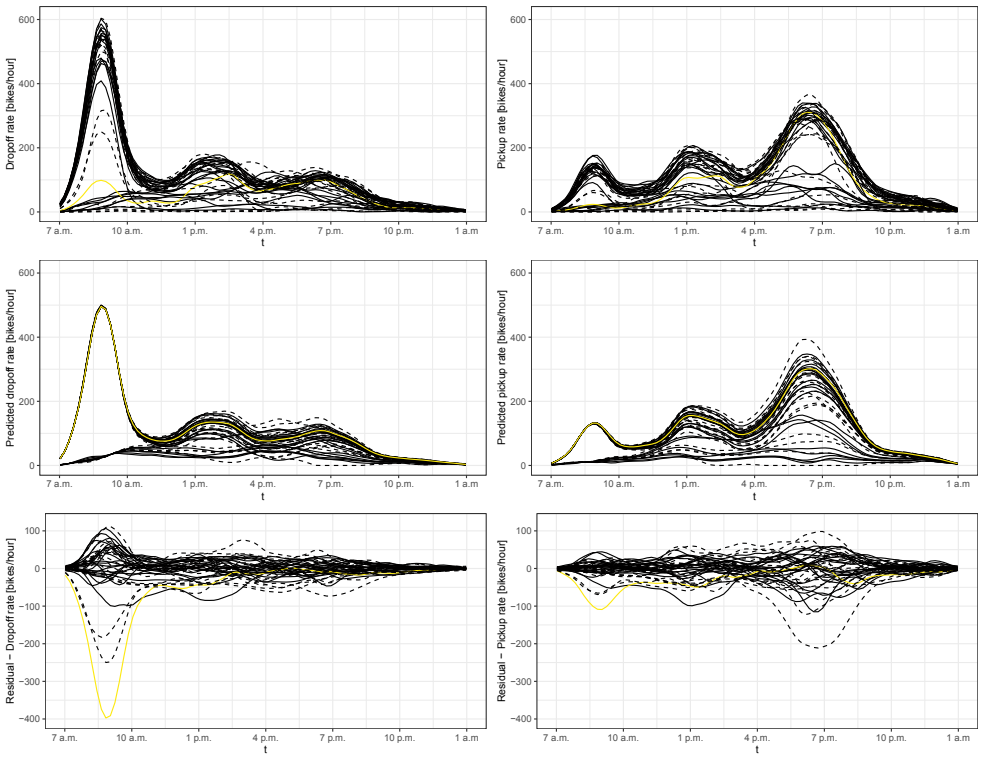

FIGURE 3.3: Dropoff and pickup rates (top left, top right respectively), corresponding functional predictions (center left, center right) and functional residuals (bottom left, bottom right). Yellow curves refer to 29 February; continuous curves refer to the observations in the training set, dashed curves to those in the calibration set.

data and the context presented in Torti *et al.* (2021), the aim is to study the behavior of subscribers of Bikemi, a bike sharing system active in the city in which bikes are picked up and dropped off in specific docking stations located through the city. Starting from raw data providing various information about picked up bikes (simply pickups hereafter) and dropped off bikes (simply dropoffs hereafter) for each day considered, and focusing our attention — as an example — on the Duomo district only (i.e., the area in which Milan's cathedral is), the multivariate functional response variable $\boldsymbol{y}_i = (y_{i,1}, y_{i,2})$ representing the rate of dropoffs ($y_{i,1}$) and pickups ($y_{i,2}$) is obtained via a standard kernel density estimation smoothing method (see, e.g., Hastie *et al.*, 2009). In doing so, $y_{i,1}(t)$ ($y_{i,2}(t)$) represents the dropoff (pickup) rate at time t, with t ranging from 7 a.m. day i to 1 a.m. the next day (consequently, we assume that day i ends at 1 a.m. the next day). The period considered starts on 25 January 2016 and ends on 6 March 2016: due to an error in the data collection, 25 February is removed from the dataset in accordance with Torti *et al.* (2021), and so the sample size is $n = 41$. Data are shown in the two top panels of Fig. 3.3.

Like in Torti *et al.* (2021), the regression estimates are obtained by fitting a concurrent function-on-function linear model (Ramsay and Silverman, 2005). The model hereby used includes as covariates a functional intercept, the temperature function (after subtracting the average daily temperature function of the period considered) in degrees Celsius, and a dummy variable indicating whether day i is a weekday or not. Since the rates cannot be negative in any subinterval of the domain, the predicted functions are truncated to 0. However, as discussed in Section 3.2.1, the purpose is to construct valid, meaningful and interpretable prediction bands also when simple regression estimators are specified. In view of this, the choice of the covariates, as well as the functional form of the model, represents an aspect of limited interest in the framework considered, but carefully chosen alternative models (e.g., the nonparametric additive one of Ferraty and Vieu, 2009) could surely provide more accurate and reliable estimates.

The method presented in Section 3.2 is performed by considering the three sets of modulation functions $\{s_j^0\}_{j=1}^2$, $\{s_j^\sigma\}_{j=1}^2$, $\{\bar{s}_j\}_{j=1}^2$, $\alpha = 0.25$ and $m = 22$, $l = 19$ in order to assign, as in the simulation studies, about half of the observations to the training set and to obtain the value $1 - \lfloor (l+1)\alpha \rfloor / (l+1)$ equal to $1 - \alpha$. To remain as neutral as possible, we will consider the case in which — after having labeled the days considered with numbers from 1 to 41 — the observations referring to an odd day are assigned to the training set and those referring to an even day to the calibration set, with the observation related to day 20 assigned to the training set to satisfy $m = 22$. Two possible prediction scenarios are taken into account for the scope of visualization: in the first, we construct the multivariate prediction band for a weekday having the average temperature function of the period as temperature function; in the second, we construct it for a warmer than usual weekday. Fig. 3.4 shows, for each of the three sets of modulation functions ($\{s_j^0\}_{j=1}^2$ in the first row, $\{s_j^\sigma\}_{j=1}^2$ in the second row, $\{\bar{s}_j\}_{j=1}^2$ in the third row), the prediction sets induced by the two scenarios (first set of covariates in the first column, second in the second column). In particular, each panel shows the prediction band for the dropoff rate (light blue band) and the pickup rate (red band), with the two dashed lines representing the corresponding regression estimates. As for the predicted functions, the prediction bands are truncated to 0, as the rates cannot be negative in any subinterval of the domain. Note that this truncation does not involve any kind of drawback since the coverage reached by the prediction sets remains unchanged if a null probability portion of the bands is removed from the prediction bands. It is evident that the prediction bands for dropoffs induced by $\{s_j^\sigma\}_{j=1}^2$ are quite large in the initial portion of the domain compared to those obtained by not modulating (i.e., $\{s_j^0\}_{j=1}^2$) and by the proposed set $\{\bar{s}_j\}_{j=1}^2$. In order to clarify this aspect, let us consider Fig. 3.3.

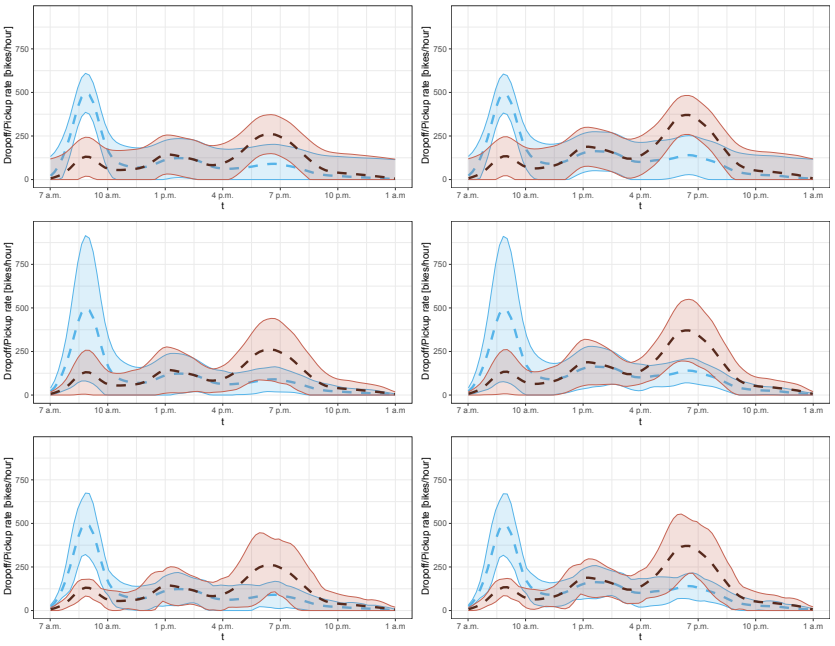

FIGURE 3.4: Prediction bands for the dropoff rate (light blue band) and the pickup rate (red band). Each panel refers to a combination of set of modulation functions ($\{s_j^0\}_{j=1}^2$ at the top, $\{s_j^\sigma\}_{j=1}^2$ in the middle, $\{\bar{s}_j\}_{j=1}^2$ at the bottom) and scenario (first set of covariates on the left, second on the right). The dashed lines indicate the corresponding regression estimates. $\alpha = 0.25$. Split into calibration/training set: even/odd(+ day 20) days.

Focusing on the residual functions of the dropoff rates (i.e., the panel at the bottom left of the figure), it is easily noticeable that the yellow curve (referring to weekday 35, i.e., 29 February, which is assigned to the training set) shows an anomalous behavior in the initial part of the domain. The panel at the top left of the same figure suggests that this is due to the fact that day 35 was characterized by an unusually low dropoff rate compared to that observed in the other weekdays (which are the curves showing a pick around 9 a.m.). Consequently, by using $\{s_j^\sigma\}_{j=1}^2$ it is natural to obtain prediction bands for dropoffs extremely wide in the first portion of the domain since this outlier has a huge impact on the modulation process, while the corresponding prediction bands obtained by not modulating are not adversely affected as $\{s_j^0\}_{j=1}^2$ does not modulate the width of the band according to the local variability of the residuals. In view of this, the set of modulation functions $\{\bar{s}_j\}_{j=1}^2$ represents an intriguing solution since, in addition to modulate the width of the band along the domains, induces a modulation process which is not misled by the anomalous behavior of day 35. However, similar considerations would have been made also if other observations than the one related to day 35 had been assigned to the training set, as can be noticed by analyzing the functional residuals of the observations assigned to the calibration set in the two panels at the bottom of

Fig. 3.3 (dashed curves for the calibration set; continuous curves for the training set). Despite the small sample size, the prediction sets of Fig. 3.4 can provide profitable information: first of all, subscribers of Bikemi seem to mainly use bikes to go to Duomo in the morning, whereas in the early evening the bike flow is reversed. Moving from the first set of covariates (weekday-temperature equal to the mean temperature of the period) to the second one (weekday-warm day), we notice that a higher temperature does not strongly affect people's behavior in the morning, whereas it involves a moderate increase in dropoffs and, at the same time, a big increase in pickups in the period of time around 7 p.m.. The information provided by the prediction bands can be indeed very useful to fleet managers in identifying the periods of time in which the imbalance between pickups and dropoffs could become critical based on the day of the week, the temperature function and other possible carefully chosen covariates.

3.5 Conclusion and Further Developments

In the present chapter we have developed a procedure aimed at creating prediction bands for multivariate functional data in a regression framework. Moving from the approach proposed in Chapter 2 for univariate i.i.d. functional data, the method presented in this chapter builds finite-sample either exact or valid prediction bands under the only assumption of exchangeable regression pairs with multivariate functional response. Despite the paramount importance of this topic both from the methodological and applied point of view, to the best of our knowledge our method represents the first proposal able to ensure such important features. These properties, together with the fact that the procedure is scalable and the bands can be easily found in closed form, allow to obtain meaningful prediction bands regardless the regression estimator used, leading to a methodology which can be applied in a wide range of application scenarios. Moreover, we have introduced a specific set of modulation functions (namely $\{\bar{s}_j\}_{j=1}^{p}$) achieving an asymptotic result in terms of efficiency regardless the sample observed $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ and inducing prediction bands whose width varies along the domains and across the components according to the local behavior. The simulation study and the real-world application provided in Section 3.3 and 3.4 respectively confirm the potential of the approach. Nevertheless, many possible directions still remain unexplored. First of all, it would be extremely interesting to extend some approaches recently developed in Conformal literature to the functional context, such as the conformalized quantile regression method of Romano *et al.* (2019) or the Distributional Conformal prediction approach of Chernozhukov *et al.* (2021). Secondly, we plan to explore the impact of the

regression estimator on the size of the prediction sets.

Chapter 4

Prediction bands for multivariate functional time series

4.1 Introduction

Since 2000 we have witnessed a revolution in terms of the availability of computational power and storage space, which are now ubiquitous and cheap. This new context has triggered a revolution in paradigm also in statistics and in the professional practice of modern statisticians and forecasters, who now face way less methodological and practical limitations with respect to their older collegues.

For instance, when dealing with continuous phenomena over a spatial or a temporal domain (e.g. a trajectory, a surface or a demand/offer curve) instead of using standard scalar statistics and working on a statistical summary of these objects, a practitioner may decide to turn to Functional Data Analysis (FDA)(Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012; Wang *et al.*, 2016).

A field of research with very promising applications is Functional Time Series (FTS), namely the study of methods and the development of applications to deal with functional data characterised by some kind of temporal dependency. The interested reader may refer to Hörmann and Kokoszka (2012) for a review of the theoretical underpinnings and some definitions on the field. The main focus of FTS, as testified by the very rich outstanding literature, has been the issue of one-step ahead forecasting. Among the many contributions that may be found in the literature, Chen *et al.* (2021) provides a review of some of the work on point prediction for FTS using autoregressive models, Ferraty *et al.* (2002); Ferraty and Vieu (2004) present nonparametric methods and Canale and Vantini (2016) presents an extension for the forecasting of FTS with constraints. On a slightly different line of reasoning Hyndman and Shang (2009) proposes a method based

on dimensionality reduction via weighted functional principal component and weighted functional partial least squares regression while Gao and Shang (2017) proposes a vector error-correcting model, still based on dimensionality reduction of functional data.

The great majority of the presented work in FTS forecasting focuses on point predictions: the issue of providing quantification of uncertainty is usually addressed using extensions of the Bootstrap to non $i.i.d$ cases (see e.g. Paparoditis and Shang (2021), which also provides a good introduction to the field, as well as Rossini and Canale (2019) where a remarkable extension to the constrained FTS case is presented). Bootstrap-based methods, though, are shown to have relatively weak finite-sample properties, and are of course very computationally intensive.

The aim of the present chapter is to propose distribution-free prediction bands for multivariate functional time series guaranteeing finite-sample performance bounds in terms of coverage and asymptotic exactness, i.e. coverage asymptotically equal to the nominal confidence level. To do that, we move from a relatively new contribution by Chernozhukov *et al.* (2018) showing an extension of Conformal methods to time dependent data.

From the application point of view, we focus on the issue of forecasting in energy markets, and specifically in the Italian gas market. The specific regulatory framework of Italian energy markets has given birth to big and novel challenges to the players in the market: producers, brokers and utilities need to be able to forecast with great accuracy exchanged quantities as well as prices on the markets. This is of key importance for tactical and strategic planning in producing, storing and trading energy. Uncertainties and their quantification are of key importance in this context: having reliable assessments regarding the uncertainty of the predictions obtained provides invaluable insights to risk management and represents the basis for any reliable hedging strategy in trading.

Forecasts, though, should not only be constrained to spot quantities and prices: the market is a dynamic object, and it is crucial to be able to understand not only its behaviour at equilibrium but also its "shape" in terms of position and slope of its demand and offer curves, allowing traders to evaluate the possible effect of their offers/bids on the market. To do so, the FDA approach has proven to be a very powerful method: the already cited Canale and Vantini (2016) alongside Shah and Lisi (2020) represent some recent contributions in the field of FDA with respect to jointly predicting demand and offer curves for the gas/electricity market. One may refer to the references contained in the two papers for a thorough representation of the history of the field in terms of price and demand forecasting in the electricity and gas markets. Expanding on Canale and Vantini (2016), Rossini and Canale (2019) adds uncertainty quantification to the

framework presented for the gas market. As previously said though, these bands have no finite-sample coverage guarantee. The aim of our application test case is thus to show how our methodological proposal may provide more reliable information in terms of prediction uncertainty to energy traders.

The chapter is structured as follows: Section 4.2 presents the proposed method and its theoretical underpinnings in detail, while in Section 4.3 and 4.3.1 we present a simulation study to assess the properties of the method and its results. The application is presented in Section 4.4 while Section 4.5 presents conclusions and draws further developments.

4.2 Methods

Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ be a time series such that $\boldsymbol{Z}_t = (X_t, \boldsymbol{Y}_t)$ consists of a set of covariates X_t and a multivariate functional response variable $\boldsymbol{Y}_t$, $t \in \{1, \ldots, T\}$. Let $\boldsymbol{Y}_t = (Y_{t,1}, \ldots, Y_{t,p})$ be a multivariate random function where its j-th component $Y_{t,j}$ $(j = 1, \ldots, p)$ is a random function which takes values in $L^\infty(\mathcal{Q}_j)$, that is the family of limited functions $y : \mathcal{Q}_j \to \mathbb{R}$ with $\mathcal{Q}_j$ closed and bounded subset of $\mathbb{R}^{d_j}$, $d_j \in \mathbb{N}_{>0}$. For simplicity, later in the discussion we will use $\prod_{j=1}^{p} L^\infty(\mathcal{Q}_j)$ to indicate the space $L^\infty(\mathcal{Q}_1) \times \cdots \times L^\infty(\mathcal{Q}_p)$ in which $\boldsymbol{Y}_t$ takes values. The framework considered is very general since it includes the case of univariate functional response variable (when $p = 1$) as a special case, but it also allows the $p > 1$ domains $\mathcal{Q}_j$ and images of $Y_{t,j}$ to be greatly different when j changes. X_t belongs to a measurable space and contains the lagged functional response variable, in addition to possible external predictors (which can be, for example, scalar or functional covariates). As regards the notation, it is important to notice that two components of $\boldsymbol{Y}_t$ ($Y_{t,1}$ and $Y_{t,2}$ for example) may share some (or even all) the covariates, but generally speaking X_t represents the set of all the covariates related to $Y_{t,1}, \ldots, Y_{t,p}$. Let $\mu^j(x_t) = \mathbb{E}(Y_{t,j}|X_t = x_t)$ be the regression map for the j-th component evaluated at x_t, and similarly let us define the scalar quantity $[\mu^j(x_t)](q) = \mathbb{E}(Y_{t,j}(q)|X_t = x_t)$.

Our aim is to introduce a procedure able to create simultaneous prediction sets for $\boldsymbol{Y}_{T+1}$ (i.e. prediction sets holding for the multivariate random function $\boldsymbol{Y}_{T+1}$ globally, and not only for its j-th component $Y_{T+1,j}$) based on the information provided by $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ and by X_{T+1} and ensuring performance bounds in terms of unconditional coverage. By recalling the definition of prediction set introduced in Chapter 1, the purpose is to obtain prediction sets $\mathcal{C}_{T,1-\alpha}(X_{T+1})$ whose unconditional coverage $\mathbb{P}(\boldsymbol{Y}_{T+1} \in \mathcal{C}_{T,1-\alpha}(X_{T+1}))$ is close to the nominal confidence level $1-\alpha$ under mild conditions on the data generating process and being characterized - as in Chapter 3 - by a

particular shape, i.e. they are multivariate functional prediction bands.

Moving from the literature on inference via permutations (see, e.g., Rubin, 1984; Romano, 1990; Lehmann and Romano, 2006) and on Conformal Prediction, we present a modification of the Conformal inference able to account for time series dependence. Intuitively, the idea is to generalize the Conformal approach traditionally used when the regression pairs $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{T+1}$ are i.i.d. by randomizing blocks of observations. Specifically, we extend the non-overlapping blocking scheme proposed by Chernozhukov *et al.* (2018) to the Split Conformal framework. This extension is mentioned as possible in Chernozhukov *et al.* (2018), nevertheless to the best of our knowledge it has never been formally built - or even taken into account - in the literature. In light of this, first of all the procedure we built is presented, then its logic is explained.

Let m, l be two strictly positive integers such that $T = m+l$, and let us define $\mathcal{I}_1$ and $\mathcal{I}_2$ as two sets of size m and l respectively such that $\mathcal{I}_1 \cup \mathcal{I}_2 = \{1, \ldots, T\}$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$. Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$ be realizations of $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$, with $\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}$ denoting the *training set* of size m and $\{\boldsymbol{z}_h : h \in \mathcal{I}_2\}$ denoting the *calibration set* of size l, and let $b \in \{1, \ldots, l+1\}$ be a value such that $(l+1)/b$ is an integer[1]. In accordance with the Conformal Prediction framework, let us also define any measurable function $A(\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}, \boldsymbol{z})$ which takes values in $\bar{\mathbb{R}}$ as *nonconformity measure*. As suggested by the name, the purpose of the nonconformity measure is to score how different the generic element $\boldsymbol{z}$ is from the elements of the training set: for example, in the traditional non-functional regression framework in which the response variable is scalar and the set of covariates is a vector, a popular choice of nonconformity measure is the absolute value of the regression residual obtained by fitting the regression algorithm on the training set. For any given value of b, it is therefore possible to define the collection of $(l+1)/b$ index permutations $\Pi = \{\pi_i : 1 \leq i \leq (l+1)/b\}$, whose element $\pi_i : \{1, \ldots, l+1\} \longrightarrow \{1, \ldots, l+1\}$ is the bijective function defined as:

$$\pi_i(t) = \begin{cases} t + (i-1)b & \text{if } 1 \leq t \leq l - (i-1)b + 1 \\ t + (i-1)b - l - 1 & \text{if } l - (i-1)b + 2 \leq t \leq l + 1. \end{cases}$$

The permutation scheme Π is an algebraic group containing the identity element (as $\pi_1(t) = t \ \forall t \in \{1, \ldots, l+1\}$) which naturally induces the set of scalar values $\mathcal{D}_\Pi = \{\pi_i(l+1) : 1 \leq i \leq (l+1)/b\} \subseteq \{1, \ldots, l+1\}$, which is the set of integers used to identify the observations for which the nonconformity scores will be calculated.

[1] $(l+1)/b$ is assumed to be an integer for simplicity, but the procedure can be easily generalized to include values of b such that $(l+1)/b$ is not integer-valued.

The prediction set for $\boldsymbol{Y}_{T+1}$ (which is a multivariate prediction band or not depending on the choice of the nonconformity measure) is therefore defined as

$$\mathcal{C}_{T,1-\alpha}(x_{T+1}) := \left\{ \boldsymbol{y} \in \prod_{j=1}^{p} L^{\infty}(\mathcal{Q}_j) : \delta_{\boldsymbol{y}} > \alpha \right\},$$

with

$$\delta_{\boldsymbol{y}} := \frac{|\{d \in \mathcal{D}_{\Pi} : R_{\omega_d} \geq R_{T+1}\}|}{|\mathcal{D}_{\Pi}|},$$

$|\mathcal{D}_{\Pi}| = (l+1)/b$, ω_d the dth smallest value in the set $\mathcal{I}_2 \cup \{T+1\}$ and *nonconformity scores* $R_{\omega_d} := A(\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}, \boldsymbol{z}_{\omega_d})$, $R_{T+1} := A(\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}, \boldsymbol{z}_{T+1})$, where $\boldsymbol{z}_{T+1} = (x_{T+1}, \boldsymbol{y})$. Since $T+1$ is always included in $\{\omega_d : d \in \mathcal{D}_{\Pi}\}$ (being $l+1 = \pi_1(l+1)$ and $\omega_{l+1} = T+1$), δ_y can be conveniently rewritten as

$$\delta_{\boldsymbol{y}} = \frac{1 + |\{d \in \{\pi_i(l+1) : 2 \leq i \leq (l+1)/b\} : R_{\omega_d} \geq R_{T+1}\}|}{|\mathcal{D}_{\Pi}|}.$$

Intuitively, the idea is the one introduced by Split Conformal Prediction: after randomly dividing the observed data into the training and calibration sets, the prediction set $\mathcal{C}_{T,1-\alpha}(x_{T+1})$ is defined as the set of all $\boldsymbol{y} \in \prod_{j=1}^{p} L^{\infty}(\mathcal{Q}_j)$ such that $(x_{T+1}, \boldsymbol{y})$ is similar - in terms of nonconformity measure A - to the training set $\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}$ compared to the conformity of the elements of the calibration set to the same training set. Differently from the Split Conformal Prediction framework, the permutation scheme here proposed randomizes the elements of the calibration set by considering blocks of observations of length b, and it computes the nonconformity scores for $(l+1)/b - 1$ elements of the calibration set (one for each block) and for $\boldsymbol{z}_{T+1}$. In so doing, when b increases the nonconformity scores are computed for observations more distant in time from each other (and so one is justified in expecting the dependence between the nonconformity scores to decrease under some conditions on the data generating process, a fundamental aspect as we will see shortly), but the number of nonconformity scores computed decreases. On the other hand, when $b = 1$ the approach here proposed is equivalent to the Split Conformal approach, and the nonconformity scores are computed for each observation in the calibration set.

Regardless the value of b, the permutation scheme Π guarantees that, if the regression pairs $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{T+1}$ are i.i.d. (or even exchangeable), the prediction sets obtained are finite-sample valid, i.e. $\mathbb{P}(\boldsymbol{Y}_{T+1} \in \mathcal{C}_{T,1-\alpha}(X_{T+1})) \geq 1 - \alpha \ \forall \ T, \alpha \in (0,1)$, due to the fact that the nonconformity scores are exchangeable. The proof can be trivially obtained by generalizing the well-established result holding in the Split Conformal

framework (Vovk *et al.*, 2005). As in the Conformal setting, the result concerning the validity of the prediction sets induced by the permutation scheme Π can be enriched by proving that, if the nonconformity scores have a continuous joint distribution, then $\mathbb{P}\left(\boldsymbol{Y}_{T+1} \in \mathcal{C}_{T,1-\alpha}\left(X_{T+1}\right)\right) = 1 - \frac{\lfloor \alpha(l+1)/b \rfloor}{(l+1)/b}$, i.e. the unconditional coverage is equal to an easy-to-compute value and it is not only greater than or equal to $1-\alpha$. Also in this case, the proof can be trivially obtained by generalizing Theorem 1.1.

If the regression pairs are not exchangeable, as in our case, the aforementioned results do not hold. Nevertheless, two desirable properties can still be obtained under some conditions: finite-sample performance bounds in terms of unconditional coverage and asymptotic exactness (i.e. unconditional coverage asymptotically equal to the nominal confidence level $1-\alpha$). These results, which represent an extension of a result due to Chernozhukov *et al.* (2018) to the Split framework, are reported in Theorem 4.1.

In order to introduce Theorem 4.1, let A^* be an oracle nonconformity measure inducing oracle nonconformity score $R^*_{\omega_d}$, which is typically the population counterpart of R_{ω_d}: for example, in the aforementioned non-functional regression setting, the oracle nonconformity score might be the magnitude of the error term. For notational simplicity, let us define $\bar{l} = |\mathcal{D}_\Pi| = (l+1)/b$ and let $\{\delta_{1\bar{l}}, \delta_{2m}, \gamma_{1\bar{l}}, \gamma_{2m}\}$ be sequences of positive scalar values converging to 0 when $\bar{l}, m \to 0$. Finally, let $\tilde{F}(a) := \frac{1}{\bar{l}} \sum_{d \in \mathcal{D}_\Pi} \mathbb{1}\left\{R^*_{\omega_d} < a\right\}$ and $F(a) = P(R^*_{T+1} < a)$.

Theorem 4.1. *If*

- $\sup_{a \in \mathbb{R}} |\tilde{F}(a) - F(a)| \leq \delta_{1\bar{l}}$ *with probability* $1 - \gamma_{1\bar{l}}$,
- $\frac{1}{\bar{l}} \sum_{d \in \mathcal{D}_\Pi} \left[R_{\omega_d} - R^*_{\omega_d}\right]^2 \leq \delta^2_{2m}$ *with probability* $1 - \gamma_{2m}$,
- $|R_{T+1} - R^*_{T+1}| \leq \delta_{2m}$ *with probability* $1 - \gamma_{2m}$,
- *The probability density function of* R^*_{T+1} *is bounded above by a constant* D,

then

$$\left|\mathbb{P}\left(\boldsymbol{Y}_{T+1} \in \mathcal{C}_{T,1-\alpha}\left(X_{T+1}\right)\right) - (1-\alpha)\right| \leq 6\delta_{1\bar{l}} + 2\delta_{2m} + 2D\left(\delta_{2m} + 2\sqrt{\delta_{2m}}\right) + \gamma_{1\bar{l}} + \gamma_{2m} \tag{4.1}$$

$\forall\ \alpha \in (0,1)$.

The first condition concerns the approximate ergodicity of $\tilde{F}(a)$ for $F(a)$, a condition which holds for strongly mixing time series using the permutation scheme Π (Chernozhukov *et al.*, 2018). The remaining conditions mainly concern the relationship between the nonconformity scores and the oracle nonconformity scores: intuitively,

δ_{2m} bounds the discrepancy between the nonconformity scores and their oracle counterparts. The proof of the Theorem mimics the proof of Theorem 2 in Chernozhukov *et al.* (2018) if $\{\delta_{1n}, \delta_{2n}, \gamma_{1n}, \gamma_{2n}\}$ are respectively replaced by $\{\delta_{1\bar{l}}, \delta_{2m}, \gamma_{1\bar{l}}, \gamma_{2m}\}$. We thus cross-refer to Chernozhukov *et al.* (2018) for details. Considering the Split framework of the manuscript, we require the four sequences $\{\delta_{1\bar{l}}, \delta_{2m}, \gamma_{1\bar{l}}, \gamma_{2m}\}$ to depend on m and $\bar{l}$ respectively according to their specific role: indeed, R_{ω_d} depends on the information provided by the training set, and so one is justified in requiring it to better approximate $R^*_{\omega_d}$ when the training set size m increases. As a consequence, $\{\delta_{2m}, \gamma_{2m}\}$ should depend on m. Conversely, the training set size does not affect $\tilde{F}(a)$ and $F(a)$ respectively since the training set does not affect the oracle nonconformity scores, and so the requirement is that the oracle nonconformity scores computed on the observations of the calibration set provide a proper approximation to R^*_{T+1} when the calibration set size increases. In so doing, $\{\delta_{1\bar{l}}, \gamma_{1\bar{l}}\}$ should depend on $\bar{l}$.

Theorem 4.1 provides, under some conditions, finite-sample performance bounds in terms of unconditional coverage regardless the value of b and it guarantees that the prediction sets are asymptotically exact since the right side of Inequality (4.1) converges to 0 when $m, \bar{l} \to 0$. In order to use the presented procedure in practical applications, we will consider a specific member of the family of nonconformity measures introduced in Chapter 3, i.e. the one inducing the following nonconformity scores:

$$R_{\omega_d} = \sup_{j\in\{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{q\in\mathcal{Q}_j} \left| \frac{y_{\omega_d,j}(q) - [\hat{\mu}^j(x_{\omega_d})](q)}{s_j(q)} \right| \right), \quad d \in \{\mathcal{D}_\Pi \setminus \{l+1\}\},$$

$$R_{T+1} = \sup_{j\in\{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{q\in\mathcal{Q}_j} \left| \frac{y_j(q) - [\hat{\mu}^j(x_{T+1})](q)}{s_j(q)} \right| \right),$$

with y_j the j-th component of $\boldsymbol{y}$, $[\hat{\mu}^j(x_{\omega_d})](q)$ estimate of $[\mu^j(x_{\omega_d})](q)$ based on the training set $\{\boldsymbol{z}_h : h \in \mathcal{I}_1\}$, s_j the standard deviation function of the functional regression residuals of the observations belonging to the training set, i.e.:

$$s_j(q) := \left(\sum_{h\in\mathcal{I}_1} \left(y_{h,j}(q) - [\hat{\mu}^j(x_h)](q) \right)^2 \right)^{1/2}. \tag{4.2}$$

By considering this nonconformity measure, if $\alpha \in [b/(l+1), 1)$ (which is the scenario we will consider hereafter because if $\alpha \in (0, b/(l+1))$ then $\mathcal{C}_{T,1-\alpha}(x_{T+1}) = \prod_{j=1}^p L^\infty(\mathcal{Q}_j)$

since $\delta_{\boldsymbol{y}}$ is always greater than or equal to $b/(l+1)$), then

$$\mathcal{C}_{T,1-\alpha}\left(x_{T+1}\right) := \Bigg\{ \boldsymbol{y} \in \prod_{j=1}^{p} L^{\infty}(\mathcal{Q}_j) : y_j(q) \in \big[[\hat{\mu}^j(x_{T+1})](q) - k^s \cdot s_j(q), \\ [\hat{\mu}^j(x_{T+1})](q) + k^s \cdot s_j(q)\big] \\ \forall j \in \{1,\dots,p\}, \forall q \in \mathcal{Q}_j \Bigg\},$$

with k^s the $\lceil (l+1)(1-\alpha)/b \rceil$th smallest value in the set $\{R_{\omega_d} : d \in \{\mathcal{D}_{\Pi} \setminus \{l+1\}\}\}$.

Differently from the case in which the regression pairs are i.i.d, in the framework of the manuscript the choice of the point predictors is key since it affects the relationship between the nonconformity score R_{ω_d} and its oracle counterpart $R^*_{\omega_d}$, and so the validity of Theorem 4.1: for example, strong model misspecification represents the typical case in which the validity of Theorem 4.1 is compromised, whereas the aforementioned results about the finite-sample unconditional coverage still holds in the i.i.d. setting also when the model is heavily misspecified. In addition, two further aspects depend on $[\hat{\mu}^j(x_h)](q)$. First of all, one is justified in expecting prediction bands to be smaller when accurate regression estimators are used as they usually output smaller nonconformity scores (and so a smaller k). Secondly, the regression estimators have a fundamental impact on the computational cost: indeed, the procedure here developed is highly scalable since, conditional on the computational cost required to obtain the regression estimates and s_j, the time needed to compute the prediction set increases linearly with T by assuming the ratio T/l and b fixed, and consequently the computational effort is mainly determined by the regression estimators used.

The strategy proposed in this Section represents a theoretically sound framework to obtain prediction bands when dealing with multivariate functional time series. In order to provide a comprehensive presentation of the method, in Section 4.3 we develop a simulation study whose aim is to evaluate the procedure in different scenarios, whereas in Section 4.4 the strategy is applied to real data in order to show its utility in real-world applications.

4.3 Simulation Study

In this Section we evaluate the procedure presented in Section 4.2 through a simulation study. Specifically, our aim is to analyze two different aspects: first of all (and most importantly) we estimate the coverage by computing the empirical coverage in various settings in order to compare it to the nominal confidence level $1-\alpha$; the estimation

procedure is detailed in Section 4.3.1. Secondly, we evaluate the size of the prediction bands obtained since, intuitively, a small prediction band is preferable because it includes subregions of the sample space where the probability mass is highly concentrated (Lei *et al.*, 2013) and it is typically more informative in practical applications.

We focus on a specific data generating process, that is evaluated by considering different values of T, b and kinds of model misspecification. The data generating process (obtained by setting $p = 1$, i.e. $\boldsymbol{Y}_t = Y_{t,1}$) is formally defined as follows:

$$Y_{t,1}(q) = \boldsymbol{g}'(q) \cdot \bar{\boldsymbol{Y}}_t = \bar{Y}_{t,1} + \bar{Y}_{t,2}\frac{sin(2\pi q)}{\sqrt{1/2}} + \bar{Y}_{t,3}\frac{cos(2\pi q)}{\sqrt{1/2}}$$

with $\bar{\boldsymbol{Y}}'_t = [\bar{Y}_{t,1}, \bar{Y}_{t,2}, \bar{Y}_{t,3}]$ a VAR(2) process, i.e.:

$$\bar{\boldsymbol{Y}}_t = \bar{\Psi}_1\bar{\boldsymbol{Y}}_{t-1} + \bar{\Psi}_2\bar{\boldsymbol{Y}}_{t-2} + \bar{\boldsymbol{\epsilon}}_t,$$

with $\bar{\Psi}_i = \frac{\Upsilon_i}{2\cdot||\Upsilon_i||}$ for $i = 1, 2$ and

$$\Upsilon_1 = \begin{bmatrix} 0.8 & 0.3 & 0.3 \\ 0.3 & 0.8 & 0.3 \\ 0.3 & 0.3 & 0.8 \end{bmatrix},$$

$$\Upsilon_2 = \begin{bmatrix} 0.5 & 0.1 & 0.1 \\ 0.1 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.5 \end{bmatrix},$$

$||\cdot||$ the Frobenius norm and $\bar{\boldsymbol{\epsilon}}_t$ multivariate Student's T random variable with 4 degrees of freedom and scale matrix:

$$\Sigma = \begin{bmatrix} 0.5 & 0.3 & 0.3 \\ 0.3 & 0.5 & 0.3 \\ 0.3 & 0.3 & 0.5 \end{bmatrix}.$$

In so doing, the VAR(2) process is stable since $det(I_3 - \bar{\Psi}_1 \cdot u - \bar{\Psi}_2 \cdot u^2) \neq 0 \ \forall |u| \leq 1$. A graphical representation of a replication with $T = 25$ is provided in Figure 4.1.

Seven models are taken into account:

- *Oracle Model* The point predictions are obtained by considering both $\boldsymbol{g}(q)$ and $\bar{\Psi}_1, \bar{\Psi}_2$ as known. In other words, the point prediction for $Y_{t,1}(q)$ is simply given by $\boldsymbol{g}'(q)(\bar{\Psi}_1\bar{\boldsymbol{y}}_{t-1} + \bar{\Psi}_2\bar{\boldsymbol{y}}_{t-2}) \ \forall t \in \{3, \ldots, T+1\}$.

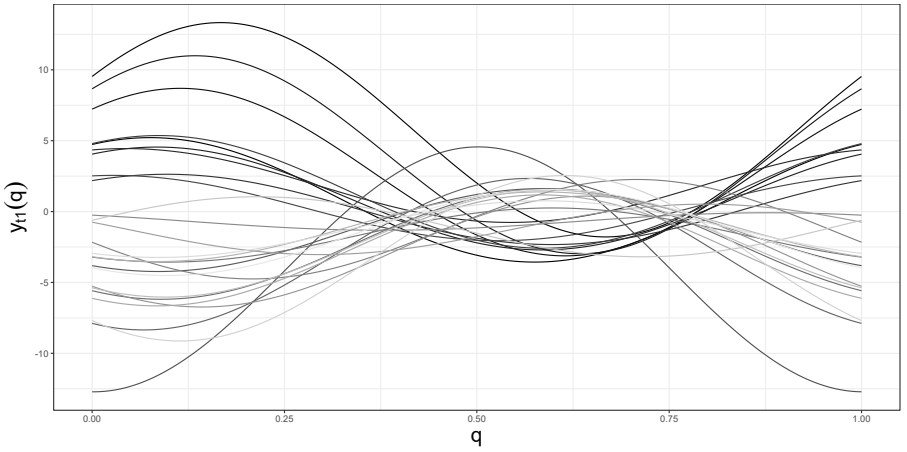

FIGURE 4.1: Graphical representation of the simulated data. The sample size is $T = 25$, with older functions being darker.

- *VAR Models* The point predictions are obtained by considering $\boldsymbol{g}(q)$ as known and $\bar{\Psi}_1, \bar{\Psi}_2$ as unknown. Specifically, the point prediction for $Y_{t,1}(q)$ is obtained by estimating r matrices $\bar{\Psi}_1, \ldots, \bar{\Psi}_r$ by fitting a VAR(r) model on $\{\bar{\boldsymbol{z}}_h : h \in \mathcal{I}_1\}$, with $\bar{\boldsymbol{z}}_h = (\bar{\boldsymbol{x}}_h, \bar{\boldsymbol{y}}_h)$, $\bar{\boldsymbol{x}}_h = \{\bar{\boldsymbol{y}}_{h-1}, \ldots, \bar{\boldsymbol{y}}_{h-r}\}$ and $r \in \{1, 2, 3\}$. In so doing, the estimation issue is converted to a problem of estimating the Fourier coefficients (Chen *et al.*, 2021).

- *FAR Models* The point predictions are obtained by considering both $\boldsymbol{g}(q)$ and $\bar{\Psi}_1, \bar{\Psi}_2$ as unknown. The point predictions are obtained by fitting (on the training set, as usual) a concurrent function-on-function autoregressive model of order $r \in \{1, 2, 3\}$, i.e.:

$$y_{t,1}(q) = \sum_{i=1}^{r} \beta_i(q) y_{t-i,1}(q) + a_t(q),$$

  with $a_t(q)$ finite-variance mean-zero error process uncorrelated with the linear systematic component and such that a_{t^1} is independent from a_{t^2}, $t^1 \neq t^2$.

The purpose is to evaluate the procedure by taking into account scenarios characterized by a decreasing knowledge of the data generating process. The first model (Oracle Model) represents the ideal case in which the oracle nonconformity scores can be computed since both $\bar{\Psi}_1, \bar{\Psi}_2$ and the subspace in which the observations lie are known, the second set of models (Var Models with $r \in \{1, 2, 3\}$) represents a more challenging case in which only the subspace in which the observations live is known, whereas the third set of models (FAR Models with $r \in \{1, 2, 3\}$) represents the general case in which the dynamic over time of the phenomenon must be derived by the available data. The simulation scheme here proposed allows to investigate, in addition to the Oracle Model

Empirical Coverage - Oracle Model		
$b = 1$	$T = 25$	0.753[0.737,0.769]
	$T = 50$	0.752[0.736,0.768]
	$T = 100$	0.748[0.732,0.764]
	$T = 1000$	0.743[0.727,0.759]
$b = 3$	$T = 50$	0.744[0.728,0.760]
	$T = 100$	0.738[0.722,0.754]
	$T = 1000$	0.743[0.727,0.759]
$b = 6$	$T = 100$	0.745[0.729,0.760]
	$T = 1000$	0.743[0.727,0.759]

TABLE 4.1: Empirical coverage (99% confidence interval in brackets). Oracle Model. α=0.25.

and the case in which the model is correctly specified (VAR Model with $r = 2$), three widespread kinds of model misspecification: the misspecification due to omitted relevant variable (VAR Model with $r = 1$, see Rao, 1971), misspecification due to inclusion of irrelevant variable (VAR Model with $r = 3$, see Rao, 1971) and the functional form misspecification (FAR Models, see Wooldridge, 1994, since the data generating process can not be rewritten as a FAR Model).

We consider $\mathcal{Q}_1 = [0, 1]$ and $\alpha = 0.25$. In order to fulfill the four conditions required ($(l + 1)/b$ integer-valued; $\alpha \in [b/(l + 1), 1)$; a training set size allowing to estimate a VAR(r) model; $\lfloor \alpha(l + 1)/b \rfloor b/(l + 1) = \alpha$ for consistency with the i.i.d. framework), we set $(T, l) = \{(25, 7), (50, 23), (100, 47), (1000, 479)\}$ when $b = 1$, $(T, l) = \{(50, 23), (100, 47), (1000, 479)\}$ when $b = 3$ and $(T, l) = \{(100, 47), (1000, 479)\}$ when $b = 6$. As usual in the time series setting, the first r observations (2 observations when the Oracle Model is considered, respectively) are taken into account only as covariates, and so the training set size is equal to $T - l - r$ ($T - l - 2$ when the Oracle Model is considered, respectively). Practically, for each value of T, we evaluate the procedure by considering $N = 5000$ replications for each combination of point predictor and value of b. The simulations are achieved by using the R Programming Language (R Core Team, 2020), and the generation of data by `fts.rar` function of `freqdom.fda` package (Hormann and Kidzinski, 2017).

4.3.1 Results

Table 4.1 and Table 4.2 show the empirical coverage, together with the related 99% confidence interval in square brackets, reached by the procedure presented in the manuscript. Specifically, the empirical coverage is simply computed as the fraction of the $N = 5000$ replications in which $\boldsymbol{y}_{T+1}$ belongs to $\mathcal{C}_{T,1-\alpha}(x_{T+1})$, and the confidence

Empirical Coverage - VAR Model				
		r=1	r=2	r=3
$b=1$	$T=25$	0.741[0.725,0.757]	0.735[0.719,0.751]	0.754[0.738,0.770]
	$T=50$	0.737[0.721,0.753]	0.746[0.730,0.762]	0.742[0.726,0.758]
	$T=100$	**0.731**[0.714,0.747]	0.743[0.727,0.759]	0.738[0.722,0.754]
	$T=1000$	0.743[0.727,0.759]	0.744[0.728,0.760]	0.742[0.726,0.758]
$b=3$	$T=50$	0.735[0.719,0.751]	0.742[0.726,0.758]	0.743[0.727,0.759]
	$T=100$	0.737[0.721,0.753]	0.738[0.722,0.754]	0.736[0.720,0.752]
	$T=1000$	0.743[0.727,0.759]	0.745[0.729,0.761]	0.744[0.728,0.760]
$b=6$	$T=100$	0.737[0.721,0.753]	0.740[0.724,0.756]	0.741[0.725,0.757]
	$T=1000$	0.746[0.730,0.762]	0.745[0.729,0.761]	0.743[0.727,0.759]
Empirical Coverage - FAR Model				
		r=1	r=2	r=3
$b=1$	$T=25$	0.745[0.729,0.761]	0.742[0.726,0.758]	0.741[0.725,0.757]
	$T=50$	0.738[0.722,0.754]	0.750[0.734,0.766]	0.747[0.731,0.763]
	$T=100$	**0.733**[0.717,0.749]	0.742[0.726,0.758]	0.740[0.724,0.756]
	$T=1000$	0.743[0.727,0.759]	0.742[0.726,0.758]	0.743[0.727,0.759]
$b=3$	$T=50$	0.736[0.720,0.752]	0.750[0.734,0.766]	0.742[0.726,0.758]
	$T=100$	0.736[0.720,0.752]	0.738[0.722,0.754]	0.743[0.727,0.759]
	$T=1000$	0.741[0.725,0.757]	0.744[0.728,0.760]	0.743[0.727,0.759]
$b=6$	$T=100$	0.736[0.720,0.752]	0.742[0.726,0.758]	0.740[0.724,0.756]
	$T=1000$	0.745[0.729,0.760]	0.744[0.728,0.760]	0.744[0.728,0.760]

TABLE 4.2: Empirical coverage (99% confidence interval in brackets). VAR Models and FAR Models. α=0.25. The values in bold indicate that the corresponding conf. intervals do not include $1-\alpha$.

interval is reported in order to provide an idea of the variability of the phenomenon, rather than to make inferential conclusion on the unconditional coverage in the various settings. The evidence is quite satisfactory as for any value of b, sample size T and model the empirical coverage is close to $1-\alpha=0.75$, as suggested by the fact that only 2 out of the 63 confidence intervals (the cells in bold in the tables) do not include the nominal confidence level. The result provided by the simulation study is particularly appealing since it suggests that an appropriate coverage is reached also when the sample size is very small, a fact that allows the procedure to be applied in many practical frameworks.

Although comparing the size of prediction bands obtained in scenarios characterized by (potentially) different unconditional coverages may lead to misleading conclusions, in light of the evidence provided so far we evaluate this aspect when the model, the value of T and the value of b vary. To do that, we define, according to the definition provided in Chapter 3 (see Equation 3.3), the size of a multivariate prediction band as the sum of the p areas between the upper and lower bound of the p univariate prediction bands, i.e. $\sum_{j=1}^{p}\int_{\mathcal{Q}_j} 2\cdot k^s\cdot s_j(q)dq$ (that, in this case, is simply $\int_{\mathcal{Q}_1} 2\cdot k^s\cdot s_1(q)dq$). Figure 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8 show the boxplots concerning the size of the $N=5000$

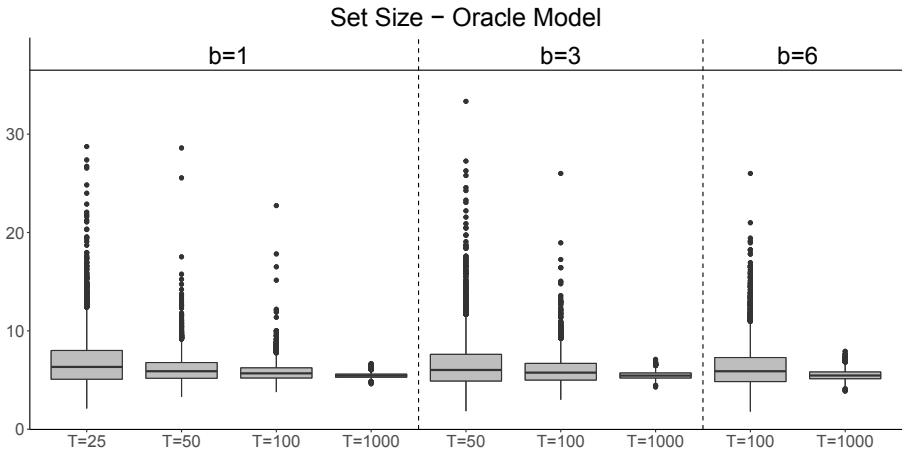

FIGURE 4.2: Set size. Oracle Model. $\alpha = 0.25$.

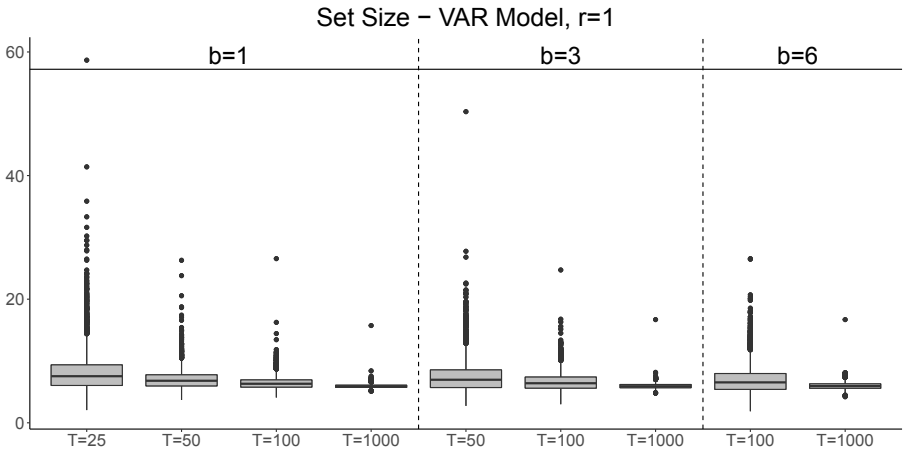

FIGURE 4.3: Set size. VAR Model, $r = 1$. $\alpha = 0.25$.

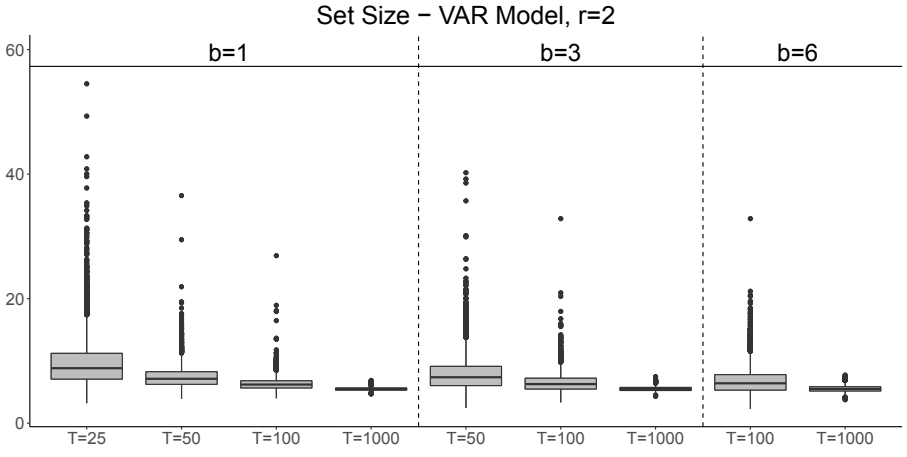

FIGURE 4.4: Set size. VAR Model, $r = 2$. $\alpha = 0.25$.

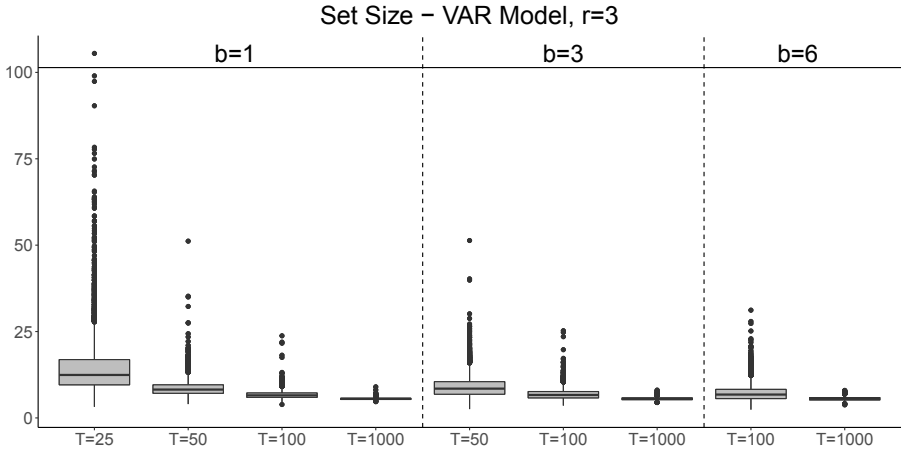

FIGURE 4.5: Set size. VAR Model, $r = 3$. $\alpha = 0.25$. For visualization purpose, the most extreme value (equal to 210.54) obtained when $b = 1$, $T = 25$ is removed.

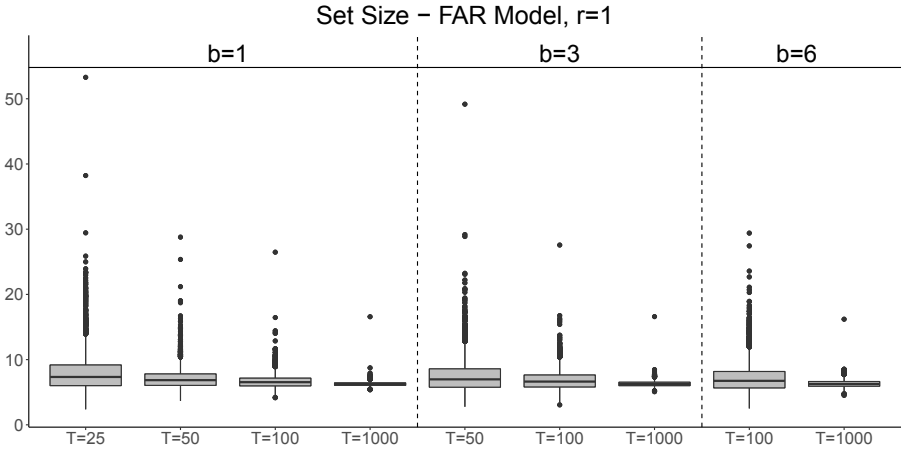

FIGURE 4.6: Set size. FAR Model, $r = 1$. $\alpha = 0.25$.

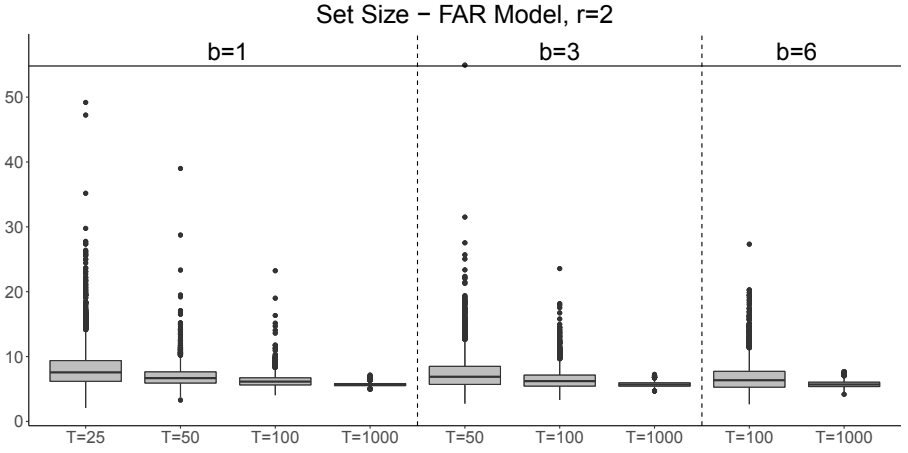

FIGURE 4.7: Set size. FAR Model, $r = 2$. $\alpha = 0.25$.

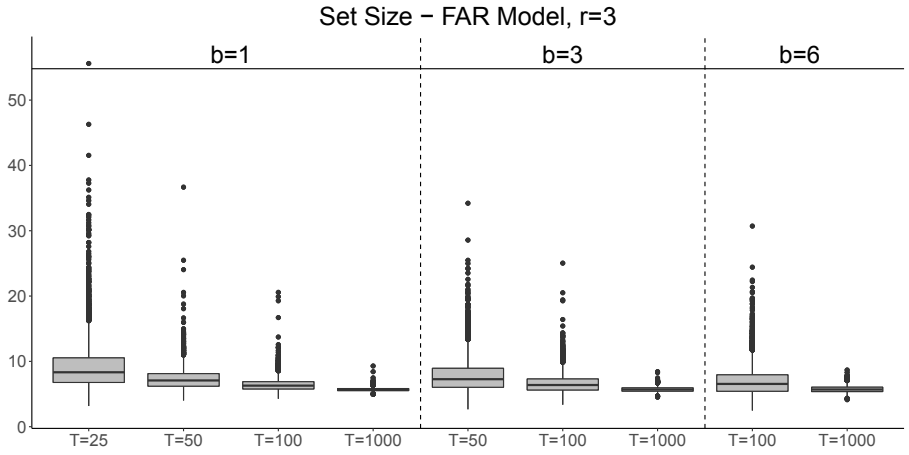

FIGURE 4.8: Set size. FAR Model, $r = 3$. $\alpha = 0.25$.

prediction bands. By considering each point predictor separately, it is possible to notice that, given b, the band size tends to decrease when T increases and, given T, it tends to decrease when b decreases: this evidence is not surprising since when T increases (and so l) and when b decreases a greater number of nonconformity scores is computed. Also the training set size has a relevant impact on the phenomenon since one is justified in expecting the band size to decrease when m grows because more accurate regression estimates provide smaller nonconformity scores, as suggested by analyzing the three couples $(T, b) = \{(25, 1), (50, 3), (100, 6)\}$ in which the number of nonconformity scores computed is constant.

The Oracle model outperforms the VAR models and the FAR models for every value of T, b, as expected. By considering the VAR models and the FAR models, when the sample size is very small (T=25) the order of the model providing the best performance in terms of size is $r = 1$ since higher values of r may provide unstable estimation procedures. Vice versa, the importance of using a model correctly specified is evident when T is large: indeed, the VAR Model with $r = 2$ represents the best choice overall (Oracle Model excluded) when $T = 1000$, and it outperforms the other two VAR models ($r = 1$, $r = 3$) also when $T = 100$. Specifically, when $T = 1000$ the VAR model with a relevant variable omitted ($r = 1$) is largely outperformed by the other two VAR models ($r = 1$, $r = 3$) since the estimation of a single matrix $\bar{\Psi}_1$ represents an undeniable limit in obtaining accurate regression estimates.

In light of the evidence provided in this Section, the procedure seems reliable in the frequent practical scenarios characterized by small sample size and/or model misspecification, whereas $b = 1$ represents the best balance between guarantee in terms of coverage and exhaustive use of the information provided by the available data.

4.4 Application to the Italian Gas Market

In this Section, we apply the procedure developed in Section 4.2 to a specific Italian gas market, namely the MGS (*Mercato Gas in Stoccaggio*), in order to create simultaneous prediction bands for the daily offer and demand curves. It should nevertheless be noted that our method can be applied to many application scenarios, such as other energy or non-energy markets. The MGS is a market in which users - authorized by the energy regulator *Gestore Mercati Energetici* (GME) - and the pipeline manager (Snam S.p.A.) day by day submit supply offers and demand bids for the gas stored, which is traded through an auction mechanism.

Specifically, for each day the supply offers (demand bids, respectively) are sorted by price in ascending (descending, respectively) order, and the demand and offer curves are built - starting from raw data provided in XML format by GME (`https://www.mercatoelettrico.org/en/`) - by considering the cumulative sum of the quantities (expressed in MWh). In doing so, by construction both daily offer and demand curves are positive monotonic (increasing and decreasing, respectively) step functions. The intersection of the two curves provides the price P_t at which the gas is traded (expressed in Euro/MWh) and the total quantity exchanged Q_t, and every offer/bid to the left of the intersection is accepted and consequently traded at price P_t.

The creation of prediction bands is strategic for energy traders' decision-making since it allows to evaluate the possible effect of offers/bids on the shape of the curves (and consequently on both the price P_t and the quantity exchanged Q_t), an aspect that cannot be directly included by usual non-functional procedures for interval price prediction. In order to show the useful insights that the procedure built in Section 4.2 can provide, we create simultaneous prediction bands for the offer ($Y_{t,1}$) and demand ($Y_{t,2}$) curve for each day in the six-month period between 1 August 2019 and 31 January 2020. For each of the 184 days we aim to predict, we build the corresponding prediction band based on the information provided by the rolling window of 90 days[2] (i.e. $T = 90$) including the most up-to-date information available. We set the function domains $\mathcal{Q}_1 = \mathcal{Q}_2 = [0, 2 \cdot 10^5]$ as all demand and offer curves are observed in this range and at the same time the total quantity exchanged Q_t always belongs to this interval in the period taken into account. The offer and demand curves considered in the analysis are displayed in Figure 4.9.

[2]We considered different values of T: 45,60,90,180,365. We chose $T = 90$ since it outputs the smallest prediction bands in the period considered.

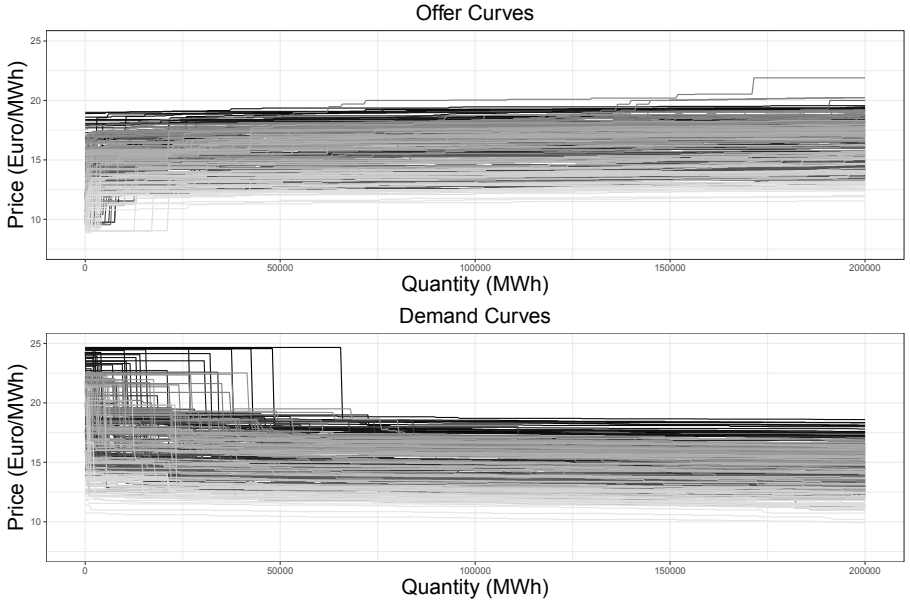

FIGURE 4.9: The offer (at the top) and demand (at the bottom) curves considered in the analysis, with older functions being darker.

In order to obtain the needed point predictions, we consider the following simple concurrent function-on-function autoregressive model with a scalar covariate:

$$y_{t,j}(q) = \beta_{1j}(q)y_{t-8,j}(q) + \beta_{2j}(q)P_{t-2} + a_t(q), \quad j \in \{1,2\}, \quad q \in \mathcal{Q}_j, \tag{4.3}$$

with $a_t(q)$ defined as in Section 4.3. The inclusion of the lagged curve at time $t-8$ and of the lagged (scalar) price at time $t-2$ is motivated by the fact that they represent the most up-to-date information available for a trader participating in the auction for day t due to GME's regulation. However, model (4.3) does not guarantee that the point predictions are monotonic functions. In view of this, after obtaining $\hat{\beta}_{1j}, \hat{\beta}_{2j}\ \forall j = 1,2$, we simply obtain monotonic point predictions by defining the point prediction for the offer curve at time t evaluated at q, i.e. $[\hat{\mu}^1_{\mathcal{I}_1}(x_t)](q)$, as follows:

$$[\hat{\mu}^1_{\mathcal{I}_1}(x_t)](q) = \begin{cases} \hat{y}_{t,1}(q) & \text{if } \hat{y}_{t,1}(q) = \max_{x\in[0,q]} \hat{y}_{t,1}(x) \\ \hat{y}_{t,1}(q') + (q-q')\left(\frac{\hat{y}_{t,1}(q'') - \hat{y}_{t,1}(q')}{q''-q'}\right) & \text{otherwise} \end{cases} \tag{4.4}$$

with $\hat{y}_{t,1}$ the predicted offer curve obtained by fitting model (4.3) using OLS, $q' := \max\left\{\text{argmax}_{x\in[0,q]}\hat{y}_{t,1}(x)\right\}$ and $q'' := \min\{x \in [q, 200000] | \hat{y}_{t,1}(x) \geq \hat{y}_{t,1}(q)\}$. The specular procedure is developed to obtain $[\hat{\mu}^2_{\mathcal{I}_1}(x_t)](q)$, i.e. the point prediction for the demand curve at time t evaluated at q. Vice versa, the fact that the point predictions are not step functions does not represent a limit since the prices at which the steps happen can

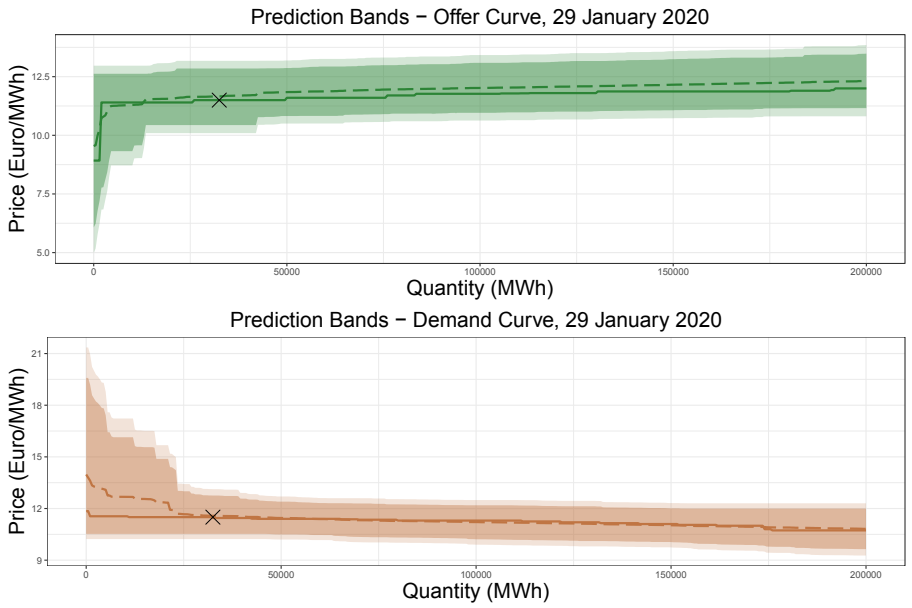

FIGURE 4.10: Multivariate prediction bands with $\alpha = 0.5$ (darker regions) and $\alpha = 0.25$ (lighter regions). The continuous lines represent the observed offer and demand curves, whereas the dashed lines represent the fitted ones. The black cross indicates (Q_t, P_t).

be absolutely continuous random variables, and consequently the expectation of this kind of random step function is a continuous function (Pelagatti, 2013). Model (4.3) and the correction induced by (4.4) certainly represent an oversimplification of the phenomenon analyzed, and one is justified in expecting other variables (such as the trading activity on the other energy markets) to have an important impact on the MGS's dynamics: however, the purpose of this Section is to illustrate the application potential of the procedure presented in Section 4.2 in a general and arbitrary prediction scenario, rather than when a particularly sophisticated model is built.

The last, fundamental step is the definition of the significance level α, of the calibration set size l, of the training set size m and of b. We consider two possible values of α, i.e. 0.5 and 0.25, and $b = 1$ in light of the evidence provided by Section 4.3. We also set $l = 39$ and, given the aforementioned delay in the information concerning lagged curves, we consider $m = T - l - 8 = 43$ in order to obtain a value of m/l close to 1 and a value of l such that $\lfloor \alpha(l+1)/b \rfloor b/(l+1) = \alpha$, as in Section 4.3.

Figure 4.10 shows the multivariate prediction bands obtained for one of the day we aim to predict (29 January 2020), with the panel at the top (at the bottom, respectively) showing the portions of the multivariate prediction bands related to the offer curve (demand curve, respectively): in both cases, the darker region indicates the prediction bands obtained by considering $\alpha = 0.5$ (i.e. nominal confidence level $1 - \alpha = 0.50$),

whereas the lighter one denotes those obtained by considering $\alpha = 0.25$ (i.e. nominal confidence level $1 - \alpha = 0.75$). For the sake of completeness, the observed (continuous line) and fitted (dashed line) curves, together with the price P_t and the quantity exchanged Q_t (black cross), are also displayed. Since the curves are monotonic by construction, the upper and the lower bounds of the prediction bands were made monotonic before being plotted: indeed, the procedure does not guarantee that such bounds are monotonic, but the fully nonparametric approach induced by the permutation scheme Π allows to made them monotonic by removing portions of the prediction bands associated to regions of the functional space that violate known features (e.g. monotonicity) of the function to be predicted, without decreasing the unconditional coverage. It is absolutely evident that the prediction bands are decidedly wider in the first part of the domain, especially in the panel at the bottom of the figure, and this is due to the fact that the behavior of the two curves in that portion of the domain is hardly predictable. The main reason of this phenomenon is the conduct of the pipeline manager Snam S.p.A.: indeed, in the period considered it typically submits extremely low supply offers and high demand bids - if compared to other traders' offers/bids - in order to be sure to sell/buy the quantity needed, but this makes the uncertainty quantification a particularly tough task if no information on Snam's trading intentions is available. As proof of that, we created a fictional scenario by removing all the offers/bids made by Snam in the period considered, and we therefore computed the corresponding 184 multivariate prediction bands in the period 1 August 2019-31 January 2020: in doing so, the size in the initial part of the domain [0,25000] of the two univariate prediction bands composing each multivariate prediction band (related to the offer and demand curve respectively, and formally defined as $\int_0^{25000} 2 \cdot k^s \cdot s_j(q)dq$, $j = 1, 2$) decreases by 45.2% (median value) for the offer curve and by 72.4% (median value) for the demand curve when $\alpha = 0.50$ is considered. A very similar result is obtained when $\alpha = 0.25$ is taken into account. In view of this, the inclusion in model (4.3) of information aimed at capturing Snam's behavior represents a possible future development that is highly likely to create smaller (and consequently more informative) prediction bands.

A further useful by-product of the procedure related to this specific application is that it allows to automatically obtain a prediction region for (Q_t, P_t) by considering the region in which the prediction band for the offer curve and that for the demand curve overlap. As an example, the region for 29 January 2020 computed with $\alpha = 0.50$ is represented in the left panel of Figure 4.11. By computing the fraction of times that the observed (Q_t, P_t) effectively belongs to the prediction region thus obtained over the 184 days considered, we obtain that 92.4% of the time the observed prediction region

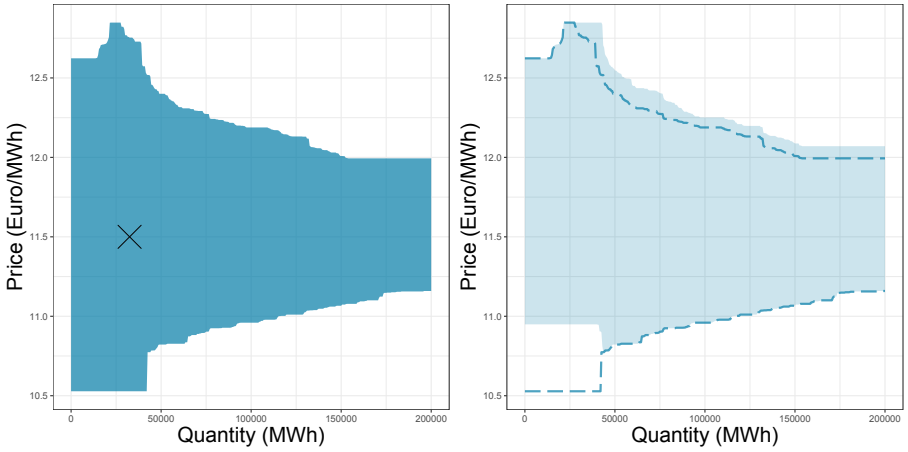

FIGURE 4.11: The left panel shows the prediction region for (Q_t, P_t) (29 January 2020, $\alpha = 0.50$), together with the value of (Q_t, P_t) effectively observed (black cross). The right panel shows the same prediction region (dashed darker lines) and the prediction region obtained by submitting an extra demand bid of 20000 MWh at 12 Euro/MWh (lighter region).

contains the observed intersection point when $\alpha = 0.50$, and that 97.8% of the time when $\alpha = 0.25$. This evidence is appealing especially when compared to the fraction of times that the observed offer and demand curves effectively belong to the observed multivariate prediction bands, that is 52.7% when $\alpha = 0.50$ (i.e. nominal confidence level $1 - \alpha = 0.50$) and 75.5% when $\alpha = 0.25$ (i.e. nominal confidence level $1 - \alpha = 0.75$), respectively. It is fundamental to notice that the last two percentages do not represent empirical coverages, and more generally provide no relevant information about the unconditional coverage reached by the procedure developed, since the prediction bands computed in this Section are obtained by repeating the procedure day after day and by considering a rolling window. However, it is still possible to obtain a theoretical result concerning the unconditional coverage of the aforementioned prediction region by simply reasoning about how it is built: indeed, by construction if the observed offer curve and the observed demand curve effectively jointly belong to the observed multivariate prediction band, and if the intersection point exists (an event always verified in the period considered), then the intersection point necessarily belongs to the area in which the two univariate prediction bands overlap. As a consequence, by construction, if the two curves intersect, than the unconditional coverage reached by the prediction region is greater or equal than $\mathbb{P}\left(\boldsymbol{Y}_{T+1} \in \mathcal{C}_{T,1-\alpha}\left(X_{T+1}\right)\right)$. In light of this and of the empirical results provided, we conclude that the prediction region naturally induced by the method described in Section 4.2 represents a promising tool - both from a theoretical and an application point of view - that can be profitably included in traders' tool kit.

The analysis here presented also allows to exploit the market from a speculative perspective: indeed, from a given trader's point of view, the procedure presented in this manuscript allows to directly evaluate the impact of any extra offer/bid on the prediction bands for tomorrow's offer and demand curves, and consequently on the prediction region for the intersection point. As an example, a trader may want to evaluate the effect of a demand bid of 20000 MWh at 12 Euro/MWh on tomorrow's intersection point: to do that, the user can add this bid to the predicted demand curve, thereby inducing a change in the multivariate prediction band and consequently in the prediction region. The resulting prediction region for 29 January 2020 is displayed in the right panel of Figure 4.11.

The evidence showed in this Section is obviously limited to a few examples. In order to provide a comprehensive overview of the results obtained in the period considered, we developed a Shiny app (available at `https://jacopodiquigiovanni.shinyapps.io/ItalianGasMarketApp/`) that allows to interactively choose the scenario of interest and to visualize the associated results. Specifically, the top left panel allows the user to dynamically change five inputs, that are: the day for which the predictions are made (between 1 August 2019 and 31 January 2020), the nominal confidence level $1 - \alpha$, the type of extra bid/offer (assuming two possible values: Demand, Offer) you want to evaluate the impact of, its quantity and its price. The top right panel and the bottom right panel show the portion of multivariate prediction band related to the offer and demand curve respectively, together with the observed curves, for the day and the nominal confidence level selected. In doing so, it is possible to verify whether the couple of curves would have been contained in the multivariate prediction band or not if the procedure had been implemented in the real world. Finally, the bottom left panel shows the prediction region for (Q_t, P_t) obtained for the day and the nominal confidence level selected (dark blue region), as well as the value of (Q_t, P_t) effectively observed (red cross), allowing the user to check the accuracy of the prediction. In addition, it shows also the prediction region obtained by including the extra bid/offer in tomorrow's predicted demand/offer curve (light blue region), allowing this tool to be used for the purposes mentioned above.

4.5 Conclusions and Further Developments

The present chapter deals with the demand for methods able to quantify uncertainty in the multivariate functional time series prediction framework. The approach developed in this chapter extends the non-overlapping blocking scheme proposed by Chernozhukov

et al. (2018) to the Split context in order to create simultaneous prediction bands for forthcoming multivariate random function $\boldsymbol{Y}_{T+1}$. The procedure inherits the guarantees for the unconditional coverage in terms of finite-sample performance bounds and of asymptotic exactness under some conditions concerning the oracle nonconformity measure A^* and the nonconformity measure A, but can be also satisfactorily applied to the multivariate functional context due to the Split process and to the specific nonconformity measure used. Theorem 4.1 provides a theoretically sound prediction framework based on assumptions similar to the ones introduced by Chernozhukov *et al.* (2018). However these assumptions could be tricky to assess in practice - and further work must be developed to find sufficient conditions that imply the conditions required by Theorem 4.1 . For this reason, we combined our theoretical work with a simulation study aimed at evaluating the procedure in situations in which Theorem 4.1 is violated, as in the case of model misspecification. The results obtained are encouraging: the empirical coverage values are close to the nominal confidence level $1 - \alpha$ also when the sample size T is small or the model is misspecified, regardless the value of b. In view of this, we applied the method described in Section 4.2 to a real-world scenario of strong interest, namely the prediction of daily offer and demand curves in the Italian gas market. We built the corresponding simultaneous prediction bands for each day in a six-month period based on a rolling window including the most up-to-date information available. Despite the fact that the point predictors considered can surely be improved to provide more accurate regression estimates, we used standard functional regression estimators to show the wide applicability of the procedure, also in a speculative perspective. In order to provide a complete overview of the study, we developed a Shiny app able to display the predicted bands, as well as the related prediction regions for (Q_t, P_t), under several operative conditions. Being based on a Split framework, our proposal shares both the strenghts (namely, the simple mathematical tractability and ease of implementation) and the weaknesses of the prediction methods based on Split Conformal. In fact the random subdivision of the sample intrinsically induces an element of randomness in the method and is not particularly efficient in its use of data. To improve on this, a very promising area of research is to employ derivations of the original Conformal approach such as the jackknife+ procedure (Barber *et al.*, 2021) and extensions (see, for example, Xu and Xie, 2020) in a functional context.

Appendix A

Appendix for Chapter 1

A.1 Proof of Theorem 1.1

Under the hypothesis of the theorem, $(l+1)\delta_{\boldsymbol{Y}} \sim U\{1,2,\ldots,l+1\}$ holds. Since $\mathcal{C}_{n,1-\alpha}(x_{n+1}) := \{\boldsymbol{y} \in \mathcal{Y}(\mathcal{T}) : \delta_{\boldsymbol{y}} > \alpha\}$, as a consequence:

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}(X_{n+1})) &= \mathbb{P}((l+1)\delta_{\boldsymbol{Y}} > (l+1)\alpha) \\ &= 1 - \mathbb{P}((l+1)\delta_{\boldsymbol{Y}} \leq (l+1)\alpha) \\ &= 1 - \frac{\lfloor (l+1)\alpha \rfloor}{l+1}.\end{aligned}$$

In addition, since

$$\frac{\lfloor (l+1)\alpha \rfloor}{l+1} \leq \frac{(l+1)\alpha}{l+1} = \alpha$$

then $\mathbb{P}(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}(X_{n+1})) \geq 1-\alpha$. Finally, since

$$\frac{\lfloor (l+1)\alpha \rfloor}{l+1} > \frac{(l+1)\alpha - 1}{l+1} = \alpha - \frac{1}{l+1}$$

then $\mathbb{P}(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}(X_{n+1})) < 1 - \alpha + \frac{1}{l+1}$.

A.2 Exactness of Smoothed Split Conformal prediction sets

Let us consider the hypothesis of Theorem 1.1. Let us notice that

$$\begin{aligned}\delta_{\boldsymbol{y},\tau_{n+1}} :=& \frac{|\{d \in \mathcal{I}_2 : R_d > R_{n+1}\}| + \tau_{n+1}\,|\{d \in \mathcal{I}_2 \cup \{n+1\} : R_d = R_{n+1}\}|}{l+1} \\ =& \frac{\tau_{n+1}}{l+1} + \frac{|\{d \in \mathcal{I}_2 : R_d \geq R_{n+1}\}|}{l+1}.\end{aligned}$$

Under the hypothesis of Theorem 1.1, $|\{d \in \mathcal{I}_2 : R_d \geq R_{n+1}\}| \sim U\{0, 1, \ldots, l\}$ holds. As a consequence:

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha,\tau_{n+1}}\left(x_{n+1}\right)|\tau_{n+1}\right) &= \mathbb{P}\left(\delta_{\boldsymbol{Y},\tau_{n+1}} > \alpha|\tau_{n+1}\right) \\
&= \mathbb{P}\left(|\{d \in \mathcal{I}_2 : R_d \geq R_{n+1}\}| > (l+1)\alpha - \tau_{n+1}|\tau_{n+1}\right) \\
&= 1 - \mathbb{P}\left(|\{d \in \mathcal{I}_2 : R_d \geq R_{n+1}\}| \leq (l+1)\alpha - \tau_{n+1}|\tau_{n+1}\right) \\
&= 1 - \frac{\lfloor (l+1)\alpha - \tau_{n+1} \rfloor + 1}{l+1}.
\end{aligned}$$

Let us call $f(\tau_{n+1}) = 1 \cdot \mathbb{1}\{\tau_{n+1} \in [0,1]\}$. Then

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha,\tau_{n+1}}\left(x_{n+1}\right)\right) &= \int_0^1 \mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha,\tau_{n+1}}\left(x_{n+1}\right)|\tau_{n+1}\right) f(\tau_{n+1}) d\tau_{n+1} \\
&= 1 - \\
&\quad \left(\int_0^{(l+1)\alpha - \lfloor (l+1)\alpha \rfloor} \frac{\lfloor (l+1)\alpha - \tau_{n+1} \rfloor + 1}{l+1} d\tau_{n+1} + \right. \\
&\quad \left. \int_{(l+1)\alpha - \lfloor (l+1)\alpha \rfloor}^1 \frac{\lfloor (l+1)\alpha - \tau_{n+1} \rfloor + 1}{l+1} d\tau_{n+1} \right).
\end{aligned}$$

Let us consider $\int_0^{(l+1)\alpha - \lfloor (l+1)\alpha \rfloor} \frac{\lfloor (l+1)\alpha - \tau_{n+1} \rfloor + 1}{l+1} d\tau_{n+1}$. Since if $\tau_{n+1} \leq (l+1)\alpha - \lfloor (l+1)\alpha \rfloor$ then $\lfloor (l+1)\alpha - \tau_{n+1} \rfloor = \lfloor (l+1)\alpha \rfloor$, we can notice that

$$\begin{aligned}
&\int_0^{(l+1)\alpha - \lfloor (l+1)\alpha \rfloor} \frac{\lfloor (l+1)\alpha - \tau_{n+1} \rfloor + 1}{l+1} d\tau_{n+1} \\
=& \int_0^{(l+1)\alpha - \lfloor (l+1)\alpha \rfloor} \frac{\lfloor (l+1)\alpha \rfloor + 1}{l+1} d\tau_{n+1} \\
=& \frac{\lfloor (l+1)\alpha \rfloor + 1}{l+1} \cdot \left((l+1)\alpha - \lfloor (l+1)\alpha \rfloor \right).
\end{aligned}$$

Let us consider $\int_{(l+1)\alpha-\lfloor(l+1)\alpha\rfloor}^{1} \frac{\lfloor(l+1)\alpha-\tau_{n+1}\rfloor+1}{l+1} d\tau_{n+1}$. Since if $\tau_{n+1} > (l+1)\alpha - \lfloor(l+1)\alpha\rfloor$ then $\lfloor(l+1)\alpha - \tau_{n+1}\rfloor = \lfloor(l+1)\alpha\rfloor - 1$, we can notice that

$$\begin{aligned}
&\int_{(l+1)\alpha-\lfloor(l+1)\alpha\rfloor}^{1} \frac{\lfloor(l+1)\alpha-\tau_{n+1}\rfloor+1}{l+1} d\tau_{n+1} \\
=&\int_{(l+1)\alpha-\lfloor(l+1)\alpha\rfloor}^{1} \frac{\lfloor(l+1)\alpha\rfloor}{l+1} d\tau_{n+1} \\
=&\frac{\lfloor(l+1)\alpha\rfloor}{l+1} \cdot (1 - ((l+1)\alpha - \lfloor(l+1)\alpha\rfloor)) .
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha,\tau_{n+1}}\left(x_{n+1}\right)\right) \\
=&1- \\
&\left(\frac{\lfloor(l+1)\alpha\rfloor+1}{l+1} \cdot ((l+1)\alpha - \lfloor(l+1)\alpha\rfloor) + \right. \\
&\left.\frac{\lfloor(l+1)\alpha\rfloor}{l+1} \cdot (1 - ((l+1)\alpha - \lfloor(l+1)\alpha\rfloor))\right) \\
=&1-\alpha.
\end{aligned}$$

Appendix B

Appendix for Chapter 2

B.1 Appendix for Chapter 2.2.1

Smoothed Split Conformal prediction set induced by nonconformity measure (2.1)

First of all let us notice that, by definition, $\mathcal{C}_{n,1-\alpha,\tau_{n+1}} = \mathcal{C}_{n,1-\alpha}$ when $\tau_{n+1} = 1$.

Since $\delta_{y,\tau_{n+1}}$ can not be less than $\tau_{n+1}/(l+1)$ and can not be greater than $(l + \tau_{n+1})/(l+1)$, we consider the case in which $\alpha \in [\tau_{n+1}/(l+1), (l+\tau_{n+1})/(l+1))$. Let us define w the $\lceil l + \tau_{n+1} - (l+1)\alpha \rceil$th smallest value in the set $\{R_d : d \in \mathcal{I}_2\}$, and r_n (v_n respectively) the number of elements in the set $\{R_d : d \in \mathcal{I}_2\}$ that are equal to w and that are to the right (left respectively) of w in the sorted version of the set. Under the assumption concerning the continuous joint distribution of $\{R_d : d \in \mathcal{I}_2\}$ made in Chapter 1 $r_n = v_n = 0$ holds, but generally speaking we assume $r_n, v_n \in \mathcal{N}_{\geq 0}$ such that $r_n + v_n \leq l - 1$. By performing calculations similar to those needed in the Split Conformal scenario, we obtain that:

- if
$$\tau_{n+1} > \frac{(l+1)\alpha - \lfloor (l+1)\alpha - \tau_{n+1} \rfloor + r_n}{r_n + v_n + 2}$$
then $y \in \mathcal{C}_{n,1-\alpha,\tau_{n+1}} \iff R_{n+1} \leq w$ and so
$$\mathcal{C}_{n,1-\alpha,\tau_{n+1}} = \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in [g_{\mathcal{I}_1}(t) - w, \\ g_{\mathcal{I}_1}(t) + w] \quad \forall t \in \mathcal{T}\}$$

- if
$$\tau_{n+1} \leq \frac{(l+1)\alpha - \lfloor (l+1)\alpha - \tau_{n+1} \rfloor + r_n}{r_n + v_n + 2}$$

then $y \in \mathcal{C}_{n,1-\alpha,\tau_{n+1}} \iff R_{n+1} < w$ and so

$$\mathcal{C}_{n,1-\alpha,\tau_{n+1}} = \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in \big(g_{\mathcal{I}_1}(t) - w, \\ g_{\mathcal{I}_1}(t) + w\big) \quad \forall t \in \mathcal{T}\}.$$

Proof that the concatenation of pointwise prediction intervals leads to a prediction band that is a subset of the simultaneous prediction band (2.2).

Let $\mathcal{U}_{n,1-\alpha}$ be the pointwise prediction set. Let us define $\tilde{R}_d(t) := |y_d(t) - g_{\mathcal{I}_1}(t)| \; \forall t \in \mathcal{T}, d \in \mathcal{I}_2$, $\tilde{R}_{n+1}(t) := |y(t) - g_{\mathcal{I}_1}(t)|$ for a given $y \in \mathcal{Y}(\mathcal{T})$ and $\tilde{k}(t)$ the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{\tilde{R}_d(t) : d \in \mathcal{I}_2\}$. By construction $R_d = \operatorname{ess\,sup}_{t \in \mathcal{T}} \tilde{R}_d(t)$, and so $R_d \geq \tilde{R}_d(t) \; \forall t \in \mathcal{T}, d \in \mathcal{I}_2$ and then $k \geq \tilde{k}(t) \; \forall t \in \mathcal{T}$. Let us consider $y \in \mathcal{U}_{n,1-\alpha}$, i.e. $y(t) \in [g_{\mathcal{I}_1}(t) - \tilde{k}(t), g_{\mathcal{I}_1}(t) + \tilde{k}(t)] \quad \forall t \in \mathcal{T}$. Since $k \geq \tilde{k}(t)$, also $y(t) \in [g_{\mathcal{I}_1}(t) - k, g_{\mathcal{I}_1}(t) + k] \quad \forall t \in \mathcal{T}$, i.e. $y \in \mathcal{C}_{n,1-\alpha}$.

Since the converse is not necessarily true (in the sense that $y \in \mathcal{C}_{n,1-\alpha}$ does not imply $y \in \mathcal{U}_{n,1-\alpha}$), we conclude that $\mathcal{U}_{n,1-\alpha} \subseteq \mathcal{C}_{n,1-\alpha}$.

B.2 Appendix for Chapter 2.2.2

Proof of the prediction set induced by the nonconformity measure $A(\{y_h : h \in \mathcal{I}_1\}, y) = \operatorname{ess\,sup}_{t \in \mathcal{T}} \left| \frac{y(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right|$.

Let us consider the Split Conformal framework. For a given $y \in \mathcal{Y}(\mathcal{T})$, let us define

$$\delta_y^s := \frac{\left|\left\{d \in \mathcal{I}_2 \cup \{n+1\} : R_d^s \geq R_{n+1}^s\right\}\right|}{l+1}.$$

The Split Conformal prediction set is defined as $\mathcal{C}_{n,1-\alpha}^s := \left\{y \in \mathcal{Y}(\mathcal{T}) : \delta_y^s > \alpha\right\}$. As a consequence, $y \in \mathcal{C}_{n,1-\alpha}^s \iff R_{n+1}^s \leq k^s$, with k^s the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d^s : d \in \mathcal{I}_2\}$. Then:

$$\begin{aligned} &\operatorname*{ess\,sup}_{t \in \mathcal{T}} \left| \frac{y(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right| \leq k^s \\ &\iff \left| \frac{y(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right| \leq k^s \quad \forall t \in \mathcal{T} \\ &\iff y(t) \in [g_{\mathcal{I}_1}(t) - k^s s_{\mathcal{I}_1}(t), g_{\mathcal{I}_1}(t) + k^s s_{\mathcal{I}_1}(t)] \quad \forall t \in \mathcal{T}. \end{aligned}$$

Therefore, the Split Conformal prediction set is

$$\mathcal{C}_{n,1-\alpha}^s := \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in [g_{\mathcal{I}_1}(t) - k^s s_{\mathcal{I}_1}(t), g_{\mathcal{I}_1}(t) + k^s s_{\mathcal{I}_1}(t)] \quad \forall t \in \mathcal{T}\}.$$

Let us consider the Smoothed Split Conformal framework. Let us define for a given $y \in \mathcal{Y}(\mathcal{T})$

$$\delta^s_{y,\tau_{n+1}} := \frac{\left|\left\{d \in \mathcal{I}_2 : R^s_d > R^s_{n+1}\right\}\right| + \tau_{n+1}\left|\left\{d \in \mathcal{I}_2 \cup \{n+1\} : R^s_d = R^s_{n+1}\right\}\right|}{l+1}$$
$$\mathcal{C}^s_{n,1-\alpha,\tau_{n+1}} := \left\{y \in \mathcal{Y}(\mathcal{T}) : \delta^s_{y,\tau_{n+1}} > \alpha\right\}.$$

By reconsidering the computations provided in Appendix B.1 and by substituting $\delta_{y,\tau_{n+1}}$ with $\delta^s_{y,\tau_{n+1}}$, w with w^s, R_d with R^s_d, r_n with r^s_n and v_n with v^s_n it is possible to notice that

- if
$$\tau_{n+1} > \frac{(l+1)\alpha - \lfloor (l+1)\alpha - \tau_{n+1} \rfloor + r^s_n}{r^s_n + v^s_n + 2}$$
then
$$\mathcal{C}^s_{n,1-\alpha,\tau_{n+1}} = \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in [g_{\mathcal{I}_1}(t) - w^s s_{\mathcal{I}_1}(t), g_{\mathcal{I}_1}(t) + w^s s_{\mathcal{I}_1}(t)] \quad \forall t \in \mathcal{T}\}$$

- if
$$\tau_{n+1} \leq \frac{(l+1)\alpha - \lfloor (l+1)\alpha - \tau_{n+1} \rfloor + r^s_n}{r^s_n + v^s_n + 2}$$
then
$$\mathcal{C}^s_{n,1-\alpha,\tau_{n+1}} = \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in \big(g_{\mathcal{I}_1}(t) - w^s s_{\mathcal{I}_1}(t), g_{\mathcal{I}_1}(t) + w^s s_{\mathcal{I}_1}(t)\big) \quad \forall t \in \mathcal{T}\}.$$

Proof of Remark 2.2.

Let us define $\mathcal{C}^{\lambda\cdot s}_{n,1-\alpha}$ the prediction set obtained by considering the modulation function $\lambda \cdot s_{\mathcal{I}_1}$. The nonconformity scores are

$$R^{\lambda\cdot s}_d = \operatorname*{ess\,sup}_{t\in\mathcal{T}} \left|\frac{y_d(t) - g_{\mathcal{I}_1}(t)}{\lambda \cdot s_{\mathcal{I}_1}(t)}\right| = \frac{1}{\lambda} R^s_d, \quad d \in \mathcal{I}_2$$
$$R^{\lambda\cdot s}_{n+1} = \operatorname*{ess\,sup}_{t\in\mathcal{T}} \left|\frac{y(t) - g_{\mathcal{I}_1}(t)}{\lambda \cdot s_{\mathcal{I}_1}(t)}\right| = \frac{1}{\lambda} R^s_{n+1}.$$

Let us also define

$$\delta^{\lambda\cdot s}_y := \frac{\left|\left\{d \in \mathcal{I}_2 \cup \{n+1\} : R^{\lambda\cdot s}_d \geq R^{\lambda\cdot s}_{n+1}\right\}\right|}{l+1}.$$

The Split Conformal prediction set is defined as $\mathcal{C}^{\lambda\cdot s}_{n,1-\alpha} := \{y \in \mathcal{Y}(\mathcal{T}) : \delta^{\lambda\cdot s}_y > \alpha\}$. As a consequence, $y \in \mathcal{C}^{\lambda\cdot s}_{n,1-\alpha} \iff R^{\lambda\cdot s}_{n+1} \leq k^{\lambda\cdot s}$, with $k^{\lambda\cdot s}$ the $\lceil (l+1)(1-\alpha)\rceil$th smallest value in the set $\{R^{\lambda\cdot s}_d : d \in \mathcal{I}_2\}$. In addition, since $R^{\lambda\cdot s}_d = R^s_d/\lambda \ \forall d \in \mathcal{I}_2$, then $k^{\lambda\cdot s} = k^s/\lambda$. Then:

$$\begin{aligned} & R^{\lambda\cdot s}_{n+1} \leq k^{\lambda\cdot s} \\ \iff & \frac{1}{\lambda} R^s_{n+1} \leq \frac{k^s}{\lambda} \\ \iff & R^s_{n+1} \leq k^s, \end{aligned}$$

and since $y \in \mathcal{C}^s_{n,1-\alpha} \iff R^s_{n+1} \leq k^s$, then $\mathcal{C}^{\lambda\cdot s}_{n,1-\alpha} = \mathcal{C}^s_{n,1-\alpha}$.

Adjustment procedure of $\bar{s}^c_{\mathcal{I}_1}$ ***and*** $\bar{s}_{\mathcal{I}_1}$

If $\max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| = 0$ for at least one value t but the condition $\int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt \neq 0$ still holds, in order to ensure that $\bar{s}^c_{\mathcal{I}_1}(t) > 0$ $\forall t \in \mathcal{T}$ it is sufficient to add an arbitrarily (small) positive value to $\bar{s}^c_{\mathcal{I}_1}(t)$ $\forall t \in \mathcal{T}$ and to adjust the normalization constant accordingly. The pathological case in which $\int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt = 0$ is addressed only when $y_d(t) = g_{\mathcal{I}_1}(t)$ $\forall d \in \mathcal{H}_2$ and almost every $t \in \mathcal{T}$ and it represents a case of no practical interest.

Should $\exists\ t \in \mathcal{T}$ such that $\max_{h\in\mathcal{H}_1} |y_h(t) - g_{\mathcal{I}_1}(t)| = 0$, the same procedure is developed.

Proof of Theorem 2.4.

Let us focus on $\bar{s}_{\mathcal{I}_1}(t)$. Since $m/n = \theta$ with $0 < \theta < 1$, if $n \to +\infty$ then $m \to +\infty$. By definition, the scalar γ is the empirical quantile of order $\lceil (m+1)(1-\alpha)\rceil$) of $\{\operatorname{ess\,sup}_{t\in\mathcal{T}} |y_h(t) - g_{\mathcal{I}_1}(t)| : h \in \mathcal{I}_1\}$. First of all note that

$$\lim_{m\to+\infty} \frac{\lceil (m+1)(1-\alpha)\rceil}{m} = \lim_{m\to+\infty} \frac{m+1-\lfloor (m+1)\alpha\rfloor}{m}$$

and since

$$\frac{(m+1)\alpha - 1}{m} \leq \frac{\lfloor (m+1)\alpha\rfloor}{m} \leq \frac{(m+1)\alpha}{m} \quad \forall m \in \mathbb{N}$$

and

$$\lim_{m\to+\infty} \frac{(m+1)\alpha - 1}{m} = \lim_{m\to+\infty} \frac{(m+1)\alpha}{m} = \alpha$$

then by the squeeze theorem (also known as the sandwich theorem) we obtain that

$$\lim_{m\to+\infty} \frac{\lfloor (m+1)\alpha \rfloor}{m} = \alpha$$

and then

$$\lim_{m\to+\infty} \frac{\lceil (m+1)(1-\alpha) \rceil)}{m} = 1-\alpha.$$

As a consequence, γ is the empirical quantile of order $1-\alpha$ when $m \to +\infty$.

For convenience, let us define $x_h := \operatorname{ess\,sup}_{t\in\mathcal{T}} |y_h(t) - g_{\mathcal{I}_1}(t)| \ \forall\ h \in \mathcal{I}_1$. The random variables $\{X_h : h \in \mathcal{I}_1\}$ from which $\{x_h : h \in \mathcal{I}_1\}$ are drawn are continuous and they are asymptotically i.i.d. as $\mathrm{Var}[g_{\mathcal{I}_1}(t)] \to 0$. The Glivenko-Cantelli theorem ensures that the empirical distribution function of these variables converges uniformly (and almost surely pointwise) to its distribution function, and then also the empirical quantiles converge in distribution (and so in probability) to the corresponding theoretical quantiles, as shown for example by Van der Vaart (2000, chap. 21). Specifically, empirical quantile γ converges to $q_{1-\alpha}$, the theoretical quantile of order $1-\alpha$. As a consequence, when $m \to +\infty$:

$$\mathcal{H}_1 := \{h \in \mathcal{I}_1 : \operatorname*{ess\,sup}_{t\in\mathcal{T}} |y_h(t) - g_{\mathcal{I}_1}(t)| \le q_{1-\alpha}\}$$

with $q_{1-\alpha}$ deterministic quantity. Let us focus on the numerator of $\bar{s}_{\mathcal{I}_1}(t)$ since the denominator is just a normalizing constant. $\forall t \in \mathcal{T}$, the sequence $\{\max_{h\in\mathcal{H}_1} |y_h(t) - g_{\mathcal{I}_1}(t)|\}_m$ is eventually bounded by $q_{1-\alpha}$ and is eventually increasing since $\{|\mathcal{H}_1|\}_m$ is eventually increasing. By the monotone convergence theorem, the sequence converges to its supremum.

In order to prove the convergence of the numerator of $\bar{s}^c_{\mathcal{I}_1}$ to the same limit function, it is sufficient to consider the previous computations by noting that if $n \to +\infty$ then $l = n(1-\theta) \to +\infty$ and by substituting γ with k, m with l, $\mathcal{H}_1$ with $\mathcal{H}_2$ and $\mathcal{I}_1$ with $\mathcal{I}_2$ (except for $g_{\mathcal{I}_1}$ that is naturally not substituted by $g_{\mathcal{I}_2}$). Since the numerators of $\bar{s}_{\mathcal{I}_1}$ and $\bar{s}^c_{\mathcal{I}_1}$ converge to the same function, also the two normalizing constants converge to the same quantity.

Proof of Theorem 2.5.

The proof consists of two steps. At the first step we show that $k^{\bar{s}^c} = \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt$, a fundamental result to obtain, at the second step, the proof of the theorem.

I step

In order not to overcomplicate the proof, first of all let us consider the case in which

$|\mathcal{H}_2| = \lceil (l+1)(1-\alpha) \rceil$. It is important to notice that under the assumption concerning the continuous joint distribution of $\{R_d : d \in \mathcal{I}_2\}$ made in Section 1 such condition is always satisfied. However, the result proved at this first step holds also when this assumption is violated, and its proof requires just minor changes. Therefore, for the sake of completeness such proof is addressed below.

- $\forall i \in \mathcal{H}_2$ the following relationship holds $\forall t \in \mathcal{T}$:

$$\begin{aligned} &\left| \frac{y_i(t) - g_{\mathcal{I}_1}(t)}{\bar{s}^c_{\mathcal{I}_1}(t)} \right| \\ &= \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt \cdot \frac{|y_i(t) - g_{\mathcal{I}_1}(t)|}{\max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|} \\ &\leq \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt, \end{aligned}$$

  and then

$$R_i^{\bar{s}^c} := \operatorname*{ess\,sup}_{t \in \mathcal{T}} \left| \frac{y_i(t) - g_{\mathcal{I}_1}(t)}{\bar{s}^c_{\mathcal{I}_1}(t)} \right| \leq \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt.$$

  Specifically, $\exists\ \underline{i} \in \mathcal{H}_2$ such that $R_{\underline{i}}^{\bar{s}^c} = \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$ since $\forall t \in \mathcal{T}$ at least one function $y_{\underline{i}}$ satisfies $|y_{\underline{i}}(t) - g_{\mathcal{I}_1}(t)| = \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|$.

- Let us define $\mathcal{CH}_2 := \mathcal{I}_2 \setminus \mathcal{H}_2$ and let t_d^* be the value such that

$$|y_d(t_d^*) - g_{\mathcal{I}_1}(t_d^*)| = \operatorname*{ess\,sup}_{t \in \mathcal{T}} |y_d(t) - g_{\mathcal{I}_1}(t)| \quad \forall d \in \mathcal{I}_2.$$

  If t_d^* is not unique, it is randomly chosen from the values that satisfy that condition. $\forall i \in \mathcal{CH}_2$, by definition of $\mathcal{H}_2$ we obtain that $|y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)| > \max_{d \in \mathcal{H}_2} |y_d(t_i^*) - g_{\mathcal{I}_1}(t_i^*)|$ and so the following relationship holds:

$$\begin{aligned} &\left| \frac{y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)}{\bar{s}^c_{\mathcal{I}_1}(t_i^*)} \right| \\ &= \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt \cdot \frac{|y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)|}{\max_{d \in \mathcal{H}_2} |y_d(t_i^*) - g_{\mathcal{I}_1}(t_i^*)|} \\ &> \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt. \end{aligned}$$

  As a consequence,

$$R_i^{\bar{s}^c} := \operatorname*{ess\,sup}_{t \in \mathcal{T}} \left| \frac{y_i(t) - g_{\mathcal{I}_1}(t)}{\bar{s}^c_{\mathcal{I}_1}(t)} \right| > \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt.$$

Since:

- $|\mathcal{H}_2| = \lceil (l+1)(1-\alpha) \rceil$
- $\forall i \in \mathcal{H}_2 \; R_i^{\bar{s}^c} \leq \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$ and $\exists \; \underline{i} \in \mathcal{H}_2$ such that $R_{\underline{i}}^{\bar{s}^c} = \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$
- $\forall i \in \mathcal{CH}_2 \; R_i^{\bar{s}^c} > \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$

we conclude that $k^{\bar{s}^c} = \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$, with $k^{\bar{s}^c}$ the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d^{\bar{s}^c} : d \in \mathcal{I}_2\}$.

If $|\mathcal{H}_2| > \lceil (l+1)(1-\alpha) \rceil$, then $R_i^{\bar{s}^c} = \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$ is valid $\forall i \in \mathcal{H}_2$ such that $\operatorname{ess\,sup}_{t \in \mathcal{T}} |y_i(t) - g_{\mathcal{I}_1}(t)| = k$ and in the same way we can conclude that $k^{\bar{s}^c} = \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$.

II step

Let us define $\forall d \in \mathcal{I}_2$

$$R_d^{s^0} := \operatorname*{ess\,sup}_{t \in \mathcal{T}} \left| \frac{y_d(t) - g_{\mathcal{I}_1}(t)}{s^0(t)} \right| = |\mathcal{T}| \operatorname*{ess\,sup}_{t \in \mathcal{T}} |y_d(t) - g_{\mathcal{I}_1}(t)| \, .$$

Since k^{s^0} is the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d^{s^0} : d \in \mathcal{I}_2\}$, by definition of $\mathcal{H}_2$ we obtain that

$$\begin{aligned} k^{s^0} &= |\mathcal{T}| \max_{d \in \mathcal{H}_2} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}} |y_d(t) - g_{\mathcal{I}_1}(t)| \right) \\ &= |\mathcal{T}| \operatorname*{ess\,sup}_{t \in \mathcal{T}} \left(\max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| \right) . \end{aligned}$$

Since at the first step we proved that $k^{\bar{s}^c} = \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$, we obtain that

$$k^{s^0} - k^{\bar{s}^c} = |\mathcal{T}| \operatorname*{ess\,sup}_{t \in \mathcal{T}} \left(\max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| \right) - \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt.$$

Since the right side of the equation is greater than or equal to 0 by the integral mean value theorem, then $\mathcal{Q}(s^0) \geq \mathcal{Q}(\bar{s}^c_{\mathcal{I}_1})$.

The same theorem ensures that

$$\begin{aligned} |\mathcal{T}| \operatorname*{ess\,sup}_{t \in \mathcal{T}} \left(\max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| \right) &= \int_{\mathcal{T}} \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt \\ \iff \max_{d \in \mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| &\quad \text{is constant almost everywhere,} \end{aligned}$$

i.e. if and only if $\bar{s}^c_{\mathcal{I}_1}(t) = \bar{s}^0(t)$ almost everywhere.

Proof of Theorem 2.6.

We have already shown at the first step of previous proof ot Theorem 2.5 that $k^{\bar{s}^c} = \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt$. Since by assumption $s^{\zeta}_{\mathcal{I}_1}(t^*_i) \leq \bar{s}^c_{\mathcal{I}_1}(t^*_i)\ \forall i \in \mathcal{CH}_2$ and $|\mathcal{H}_2| = \lceil (l+1)(1-\alpha) \rceil$, let us define $a_i \geq 0\ \forall i \in \mathcal{CH}_2$ the value such that $s^{\zeta}_{\mathcal{I}_1}(t^*_i) = \bar{s}^c_{\mathcal{I}_1}(t^*_i) - a_i$.

- *Case 1*: If $\exists\ x \in \mathcal{CH}_2$ s.t. $a_x > 0$, $\exists\ \underline{i} \in \mathcal{H}_2$ such that

$$
\begin{aligned}
&\left| \frac{y_{\underline{i}}(t^*_x) - g_{\mathcal{I}_1}(t^*_x)}{s^{\zeta}_{\mathcal{I}_1}(t^*_x)} \right| \\
&= \left| \frac{y_{\underline{i}}(t^*_x) - g_{\mathcal{I}_1}(t^*_x)}{\bar{s}^c_{\mathcal{I}_1}(t^*_x) - a_x} \right| \\
&= \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt \quad \times \\
&\quad \frac{|y_{\underline{i}}(t^*_x) - g_{\mathcal{I}_1}(t^*_x)|}{\max_{d\in\mathcal{H}_2} |y_d(t^*_x) - g_{\mathcal{I}_1}(t^*_x)| - a_x \cdot \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt} \\
&> \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt
\end{aligned}
$$

since $\forall t \in \mathcal{T}$ (and specifically for t^*_x) at least one function $y_{\underline{i}}$ satisfies $|y_{\underline{i}}(t) - g_{\mathcal{I}_1}(t)| = \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|$.

Case 2: If $a_i = 0\ \forall i \in \mathcal{CH}_2$, there exist at least two values $t_{\downarrow}, t_{\uparrow} \in \mathcal{T}^*$ such that $s^{\zeta}_{\mathcal{I}_1}(t_{\downarrow}) < \bar{s}^c_{\mathcal{I}_1}(t_{\downarrow})$ and $s^{\zeta}_{\mathcal{I}_1}(t_{\uparrow}) > \bar{s}^c_{\mathcal{I}_1}(t_{\uparrow})$ since otherwise $s^{\zeta}_{\mathcal{I}_1}(t) = \bar{s}^c_{\mathcal{I}_1}(t)\ \forall t \in \mathcal{T}^*$. Let us define $a_{\downarrow} > 0$ the value such that $s^{\zeta}_{\mathcal{I}_1}(t_{\downarrow}) = \bar{s}^c_{\mathcal{I}_1}(t_{\downarrow}) - a_{\downarrow}$. Therefore $\exists\ \underline{i} \in \mathcal{H}_2$ such that

$$
\begin{aligned}
&\left| \frac{y_{\underline{i}}(t_{\downarrow}) - g_{\mathcal{I}_1}(t_{\downarrow})}{s^{\zeta}_{\mathcal{I}_1}(t_{\downarrow})} \right| \\
&= \left| \frac{y_{\underline{i}}(t_{\downarrow}) - g_{\mathcal{I}_1}(t_{\downarrow})}{\bar{s}^c_{\mathcal{I}_1}(t_{\downarrow}) - a_{\downarrow}} \right| \\
&= \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt \quad \times \\
&\quad \frac{|y_{\underline{i}}(t_{\downarrow}) - g_{\mathcal{I}_1}(t_{\downarrow})|}{\max_{d\in\mathcal{H}_2} |y_d(t_{\downarrow}) - g_{\mathcal{I}_1}(t_{\downarrow})| - a_{\downarrow} \cdot \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt} \\
&> \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|dt
\end{aligned}
$$

since $\forall t \in \mathcal{T}$ (and specifically for $t_{\downarrow}$) at least one function $y_{\underline{i}}$ satisfies $|y_{\underline{i}}(t) - g_{\mathcal{I}_1}(t)| = \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)|$.

As a consequence, in both cases ($\exists x \in \mathcal{CH}_2$ s.t. $a_x > 0$ and $a_i = 0\ \forall i \in \mathcal{CH}_2$) we obtain that $\exists\ \underline{i} \in \mathcal{H}_2$ such that

$$R_{\underline{i}}^{s^\zeta} := \operatorname*{ess\,sup}_{t\in\mathcal{T}} \left| \frac{y_{\underline{i}}(t) - g_{\mathcal{I}_1}(t)}{s^\zeta_{\mathcal{I}_1}(t)} \right| > \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt.$$

- $\forall i \in \mathcal{CH}_2$, by definition of $\mathcal{H}_2$ we obtain that $|y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)| > \max_{d\in\mathcal{H}_2} |y_d(t_i^*) - g_{\mathcal{I}_1}(t_i^*)|$ and so the following relationship holds:

$$\begin{aligned}
&\left| \frac{y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)}{s^\zeta_{\mathcal{I}_1}(t_i^*)} \right| \\
&= \left| \frac{y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)}{\bar{s}^c_{\mathcal{I}_1}(t_i^*) - a_i} \right| \\
&= \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt \quad \times \\
&\quad \frac{|y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)|}{\max_{d\in\mathcal{H}_2} |y_d(t_i^*) - g_{\mathcal{I}_1}(t_i^*)| - a_i \cdot \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt} \\
&> \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt.
\end{aligned}$$

As a consequence,

$$R_i^{s^\zeta} := \operatorname*{ess\,sup}_{t\in\mathcal{T}} \left| \frac{y_i(t) - g_{\mathcal{I}_1}(t)}{s^\zeta_{\mathcal{I}_1}(t)} \right| > \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt.$$

Since:

- $|\mathcal{H}_2| = \lceil (l+1)(1-\alpha) \rceil$
- $\exists\ \underline{i} \in \mathcal{H}_2$ such that $R_{\underline{i}}^{s^\zeta} > \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$
- $\forall i \in \mathcal{CH}_2\ R_i^{s^\zeta} > \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$

we conclude that $k^{s^\zeta} > \int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt$, i.e. $k^{s^\zeta} > k^{\bar{s}^c}$, with k^{s^ζ} the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d^{s^\zeta} : d \in \mathcal{I}_2\}$.

Proof that Theorem 2.6 does not imply Theorem 2.5.

Theorem 2.6 does not imply Theorem 2.5 since s^0 may not fulfill $s^0(t_i^*) \leq \bar{s}^c_{\mathcal{I}_1}(t_i^*)$ $\forall i \in \mathcal{CH}_2$. In fact, $\forall i \in \mathcal{CH}_2$:

$$s^0(t_i^*) \leq \bar{s}^c_{\mathcal{I}_1}(t_i^*) \iff \frac{\int_{\mathcal{T}} \max_{d\in\mathcal{H}_2} |y_d(t) - g_{\mathcal{I}_1}(t)| dt}{|\mathcal{T}|} \leq \max_{d\in\mathcal{H}_2} |y_d(t_i^*) - g_{\mathcal{I}_1}(t_i^*)|$$

and the condition on the right side is not always satisfied because no constraints are imposed on $y_d(t_i^*)$, with $d \in \mathcal{H}_2$, $i \in \mathcal{CH}_2$.

Generalization of functions (2.6)**,** (2.7)***, Theorems 2.4, 2.5 and 2.6 to the Smoothed Split Conformal framework***.

The functions $\bar{s}^c_{\mathcal{I}_1}$ and $\bar{s}_{\mathcal{I}_1}$ are defined as in Section 2.2.2 except for k (γ respectively) that is the $\lceil l+\tau_{n+1}-(l+1)\alpha \rceil$th ($\lceil m+\tau_{n+1}-(m+1)\alpha \rceil$th respectively) smallest value in the corresponding set; similarly, if $\lceil m+\tau_{n+1}-(m+1)\alpha \rceil > m$ then $\mathcal{H}_1 = \mathcal{I}_1$ and if $\lceil m+\tau_{n+1}-(m+1)\alpha \rceil \leq 0$ we arbitrarily set $\bar{s}_{\mathcal{I}_1} = s^0$. The theorems of Section 2.2.2 still hold by substituting $\lceil (l+1)(1-\alpha) \rceil, \lceil (m+1)(1-\alpha) \rceil$ with $\lceil l+\tau_{n+1}-(l+1)\alpha \rceil, \lceil m+\tau_{n+1}-(m+1)\alpha \rceil$.

Appendix C

Appendix for Chapter 3

C.1 Appendix for Chapter 3.2.1

Computation to find $\mathcal{C}_{n,1-\alpha}(x_{n+1})$

Since

$$\delta_{\boldsymbol{y}} = \frac{|\{d \in \mathcal{I}_2 \cup \{n+1\} : R_d \geq R_{n+1}\}|}{l+1},$$
$$\mathcal{C}_{n,1-\alpha}(x_{n+1}) = \left\{ \boldsymbol{y} \in \prod_{j=1}^{p} L^{\infty}(\mathcal{T}_j) : \delta_{\boldsymbol{y}} > \alpha \right\},$$

if $\alpha \in [1/(l+1), 1)$, then $\boldsymbol{y} \in \mathcal{C}_{n,1-\alpha}(x_{n+1}) \iff R_{n+1} \leq k^s$, with k^s the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d : d \in \mathcal{I}_2\}$. Then

$$\begin{aligned}
&\sup_{j \in \{1,\ldots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left| \frac{y_j(t) - [\hat{\mu}^j(x_{n+1})](t)}{s_j(t)} \right| \right) \leq k^s \\
\iff &\left| \frac{y_j(t) - [\hat{\mu}^j(x_{n+1})](t)}{s_j(t)} \right| \leq k^s \quad \forall j \in \{1,\ldots,p\}, \forall t \in \mathcal{T}_j \\
\iff &y_j(t) \in \big[[\hat{\mu}^j(x_{n+1})](t) - k^s \cdot s_j(t), \\
&\qquad [\hat{\mu}^j(x_{n+1})](t) + k^s \cdot s_j(t) \big] \quad \forall j \in \{1,\ldots,p\}, \forall t \in \mathcal{T}_j.
\end{aligned}$$

As a consequence, the Split Conformal prediction set is

$$\begin{aligned}
\mathcal{C}_{n,1-\alpha}(x_{n+1}) := \Bigg\{ \boldsymbol{y} \in \prod_{j=1}^{p} L^{\infty}(\mathcal{T}_j) : y_j(t) \in \big[&[\hat{\mu}^j(x_{n+1})](t) - k^s \cdot s_j(t), \\
&[\hat{\mu}^j(x_{n+1})](t) + k^s \cdot s_j(t) \big] \\
&\forall j \in \{1,\ldots,p\}, \forall t \in \mathcal{T}_j \Bigg\}.
\end{aligned}$$

Computation to find $\mathcal{C}_{n,1-\alpha,\tau_{n+1}}(x_{n+1})$

Consistently with the Split Conformal scenario, let us define

$$\delta_{\boldsymbol{y},\tau_{n+1}} := \frac{|\{d \in \mathcal{I}_2 : R_d > R_{n+1}\}| + \tau_{n+1}\,|\{d \in \mathcal{I}_2 \cup \{n+1\} : R_d = R_{n+1}\}|}{l+1}$$

$$\mathcal{C}_{n,1-\alpha,\tau_{n+1}}(x_{n+1}) := \left\{ \boldsymbol{y} \in \prod_{j=1}^{p} L^{\infty}(\mathcal{T}_j) : \delta_{\boldsymbol{y},\tau_{n+1}} > \alpha \right\}.$$

By definition, $\mathcal{C}_{n,1-\alpha,1}(x_{n+1}) = \mathcal{C}_{n,1-\alpha}(x_{n+1})$.

Since $\delta_{\boldsymbol{y},\tau_{n+1}} \in [\tau_{n+1}/(l+1), (l+\tau_{n+1})/(l+1)]$, we will focus on the scenario in which $\alpha \in [\tau_{n+1}/(l+1), (l+\tau_{n+1})/(l+1))$. Let us define w^s the $\lceil l+\tau_{n+1}-(l+1)\alpha \rceil$th smallest value in the set $\{R_d : d \in \mathcal{I}_2\}$, and r_n^s (v_n^s respectively) the number of elements in the set $\{R_d : d \in \mathcal{I}_2\}$ that are equal to w^s and that are to the right (left respectively) of w^s in the sorted version of the set. Note that $r_n^s = v_n^s = 0$ when the assumption about the continuous joint distribution of $\{R_d : d \in \mathcal{I}_2\}$ is satisfied, but generally speaking we will consider $r_n^s, v_n^s \in \mathcal{N}_{\geq 0}$ such that $r_n^s + v_n^s \leq l-1$. By replicating calculations similar to those performed in the Split Conformal framework, we obtain that:

- if
$$\tau_{n+1} > \frac{(l+1)\alpha - \lfloor (l+1)\alpha - \tau_{n+1} \rfloor + r_n^s}{r_n^s + v_n^s + 2}$$
then $\boldsymbol{y} \in \mathcal{C}_{n,1-\alpha,\tau_{n+1}}(x_{n+1}) \iff R_{n+1} \leq w^s$ and so
$$\mathcal{C}_{n,1-\alpha,\tau_{n+1}}(x_{n+1}) := \Bigg\{ \boldsymbol{y} \in \prod_{j=1}^{p} L^{\infty}(\mathcal{T}_j) : y_j(t) \in \big[[\hat{\mu}^j(x_{n+1})](t) - w^s \cdot s_j(t), [\hat{\mu}^j(x_{n+1})](t) + w^s \cdot s_j(t) \big] \; \forall j \in \{1,\dots,p\}, \forall t \in \mathcal{T}_j \Bigg\}.$$

- if
$$\tau_{n+1} \leq \frac{(l+1)\alpha - \lfloor (l+1)\alpha - \tau_{n+1} \rfloor + r_n^s}{r_n^s + v_n^s + 2}$$
then $\boldsymbol{y} \in \mathcal{C}_{n,1-\alpha,\tau_{n+1}}(x_{n+1}) \iff R_{n+1} < w^s$ and so
$$\mathcal{C}_{n,1-\alpha,\tau_{n+1}}(x_{n+1}) := \Bigg\{ \boldsymbol{y} \in \prod_{j=1}^{p} L^{\infty}(\mathcal{T}_j) : y_j(t) \in \big([\hat{\mu}^j(x_{n+1})](t) - w^s \cdot s_j(t), [\hat{\mu}^j(x_{n+1})](t) + w^s \cdot s_j(t) \big) \; \forall j \in \{1,\dots,p\}, \forall t \in \mathcal{T}_j \Bigg\}.$$

Proof that prediction bands induced by $\{s_{j,}\}_{j=1}^{p}$ ***and by*** $\{\lambda \cdot s_j\}_{j=1}^{p}$ ***coincide*** $\forall \lambda \in \mathbb{R}_{>0}$

Let $\mathcal{C}_{n,1-\alpha}^{\lambda \cdot s}(x_{n+1})$ ($\mathcal{C}_{n,1-\alpha}^{s}(x_{n+1})$ respectively) be the prediction band induced by the set of modulation functions $\{\lambda \cdot s_j\}_{j=1}^{p}$ ($\{s_j\}_{j=1}^{p}$ respectively) and $R_d^{\lambda \cdot s}$ (R_d^s respectively) the nonconformity score induced by the corresponding set of modulation funcions. The nonconformity scores induced by $\{\lambda \cdot s_j\}_{j=1}^{p}$ are:

$$R_d^{\lambda \cdot s} = \sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left| \frac{y_{dj}(t) - [\hat{\mu}^j(x_{dj})](t)}{\lambda \cdot s_j(t)} \right| \right) = \frac{1}{\lambda} R_d^s, \quad d \in \mathcal{I}_2$$

$$R_{n+1}^{\lambda \cdot s} = \sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left| \frac{y_j(t) - [\hat{\mu}^j(x_{n+1,j})](t)}{\lambda \cdot s_j(t)} \right| \right) = \frac{1}{\lambda} R_{n+1}^s.$$

Moreover, let us define:

$$\delta_{\boldsymbol{y}}^{\lambda \cdot s} := \frac{\left| \left\{ d \in \mathcal{I}_2 \cup \{n+1\} : R_d^{\lambda \cdot s} \geq R_{n+1}^{\lambda \cdot s} \right\} \right|}{l+1},$$

with, as usual, $\mathcal{C}_{n,1-\alpha}^{\lambda \cdot s}(x_{n+1}) := \left\{ \boldsymbol{y} \in \prod_{j=1}^{p} L^{\infty}(\mathcal{T}_j) : \delta_{\boldsymbol{y}}^{\lambda \cdot s} > \alpha \right\}$. As a consequence, $\boldsymbol{y} \in \mathcal{C}_{n,1-\alpha}^{\lambda \cdot s}(x_{n+1}) \iff R_{n+1}^{\lambda \cdot s} \leq k^{\lambda \cdot s}$, with $k^{\lambda \cdot s}$ the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d^{\lambda \cdot s} : d \in \mathcal{I}_2\}$. Since $R_d^{\lambda \cdot s} = R_d^s / \lambda \ \forall d \in \mathcal{I}_2$, then $k^{\lambda \cdot s} = k^s / \lambda$. Then:

$$\begin{aligned} & R_{n+1}^{\lambda \cdot s} \leq k^{\lambda \cdot s} \\ \iff & \frac{1}{\lambda} R_{n+1}^s \leq \frac{k^s}{\lambda} \\ \iff & R_{n+1}^s \leq k^s, \end{aligned}$$

and since $\boldsymbol{y} \in \mathcal{C}_{n,1-\alpha}^{s}(x_{n+1}) \iff R_{n+1}^s \leq k^s$, then $\mathcal{C}_{n,1-\alpha}^{\lambda \cdot s}(x_{n+1})$ coincides with $\mathcal{C}_{n,1-\alpha}^{s}(x_{n+1})$.

C.2 Appendix for Chapter 3.2.2

Proof of Theorem 3.1

Let us consider $\bar{s}_j(t)$, with $j \in \{1, \dots, p\}$. Since $m/n = \theta$ with $0 < \theta < 1$, if $n \to +\infty$ then $m \to +\infty$. The scalar γ is the empirical quantile of order $\lceil (m+1)(1-\alpha) \rceil$) of

$\{\sup_{j\in\{1,\dots,p\}}\left(\operatorname{ess\,sup}_{t\in\mathcal{T}_j}|y_{h,j}(t)-[\hat{\mu}^j(x_h)](t)|\right):h\in\mathcal{I}_1\}$. First of all note that

$$\lim_{m\to+\infty}\frac{\lceil(m+1)(1-\alpha)\rceil}{m}=\lim_{m\to+\infty}\frac{m+1-\lfloor(m+1)\alpha\rfloor}{m}$$

and since

$$\frac{(m+1)\alpha-1}{m}\leq\frac{\lfloor(m+1)\alpha\rfloor}{m}\leq\frac{(m+1)\alpha}{m}\quad\forall m\in\mathbb{N},$$

$$\lim_{m\to+\infty}\frac{(m+1)\alpha-1}{m}=\lim_{m\to+\infty}\frac{(m+1)\alpha}{m}=\alpha$$

then by the squeeze theorem we know that

$$\lim_{m\to+\infty}\frac{\lfloor(m+1)\alpha\rfloor}{m}=\alpha$$

and then

$$\lim_{m\to+\infty}\frac{\lceil(m+1)(1-\alpha)\rceil)}{m}=1-\alpha.$$

Consequently, γ is the empirical quantile of order $1-\alpha$ when $m\to+\infty$.

Let us define $w_h:=\sup_{j\in\{1,\dots,p\}}\left(\operatorname{ess\,sup}_{t\in\mathcal{T}_j}|y_{h,j}(t)-[\hat{\mu}^j(x_h)](t)|\right)\ \forall\ h\in\mathcal{I}_1$. The random variables $\{W_h:h\in\mathcal{I}_1\}$ from which $\{w_h:h\in\mathcal{I}_1\}$ are drawn are continuous and after $\operatorname{Var}\big[[\hat{\mu}^j(X_h)](t)\big]\to 0\ \forall j\in\{1,\dots,p\}$ they become i.i.d.. The Glivenko-Cantelli theorem guarantees that the empirical distribution function of these variables converges uniformly and almost surely pointwise to its distribution function, and so also the empirical quantiles converge in distribution - and so in probability - to the corresponding theoretical quantiles (see, for example, Van der Vaart, 2000, chap. 21). In so doing, empirical quantile γ converges to $q_{1-\alpha}$, the theoretical quantile of order $1-\alpha$. As a consequence:

$$\mathcal{H}_1:=\{h\in\mathcal{I}_1:\sup_{j\in\{1,\dots,p\}}\left(\operatorname*{ess\,sup}_{t\in\mathcal{T}_j}\left|y_{h,j}(t)-[\hat{\mu}^j(x_h)](t)\right|\right)\leq q_{1-\alpha}\}$$

when $m\to+\infty$, with $q_{1-\alpha}$ non-random quantity. Let us consider the numerator of $\bar{s}_j(t)$ $\forall j\in\{1,\dots,p\}$ as the denominator is a normalizing constant. $\forall t\in\mathcal{T}_j$, the sequence $\{\max_{h\in\mathcal{H}_1}|y_{h,j}(t)-[\hat{\mu}^j(x_h)](t)|\}_m$ is eventually bounded by $q_{1-\alpha}$ and is eventually increasing since $\{|\mathcal{H}_1|\}_m$ is eventually increasing. Therefore the sequence converges to its supremum by the monotone convergence theorem.

As regards $\bar{s}_j^c$, first of all it is possible to notice that if $n\to+\infty$ then $l=n(1-\theta)\to+\infty$. In order to show the convergence of the numerator of $\bar{s}_j^c$ to the same limit function,

it is sufficient to consider the previous calculations by substituting γ with k, m with l, $\mathcal{H}_1$ with $\mathcal{H}_2$ and $\mathcal{I}_1$ with $\mathcal{I}_2$. Finally, as the numerators of $\bar{s}_j$ and $\bar{s}_j^c$ converge to the same function $\forall j \in \{1, \ldots, p\}$, also the two normalizing constants converge to the same value.

Proof of Theorem 3.2

For the sake of simplicity, let us focus on the case in which $|\mathcal{H}_2| = \lceil (l+1)(1-\alpha) \rceil$. Under the assumption concerning the continuous joint distribution of $\{R_d : d \in \mathcal{I}_2\}$ made in Chapter 1 such condition is always satisfied, but for the sake of completeness the proof when this assumption is violated is addressed below.

- $\forall d \in \mathcal{H}_2$, $\forall j \in \{1, \ldots, p\}$ the following relationship holds $\forall t \in \mathcal{T}_j$:

$$\begin{aligned}
&\left| \frac{y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)}{\bar{s}_j^c(t)} \right| \\
&= \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| dt \cdot \frac{|y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|}{\max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|} \\
&\leq \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| dt.
\end{aligned}$$

  By indicating with $R_d^{\bar{s}^c}$ the nonconformity score induced by the set of functions $\bar{s}^c$, then

$$R_d^{\bar{s}^c} := \sup_{j \in \{1,\ldots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left| \frac{y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)}{\bar{s}_j^c(t)} \right| \right) \leq \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| dt.$$

  Specifically, $\exists\ \underline{d} \in \mathcal{H}_2$ such that $R_{\underline{d}}^{\bar{s}^c} = \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| dt$ since $\forall j \in \{1, \ldots, p\}$ and $\forall t \in \mathcal{T}_j$ at least one function $y_{\underline{d},j}$ satisfies $|y_{\underline{d},j}(t) - [\hat{\mu}^j(x_{\underline{d}})](t)| = \max_{d \in \mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_{d,j})](t)|$.

- Let us define $\mathcal{CH}_2 := \mathcal{I}_2 \setminus \mathcal{H}_2$ and let (t_d^*, j_d^*) be the couple of values such that

$$\left| y_{d,j_d^*}(t_d^*) - [\hat{\mu}^{j_d^*}(x_d)](t_d^*) \right| = \sup_{j \in \{1,\ldots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \quad \forall d \in \mathcal{I}_2.$$

  If (t_d^*, b_d^*) is not unique, it is randomly chosen from the couples satisfying that condition.

$\forall b \in \mathcal{CH}_2$, by definition of $\mathcal{H}_2$ it is possible to notice that $\left|y_{b,j_b^*}(t_b^*) - [\hat{\mu}^{j_b^*}(x_b)](t_b^*)\right| > \max_{d\in\mathcal{H}_2} |y_{d,j_b^*}(t_b^*) - [\hat{\mu}^{j_b^*}(x_d)](t_b^*)|$ and so the following relationship holds:

$$\begin{aligned}
&\left|\frac{y_{b,j_b^*}(t_b^*) - [\hat{\mu}^{j_b^*}(x_b)](t_b^*)}{\bar{s}^c_{j_b^*}(t_b^*)}\right| \\
&= \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d\in\mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|dt \cdot \frac{\left|y_{b,j_b^*}(t_b^*) - [\hat{\mu}^{j_b^*}(x_b)](t_b^*)\right|}{\max_{d\in\mathcal{H}_2} |y_{d,j_b^*}(t_b^*) - [\hat{\mu}^{j_b^*}(x_d)](t_b^*)|} \\
&> \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d\in\mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|dt.
\end{aligned}$$

Consequently,

$$R_b^{\bar{s}^c} := \sup_{j\in\{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t\in\mathcal{T}_j} \left| \frac{y_{b,j}(t) - [\hat{\mu}^j(x_b)](t)}{\bar{s}^c_j(t)} \right| \right) > \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d\in\mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|dt.$$

Since:

- $|\mathcal{H}_2| = \lceil (l+1)(1-\alpha) \rceil$
- $\forall d \in \mathcal{H}_2 \; R_d^{\bar{s}^c} \leq \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d\in\mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|dt$ and $\exists \; \underline{d} \in \mathcal{H}_2$ such that $R_{\underline{d}}^{\bar{s}^c} = \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d\in\mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|dt$
- $\forall b \in \mathcal{CH}_2 \; R_b^{\bar{s}^c} > \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d\in\mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|dt$

we conclude that $k^{\bar{s}^c} = \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d\in\mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|dt$, with $k^{\bar{s}^c}$ the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d^{\bar{s}^c} : d \in \mathcal{I}_2\}$.

If $|\mathcal{H}_2| > \lceil (l+1)(1-\alpha) \rceil$, then $R_d^{\bar{s}^c} = \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d\in\mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|dt$ is valid $\forall d \in \mathcal{H}_2$ such that $\sup_{j\in\{1,\dots,p\}} \left(\operatorname{ess\,sup}_{t\in\mathcal{T}_j} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)| \right) = k$ and we can conclude also in this case that $k^{\bar{s}^c} = \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d\in\mathcal{H}_2} |y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)|dt$.

Focusing now on the set of modulation functions s^0, $\forall d \in \mathcal{I}_2$:

$$R_d^{s^0} := \sup_{j\in\{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t\in\mathcal{T}_j} \left| \frac{y_{d,j}(t) - [\hat{\mu}^j(x_d)](t)}{s_j^0(t)} \right| \right) = \sup_{j\in\{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t\in\mathcal{T}_j} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \cdot \sum_{j=1}^{p} |\mathcal{T}_j| \,.$$

Since k^{s^0} is the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_d^{s^0} : d \in \mathcal{I}_2\}$, by definition of $\mathcal{H}_2$ we can notice that

$$\begin{aligned} k^{s^0} &= \max_{d \in \mathcal{H}_2} R_d^{s^0} \\ &= \max_{d \in \mathcal{H}_2} \left(\sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \right) \cdot \sum_{j=1}^{p} |\mathcal{T}_j| \\ &= \sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \right) \cdot \sum_{j=1}^{p} |\mathcal{T}_j| \,. \end{aligned}$$

Since by the integral mean value theorem we know that $\forall j \in \{1, \dots, p\}$

$$\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \cdot |\mathcal{T}_j| \geq \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| dt,$$

then the following relationship is valid:

$$\sum_{j=1}^{p} \operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \cdot |\mathcal{T}_j| \geq \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| dt. \tag{C.1}$$

In addition, by definition $\forall j \in \{1, \dots, p\}$

$$\sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \right) \geq \operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right)$$

and so:

$$\begin{aligned} &\sum_{j=1}^{p} \sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \right) \cdot |\mathcal{T}_j| \\ &= \sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \right) \cdot \sum_{j=1}^{p} |\mathcal{T}_j| \\ &\geq \sum_{j=1}^{p} \operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \cdot |\mathcal{T}_j| \,. \end{aligned} \tag{C.2}$$

By combining Inequality C.1 and Inequality C.2 we can notice that

$$\sup_{j \in \{1,\dots,p\}} \left(\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \right) \cdot \sum_{j=1}^{p} |\mathcal{T}_j| \geq \sum_{j=1}^{p} \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| dt,$$

i.e. $k^{s^0} \geq k^{\bar{s}^c}$. Then, $\mathcal{Q}(s^0) \geq \mathcal{Q}(\bar{s}^c)$.

Specifically, the integral mean value theorem guarantees that $\forall j \in \{1, \ldots, p\}$

$$\operatorname*{ess\,sup}_{t \in \mathcal{T}_j} \left(\max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \right) \cdot |\mathcal{T}_j| = \int_{\mathcal{T}_j} \max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| dt$$
$$\iff \max_{d \in \mathcal{H}_2} \left| y_{d,j}(t) - [\hat{\mu}^j(x_d)](t) \right| \quad \text{is constant almost everywhere,}$$

i.e. if and only if $\bar{s}^c_j(t)$ is constant almost everywhere over $\mathcal{T}_j$. Consequently, if at least one of the functions $\bar{s}^c_1(t), \ldots, \bar{s}^c_p(t)$ is not constant almost everywhere over its domain then the left side of Inequality C.1 is strictly greater than the right side (implying $\mathcal{Q}(s^0) > \mathcal{Q}(\bar{s}^c)$); otherwise, $\mathcal{Q}(s^0) = \mathcal{Q}(\bar{s}^c)$.

Generalization of $(\bar{s}, \bar{s}^c)$***, Theorem 3.1 and Theorem 3.2 to the Smoothed Split Conformal framework***

The functions $\bar{s}^c$ and $\bar{s}$ are defined as in the Split Conformal framework, except for: k (γ respectively) that is the $\lceil l + \tau_{n+1} - (l+1)\alpha \rceil$th ($\lceil m + \tau_{n+1} - (m+1)\alpha \rceil$th respectively) smallest value in the corresponding set; similarly to the Split Conformal framework, if $\lceil m + \tau_{n+1} - (m+1)\alpha \rceil > m$ then $\mathcal{H}_1 = \mathcal{I}_1$ and if $\lceil m + \tau_{n+1} - (m+1)\alpha \rceil \leq 0$ we arbitrarily set $\bar{s}_j = s^0_j$. Theorem 3.1 and Theorem 3.2 still hold by substituting $\lceil (l+1)(1-\alpha) \rceil, \lceil (m+1)(1-\alpha) \rceil$ with $\lceil l + \tau_{n+1} - (l+1)\alpha \rceil, \lceil m + \tau_{n+1} - (m+1)\alpha \rceil$.

Bibliography

Aneiros, G., Cao, R., Fraiman, R., Genest, C. and Vieu, P. (2019) Recent advances in functional data analysis and high-dimensional statistics. *J. Multivariate Anal.* **170**, 3–9.

Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J.-M. (2016) A prediction interval for a function-valued forecast model: Application to load forecasting. *Int. J. Forecast.* **32**(3), 939–947.

Balasubramanian, V., Ho, S.-S. and Vovk, V. (2014) *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Boston: Morgan Kaufmann.

Barber, R. F., Candes, E. J., Ramdas, A. and Tibshirani, R. J. (2021) Predictive inference with the jackknife+. *Ann. Statist.* **49**(1), 486–507.

Canale, A. and Vantini, S. (2016) Constrained functional time series: Applications to the Italian gas market. *Int. J. Forecast.* **32**(4), 1340–1351.

Cao, G., Yang, L. and Todem, D. (2012) Simultaneous Inference For The Mean Function Based on Dense Functional Data. *J. Nonparametr. Stat.* **24**(2), 359–377.

Chen, Y., Koch, T., Lim, K. G., Xu, X. and Zakiyeva, N. (2021) A review study of functional autoregressive models with application to energy forecasting. *Wiley Interdiscip. Rev. Comput. Stat.* **13**(3), e1525.

Chernozhukov, V., Fernandez-Val, I., Melly, B. and Wüthrich, K. (2019) Generic inference on quantile and quantile effect functions for discrete outcomes. *Journal of the American Statistical Association* .

Chernozhukov, V., Wüthrich, K. and Zhu, Y. (2018) Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory*, pp. 732–749.

Chernozhukov, V., Wüthrich, K. and Zhu, Y. (2021) Distributional conformal prediction. *Proceedings of the National Academy of Sciences* **118**(48).

Choi, H. and Reimherr, M. (2018) A geometric approach to confidence regions and bands for functional parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80**(1), 239–260.

Degras, D. (2017) Simultaneous confidence bands for the mean of functional data. *Wiley Interdiscip. Rev. Comput. Stat.* **9**(3), e1397.

Degras, D. A. (2011) Simultaneous confidence bands for nonparametric regression with functional data. *Statist. Sinica* **21**(4), 1735–1765.

Delaigle, A., Hall, P. *et al.* (2010) Defining probability density for a distribution of random functions. *Ann. Statist.* **38**(2), 1171–1193.

Ferraty, F., Goia, A. and Vieu, P. (2002) Functional nonparametric model for time series: a fractal approach for dimension reduction. *Test* **11**(2), 317–344.

Ferraty, F. and Vieu, P. (2004) Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *J. Nonparametr. Stat.* **16**(1-2), 111–125.

Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. New York: Springer-Verlag.

Ferraty, F. and Vieu, P. (2009) Additive prediction and boosting for functional data. *Comput. Statist. Data Anal.* **53**(4), 1400–1413.

Fontana, M., Zeni, G. and Vantini, S. (2022) Conformal prediction: a unified review of theory and new challenges. *Bernoulli* (forthcoming) .

Gammerman, A., Vovk, V. and Vapnik, V. (1998) Learning by Transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pp. 148–155. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. event-place: Madison, Wisconsin.

Gao, Y. and Shang, H. L. (2017) Multivariate Functional Time Series Forecasting: Application to Age-Specific Mortality Rates. *Risks* **5**(2), 21.

Gijbels, I. and Nagy, S. (2017) On a general definition of depth for functional data. *Statist. Sci.* **32**(4), 630–639.

Goia, A. and Vieu, P. (2016) An introduction to recent advances in high/infinite dimensional statistics. *J. Multivariate Anal.* **146**, 1–6.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hormann and Kidzinski (2017) *freqdom.fda: Functional Time Series: Dynamic Functional Principal Components*. R package version 0.9.1.

Horváth, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications*. Springer Series in Statistics. New York: Springer-Verlag.

Hyndman, R. J. and Shahid Ullah, M. (2007) Robust forecasting of mortality and fertility rates: A functional data approach. *Comput. Statist. Data Anal.* **51**(10), 4942–4956.

Hyndman, R. J. and Shang, H. L. (2009) Forecasting functional time series. *J. Korean Statist. Soc.* **38**(3), 199–211.

Hörmann, S. and Kokoszka, P. (2012) Functional Time Series. In *Handbook of Statistics*, volume 30 of *Time Series Analysis: Methods and Applications*, pp. 157–186. Elsevier.

Inselberg, A. (1985) The plane with parallel coordinates. *Vis Comput* **1**(2), 69–91.

Lehmann, E. L. and Romano, J. P. (2006) *Testing statistical hypotheses*. Springer Science & Business Media.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018) Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113**(523), 1094–1111.

Lei, J., Rinaldo, A. and Wasserman, L. (2015) A conformal prediction approach to explore functional data. *Ann. Math. Artif. Intell.* **74**(1-2), 29–43.

Lei, J., Robins, J. and Wasserman, L. (2013) Distribution-free prediction sets. *J. Amer. Statist. Assoc.* **108**(501), 278–287.

Liebl, D. and Reimherr, M. (2019) Fast and fair simultaneous confidence bands for functional parameters. *arXiv:1910.00131* .

Lopez-Pintado, S. and Qian, K. (2021) A depth-based global envelope test for comparing two groups of functions with applications to biomedical data. *Stat. Med.* **40**(7), 1639–1652.

López-Pintado, S. and Romo, J. (2009) On the concept of depth for functional data. *J. Amer. Statist. Assoc.* **104**(486), 718–734.

Mosler, K. (2013) Depth statistics. In *Robustness and complex data structures*, pp. 17–34. New York, NY: Springer.

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H. and Hahn, U. (2017) Global envelope tests for spatial processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79**(2), 381–404.

Nagy, S. (2016) *Statistical depth for functional data.* Ph.D. thesis, KU Leuven and Charles University, xxvi+216 pp.

Nagy, S., Gijbels, I. and Hlubinka, D. (2017) Depth-based recognition of shape outlying functions. *J. Comput. Graph. Statist.* **26**(4), 883–893.

Narisetty, N. N. and Nair, V. N. (2016) Extremal depth for functional data and applications. *J. Amer. Statist. Assoc.* **111**(516), 1705–1714.

Papadopoulos, H., Proedrou, K., Vovk, V. and Gammerman, A. (2002) Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356.

Paparoditis, E. and Shang, H. L. (2021) Bootstrap Prediction Bands for Functional Time Series. *arXiv:2004.03971* .

Pelagatti, M. (2013) Supply function prediction in electricity auctions. In *Complex Models and Computational Methods in Statistics*, pp. 203–213. Springer.

R Core Team (2020) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. (1982) When the data are functions. *Psychometrika* **47**(4), 379–396.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis.* Second edition edition. Springer series in statistics. New York, NY: Springer.

Ramsay, J. O., Wickham, H., Graves, S. and Hooker, G. (2020) *fda: Functional Data Analysis.* R package version 2.4.8.1.

Rao, P. (1971) Some notes on misspecification in multiple regressions. *Amer. Statist.* **25**(5), 37–39.

Romano, J. P. (1990) On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.* **85**(411), 686–692.

Romano, Y., Patterson, E. and Candes, E. (2019) Conformalized quantile regression. *Advances in neural information processing systems* **32**.

Rossini, J. and Canale, A. (2019) Quantifying prediction uncertainty for functional-and-scalar to functional autoregressive models under shape constraints. *J. Multivariate Anal.* **170**, 221–231.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applies statistician. *Ann. Statist* **12**(4), 1151–1172.

Shafer, G. and Vovk, V. (2008) A Tutorial on Conformal Prediction. *J. Mach. Learn. Res.* **9**, 371–421.

Shah, I. and Lisi, F. (2020) Forecasting of electricity price through a functional prediction of sale and purchase curves. *J. Forecast.* **39**(2), 242–259.

Sun, Y. and Genton, M. G. (2011) Functional Boxplots. *J. Comput. Graph. Statist.* **20**(2), 316–334.

Tarabelloni, N., Arribas-Gil, A., Ieva, F., Paganoni, A. M. and Romo, J. (2018) *roahd: Robust Analysis of High Dimensional Data*. R package version 1.4.1.

Telschow, F. J. and Schwartzman, A. (2022) Simultaneous confidence bands for functional data using the Gaussian kinematic formula. *J. Statist. Plann. Inference* **216**, 70–94.

Torti, A., Pini, A. and Vantini, S. (2021) Modelling time-varying mobility flows using function-on-function regression: Analysis of a bike sharing system in the city of milan. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70**(1), 226–247.

Tuddenham, R. D. and Snyder, M. M. (1954) Physical growth of california boys and girls from birth to eighteen years. *University of California publications in child development* **1**, 183–364.

Van der Vaart, A. W. (2000) *Asymptotic statistics*. Volume 3. Cambridge university press.

Vapnik, V. (1992) Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pp. 831–838.

Vovk, V., Gammerman, A. and Shafer, G. (2005) *Algorithmic learning in a random world*. Springer Science & Business Media.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016) Functional Data Analysis. *Annu. Rev. Stat. Appl.* **3**, 257–295.

Wooldridge, J. M. (1994) A simple specification test for the predictive ability of transformation models. *Rev. Econ. Stat.* **76**(1), 59–65.

Xu, C. and Xie, Y. (2020) Conformal prediction interval for dynamic time-series. *arXiv:2010.09107* .

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**(470), 577–590.

Zuo, Y. (2003) Projection-based depth functions and associated medians. *Ann. Statist.* **31**(5), 1460–1490.