MOX-Report No. 07/2021

The Importance of Being a Band: Finite-Sample Exact Distribution-Free Prediction Sets for Functional Data

Diquigiovanni, J.; Fontana, M.; Vantini, S.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

The Importance of Being a Band: Finite-Sample Exact Distribution-Free Prediction Sets for Functional Data

Jacopo Diquigiovanni[1,4], Matteo Fontana[2,3], Simone Vantini[2]

[1]Department of Statistical Sciences, University of Padova, Italy.

[2]MOX - Department of Mathematics, Politecnico di Milano, Italy

[3]now at European Commission, Joint Research Centre (JRC), Ispra (VA), Italy

[4]Corresponding Author: jacopo.diquigiovanni@phd.unipd.it

Abstract: Functional Data Analysis represents a field of growing interest in statistics. Despite several studies have been proposed leading to fundamental results, the problem of obtaining valid and efficient prediction sets has not been thoroughly covered. Indeed, the great majority of methods currently in the literature rely on strong distributional assumptions (e.g, Gaussianity), dimension reduction techniques and/or asymptotic arguments. We propose a new nonparametric approach in the field of Conformal Prediction, based on a new family of nonconformity measures inducing conformal predictors able to create closed-form finite-sample valid or exact prediction sets for functional data under very minimal distributional assumptions. In addition, our proposal ensures that the prediction

The present manuscript is a working paper, the final version has been currently submitted to a journal

sets obtained are bands, an essential feature in the functional setting that allows the visualization and interpretation of such sets. 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 bands. Within this family of measures, we propose also a specific measure leading to prediction bands asymptotically no less efficient than those with constant width.

Key words and phrases: Conformal Prediction, Distribution-free prediction, Exact prediction set, Functional data, Prediction band, Uncertainty quantification.

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 func-

tions defined over a domain, either uni- or multidimensional, and usually characterized by some degree of smoothness. 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 $\mathbb{R}^d$, $d \in \mathbb{N}_{>0}$, 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 $d = 1$ 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: The first one is composed of works based mainly on bootstrapping techniques, either parametric (e.g., Degras, 2011; Cao et al., 2012) or, via the use of functional quantiles, via nonparametric bootstrap techniques (e.g., Cuevas et al., 2006; Berg et al., 2017; Schüssler and Trede, 2016). The first two references are involved with construction of simultaneous confidence bands for the mean of functional data, but it should be noted that in the case of Gaussian functional data this problem and the issue of forecasting a new functional observation are essentially equivalent, and one can transform the simultaneous confidence bands for the mean into simultaneous prediction bands via a simple rescaling. The second class is represented by cases in which 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.

The framework of this manuscript is Conformal Prediction (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. (2023). In univariate setting, Conformal Prediction is able to generate distribution-free, valid prediction intervals and 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 this article, we build on top of the literature about set prediction for functional data and Conformal Prediction, by introducing several theoretical and methodological innovations.

1. After having introduced the importance in interpretative terms of obtaining functional prediction sets having a specific shape (i.e. prediction bands) in Section 2.2 functional prediction sets are formally defined and the Semi-Off-Line Inductive Conformal framework, also known simply as Split Conformal, 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) 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 S1.1).

2. In Section 2.3 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.

3. In Section 2.4 we propose a family of nonconformity measures (to which the nonconformity measure introduced in Section 2.3 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).

4. In Section 2.4 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,

Theorem 3).

Finally, in Section S2 in the Supplementary Materials we propose a simulation study to compare our method with four alternatives, and in Section 3 we apply our approach to the Berkeley Growth Study data set (Tuddenham and Snyder, 1954). Section 4 provides an overview of the main results.

2. Conformal Prediction Bands

2.1 The Importance of Being a Band

Set prediction is an issue of key importance in the statistical community. Specifically, three main features characterize a prediction set: shape, coverage, and size. We start by tackling, in this section, the first issue, while the last two are explored in Section 2. In the classical multivariate 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*. Formally, a band is defined as

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

with $B_n(t) \subseteq \mathbb{R}$ 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 $\{B_n(t) : t \in \mathcal{T}\}$, 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 $\mathcal{T}$.

It should also be noted that, 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 development of a method that necessarily outputs prediction bands - instead of more general prediction sets - represents the starting point of this work.

2.2 Conformal Prediction

The framework we use to develop our prediction sets is *Conformal Prediction*, a nonparametric approach proposed in the multivariate literature for the first time by Gammerman et al. (1998) and thoroughly described in Vovk et al. (2005), that can be used to construct finite-sample either valid or exact prediction sets under no assumptions other than *i.i.d.* data (for a review of the topic see, e.g., Lei et al., 2018; Fontana et al., 2023). Even

though the theory holds also under the weaker assumption of exchangeable data, in this manuscript we will focus on the case of *i.i.d.* data which is a very common case in applications and in particular on the case of *i.i.d.* functional data taking value in $\mathcal{Y}(\mathcal{T})$.

Following the notation of Vovk et al. (2005), given a set of *i.i.d.* random functions $Y_1, \ldots, Y_n \sim P$ and an independent random function $Y_{n+1} \sim P$, a valid prediction set $\mathcal{C}_{n,1-\alpha} := \mathcal{C}_{n,1-\alpha}(Y_1, \ldots, Y_n)$ for Y_{n+1} is a set such that

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}\right) \geq 1 - \alpha \tag{2.1}$$

for any significance level $\alpha \in (0, 1)$ and with $\mathbb{P}$ the probability corresponding to the product measure induced by P (Lei et al., 2015). If the inequality in (2.1) is replaced by the equality, the prediction set is also said to be exact. In order to avoid ambiguity, later in the discussion the term *coverage* (or *unconditional coverage*) will be used to refer to $\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}\right)$, the term *conditional coverage* will be used to refer to $\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha} | \mathcal{C}_{n,1-\alpha}\right)$ 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.

Specifically, we will focus on the Semi-Off-Line Inductive Conformal framework, also known simply as *Split Conformal*, a computationally effi-

cient modification of the original Transductive Conformal method (firstly proposed in Papadopoulos et al., 2002). In order to present this approach, let us consider the following procedure: given data $y_1, \ldots, y_n$, let $\{1, \ldots, n\}$ be randomly divided into two sets $\mathcal{I}_1, \mathcal{I}_2$ and let us define the training set as $\{y_h : h \in \mathcal{I}_1\}$ and the calibration set as $\{y_h : h \in \mathcal{I}_2\}$, 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 *nonconformity measure* as any measurable function $A(\{y_h : h \in \mathcal{I}_1\}, y)$ taking values in $\bar{\mathbb{R}}$ whose aim is to score how different $y \in \mathcal{Y}(\mathcal{T})$ is from the training set. The split conformal prediction set constructed on the basis of the observed sample $y_1, \ldots, y_n$ is defined as $\mathcal{C}_{n,1-\alpha} := \{y \in \mathcal{Y}(\mathcal{T}) : \delta_y > \alpha\}$, with

$$\delta_y := \frac{|\{j \in \mathcal{I}_2 \cup \{n+1\} : R_j \geq R_{n+1}\}|}{l+1}$$

and *nonconformity scores* $R_j := A(\{y_h : h \in \mathcal{I}_1\}, y_j)$ for $j \in \mathcal{I}_2$, $R_{n+1} := A(\{y_h : h \in \mathcal{I}_1\}, y)$. In particular, hereafter we will focus on nonconformity scores $\{R_h : h \in \mathcal{I}_2\}$ having a continuous joint distribution, an assumption generally satisfied in the functional context.

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 (a direct consequence of having i.i.d. data) δ_Y is uniformly distributed over $\{1/(l+1), 2/(l+1), \ldots, 1\}$ and then (2.1) holds. Theorem 1 proves and enriches

such known result by making the exact probability reached by split prediction sets explicit. The proof is given in Appendix S1.1.

Theorem 1. *Let $\mathcal{C}_{n,1-\alpha}$ be a split conformal prediction set. If $Y_1, \ldots, Y_{n+1}$ are i.i.d. and $\{R_h : h \in \mathcal{I}_2\}$ have a continuous joint distribution, then*

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

Specifically, $\mathcal{C}_{n,1-\alpha}$ always satisfies

$$1 - \alpha \leq \mathbb{P}\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}\right) < 1 - \alpha + \frac{1}{l+1}. \tag{2.2}$$

A natural consequence of the first part of Theorem 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 states that the 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 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 (2.2) 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 (2.2).

Conformal inference is a field of deep interest as minimal assumptions are required on P to obtain prediction sets satisfying (2.1) for any finite sample size n, a property particularly appealing in the functional context. A slight modification (Vovk et al., 2005) of the aforementioned procedure even allows to obtain a stronger version of Theorem 1: in order to present it, first of all let us introduce an element of randomization τ_{n+1}, realization of a uniform random variable in $[0, 1]$. The smoothed split conformal prediction set is defined as $\mathcal{C}_{n,1-\alpha,\tau_{n+1}} := \left\{y \in \mathcal{Y}(\mathcal{T}) : \delta_{y,\tau_{n+1}} > \alpha\right\}$, with

$$(l+1) \cdot \delta_{y,\tau_{n+1}} := \left|\{j \in \mathcal{I}_2 : R_j > R_{n+1}\}\right| + \tau_{n+1} \left|\{j \in \mathcal{I}_2 \cup \{n+1\} : R_j = R_{n+1}\}\right|.$$

Smoothed split conformal prediction sets are, by construction, exact for any α, l, i.e. $\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha,\tau_{n+1}}\right) = 1 - \alpha$: 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 S1.1.

Remark 1. Our discussion was limited to the split setting because our work only focuses on it, but the results of this section are very general and require just little changes to be applied to the Transductive/Full Conformal

framework. In addition, as highlighted by Vovk et al. (2005) and briefly mentioned at the beginning of this section, Theorem 1 and the result about exactness of smoothed prediction sets hold even when the weaker assumption of exchangeability is formulated instead of the traditional hypothesis of i.i.d. data.

Remark 2. The division of data into the training and calibration sets always induces an element of randomness into the procedure, also in the non-smoothed 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 article the aforementioned single-split process is considered. The computation of the effect of splitting - as well as the impact of the specific values m and l - on the procedure has not yet been properly analyzed in the Conformal Prediction literature (see e.g. Fontana et al., 2023), but it is a topic worth of further research.

Remark 3. The Conformal approach can be also successfully applied to regression and classification problems. A detailed presentation is not included hereafter being out of scope, but an exhaustive discussion can be found in Vovk et al. (2005).

Remark 4. Although we focus on the functional setting, the Conformal

framework has initially been developed in the traditional univariate and multivariate settings and so all arguments and results presented in this section can also be applied to univariate variables and random vectors.

2.3 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 and of the discussion in Section 2.1, we propose a fast and scalable split 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, 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 al-

lows to obtain prediction bands and is based on the essential supremum:

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

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 2.4 (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 article 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 non-smoothed scenario (the minor changes needed for the smoothed case are introduced in Appendix S1.4), 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 $\{R_h : h \in \mathcal{I}_2\}$. 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 (2.3) is

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

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 (2.4)

are simultaneous by construction, i.e. bands ensuring the desired coverage globally (in addition to the pointwise validity). Similarly to the multivariate 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 (2.4) (the proof is given in Appendix S1.2); 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.

Remark 5. In application scenarios where data are characterized by specific features (e.g., positivity, monotonicity etc...), the approach presented in this Section allows to remove portions of the observed prediction bands that violate such known characteristics, without affecting the coverage. An example of this band trimming procedure is given in Section 3. 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.

2.4 Improving Efficiency: the Choice of the Modulation Function

It can be easily noted that the width of (2.4) 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. For this reason it is of key importance to create prediction bands whose width can be adapted to the local variability of functional data.

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 1 shows the prediction band obtained by the procedure presented in Section 2.3 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

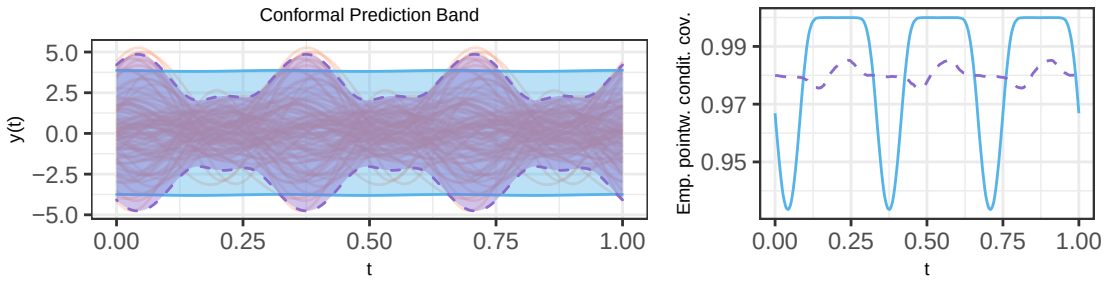

Figure 1: The left panel shows the split conformal prediction band computed as in (2.4) (solid light blue band) and that computed as in (2.6) by considering the standard deviation function as $s_{\mathcal{I}_1}$ (dashed purple band). For visualization, a random subsample of $y_1, \ldots, y_{198}$ is 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). $\alpha = 0.1$.

nonconformity measure and nonconformity scores:

$$A(\{y_h : h \in \mathcal{I}_1\}, y) = \operatorname*{ess\,sup}_{t\in\mathcal{T}} \left| \frac{y(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right|, \tag{2.5}$$

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

with $j \in \mathcal{I}_2$ 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 (2.5) can be suitably considered as the nonconformity measure (2.3) 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 (2.5), obtained by replicating the computations of Section 2.3 (see Appendix S1.3 for the proof), is

$$\mathcal{C}^s_{n,1-\alpha} := \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in \tag{2.6}$$
$$[g_{\mathcal{I}_1}(t) - k^s s_{\mathcal{I}_1}(t), g_{\mathcal{I}_1}(t) + k^s s_{\mathcal{I}_1}(t)] \forall t \in \mathcal{T}\},$$

with k^s the $\lceil (l+1)(1-\alpha) \rceil$th smallest value in the set $\{R_h^s : h \in \mathcal{I}_2\}$. In other words, the procedure presented in this section consists of modulating the data, computing the prediction band (2.4) 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 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 one in the top panel 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 6. 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 $\{R_h^s : h \in \mathcal{I}_2 \cup \{n+1\}\}$ not being exchangeable, and consequently validity not being guaranteed.

Remark 7. 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 S1.3. 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 (2.3) 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 8. 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 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 the 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:

$$\mathcal{Q}(s_{\mathcal{I}_1}) := \int_{\mathcal{T}} 2 \cdot k^s \cdot s_{\mathcal{I}_1}(t)dt = 2 \cdot k^s, \tag{2.7}$$

that is equal to k^s up to a constant and proportional to $2k^s/|\mathcal{T}|$, i.e. the average width of the prediction band over the domain $\mathcal{T}$.

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 $R^s_h, h \in \mathcal{I}_2$. In

fact, the dependency of $s_{\mathcal{I}_1}$ only on the functions of the training set and of the numerator of R_h^s (i.e. $|y_h(t) - g_{\mathcal{I}_1}(t)|\,, \quad h \in \mathcal{I}_2$) 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 $\{y_h : h \in \mathcal{I}_2\}$ 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 $\{y_h : h \in \mathcal{I}_2\}$ 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 (2.7)) 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 (2.3).

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 conforming function between those in the calibration set, in which the concept of "conformity" is induced by the metric used, being deemed as "conforming" a function whose distance from g_{I_1} is particularly high. 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:

$$\bar{s}^c_{\mathcal{I}_1}(t) := \frac{\max_{j \in \mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)|}{\int_{\mathcal{T}} \max_{j \in \mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)| dt} \tag{2.8}$$

with

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

and k defined as in Section 2.3, i.e. the $\lceil (l+1)(1-\alpha)\rceil$th smallest value in the set $\{R_h : h \in \mathcal{I}_2\}$. The corresponding t-function is

$$\bar{s}_{\mathcal{I}_1}(t) := \frac{\max_{j\in\mathcal{H}_1} |y_j(t) - g_{\mathcal{I}_1}(t)|}{\int_{\mathcal{T}} \max_{j\in\mathcal{H}_1} |y_j(t) - g_{\mathcal{I}_1}(t)| dt} \tag{2.9}$$

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

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

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 S1.3 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 (2.8) is proved by the following two theorems.

Theorem 2. *Let* $m/n = \theta$ *with* $0 < \theta < 1$ *and let* $\text{Var}[g_{\mathcal{I}_1}(t)] \to 0$ *when* $m \to +\infty$. *Then* $\bar{s}^c_{\mathcal{I}_1}$ *and* $\bar{s}_{\mathcal{I}_1}$ *converge to the same function when* $n \to +\infty$ *and* $\lim_{n\to+\infty} \mathcal{C}^{\bar{s}}_{n,1-\alpha} = \lim_{n\to+\infty} \mathcal{C}^{\bar{s}^c}_{n,1-\alpha}\ \forall\ \alpha \in (0,1)$.

Theorem 3. $\mathcal{Q}(s^0) \geq \mathcal{Q}(\bar{s}^c_{\mathcal{I}_1})$. *Specifically,* $\mathcal{Q}(s^0) = \mathcal{Q}(\bar{s}^c_{\mathcal{I}_1})$ *if and only if* $\max_{j\in\mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)|$ *is constant almost everywhere.*

Both proofs are given in Appendix S1.3. It is important to notice that Theorem 2 requires very mild conditions, an evidence that allows it to hold in many general contexts.

In light of this, the function (2.9) represents an outstanding candidate in the choice of the modulation function since the Conformal setting and the nonconformity measure (2.5) guarantee valid prediction bands - as well as all the other desirable properties highlighted in Section 2.3 - and at the same time to asymptotically obtain prediction bands no less efficient than those induced by s^0.

Remark 9. 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 theoreti-

cal justification for preferring function (2.9) to other possible modulation functions.

Theorem 4. *Let us define* $\mathcal{CH}_2 := \mathcal{I}_2 \setminus \mathcal{H}_2$ *and let* t_i^* *be the value such that*

$$|y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)| = \operatorname*{ess\,sup}_{t \in \mathcal{T}} |y_i(t) - g_{\mathcal{I}_1}(t)| \quad \forall i \in \mathcal{I}_2. \tag{2.10}$$

If t_i^* *is not unique, it is randomly chosen from the values that satisfy* (2.10).

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

1. $s^d_{\mathcal{I}_1} \neq \bar{s}^c_{\mathcal{I}_1}$ *in the sense of Lebesgue, i.e.* $\exists\ \mathcal{T}^* \subseteq \mathcal{T}$ *such that* $s^d_{\mathcal{I}_1}(t) \neq \bar{s}^c_{\mathcal{I}_1}(t)\ \forall t \in \mathcal{T}^*$ *and* $\mu(\mathcal{T}^*) > 0$, *with* μ *the Lebesgue measure*
2. $s^d_{\mathcal{I}_1}(t_i^*) \leq \bar{s}^c_{\mathcal{I}_1}(t_i^*)\ \forall i \in \mathcal{CH}_2$

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

The proof is given in Appendix S1.3, along with the demonstration that Theorem 3 is not a direct consequence of Theorem 4 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 4 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 $\{R_h : h \in \mathcal{I}_2\}$ made in Section 2.2.

Remark 10. The definitions of functions (2.8), (2.9) and Theorems 2, 3 and 4 can be easily generalized to hold also in the Smoothed Conformal framework. Technical details are provided in Appendix S1.4.

Remark 11. As it is the case with the mean function g_{I_1} that is chosen according to the specific applicative problem at hand, the choice of the modulation function s_{I_1} has to be performed in a similar fashion. Indeed it has to be selected on a case-by-case basis by taking into consideration the specific charachteristics of the modelling task at hand (e.g., homooscedasticity/heteroscedasticity along the domain or presence/absence of outliers).

3. 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 the 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 Non-Smoothed Conformal framework, the three modulation functions analyzed in Section S2 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. Note that since the application at hand 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, 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 complex scenario with strong heteroskedasticity 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

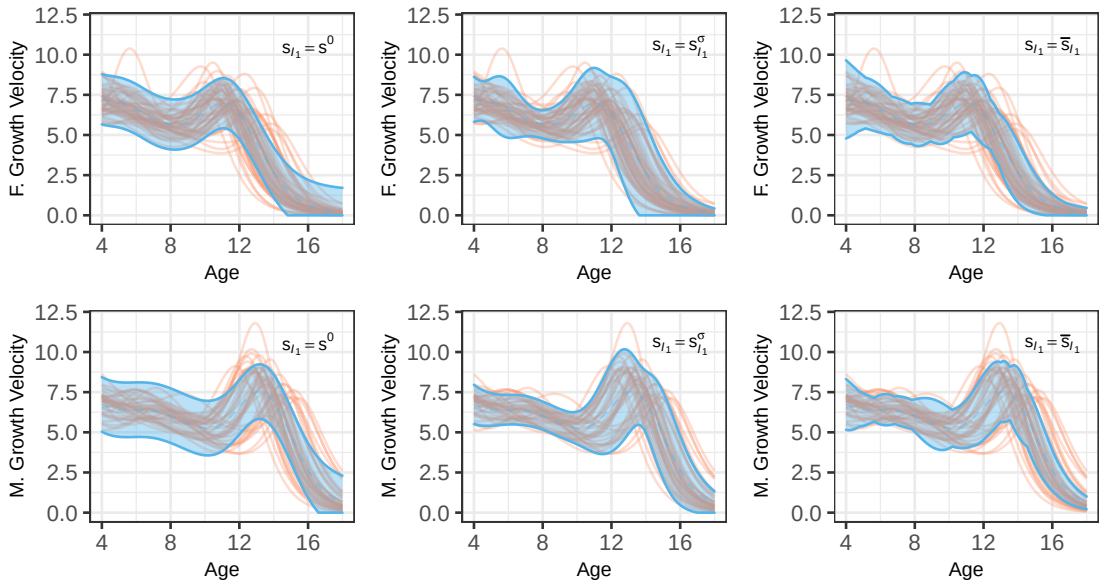

Figure 2: Berkeley Growth Study data: each panel shows the prediction band obtained considering a different modulation function (s^0 on the left, the normalised pointwise standard deviation function $s^{\sigma}_{\mathcal{I}_1}$ in the middle, $\bar{s}_{\mathcal{I}_1}$ on the right). In all cases, the dashed line represents $g_{\mathcal{I}_1}$. Predictions for girls at the top and predictions for boys at the bottom.

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 slightly narrower due to the presence of outliers. Instead focusing on boys' curves, the strong heteroskedasticity forces the prediction band induced by s^0 to be uselessly large in some parts

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

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 idiosyncratic and caused by the specific modulation function used which is not pointwise related to an average but on the specific value assumed by one of the functions. This creates narrower but less smooth bands. Both for boys and girls $\bar{s}_{\mathcal{I}_1}$ outputs the smallest prediction band, as shown in Table 1 where the quantity $\mathcal{Q}(\cdot)/|\mathcal{T}|$ is reported.

Additionally, we have explored the role that the pointwise predictor covers with respect to the prediction perfomance in this applied case. The explored methods are, similarly to the simulation study, a baseline method, represented by the sample mean (stylised "*Mean*", accompanied by the functional median (stylised "*Median*"), where the point predictor is represented by the deepest curve of the sample, according to MBD. and by

	Mean	Median	Trimmed mean
Females	2.811	3.614	2.910
Males	2.690	4.362	3.266

Table 2: Berkeley Growth Study data: average width of the prediction bands for different point predictors, using $\bar{s}_{\mathcal{I}_1}$ as the modulation function

a trimmed functional mean, computed excluding the 10% of the shallowest curves in the sample, again according to MBD. The the modulation function selected is $\bar{s}_{\mathcal{I}_1}$.

Some useful information can be also provided by the comparison between the proposed approach and its pointwise counterpart, in which the prediction band is constructed by applying a coherent univariate Conformal approach at each point t separately. Indeed, by construction the former creates prediction bands larger or equal than those obtained by the latter, but on the other hand it guarantees simultaneous (and not pointwise) validity and of course it interprets a function as a whole, a key aspect in the functional context. In order to clarify this concept, let us consider Figure 3, in which the pointwise prediction band (dark blue) is overlaid to the bottom-right panel of Figure 2. As expected, the pointwise prediction band is simply

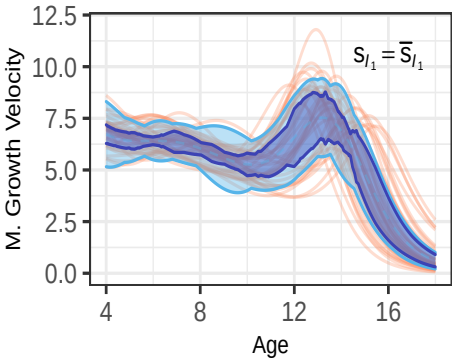

Figure 3: Berkeley Growth Study data: the prediction band represented at the bottom right of Figure 2 (light blue) and the correponding pointwise conformal prediction band (dark blue).

modulated by the local variability of the 50% central curves. Differently, the prediction bands here proposed instead take also into consideration the behavior of the functions along the domain $\mathcal{T}$ with the effect of generating narrower or wider bands also in presence of similar local variabilities and so not just obtaining a simple expansion of the pointwise prediction band.

4. Conclusion

The creation of prediction sets for 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 article 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. The focus of this article was on i.i.d. data, but we envision an extension of the procedure to regression and classification problems.

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 article 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.

Supplementary Materials

The supporting material contains the technical proofs and the supplementary material for the simulation study.

Acknowledgements

The authors would like to thank Giulio Pegorer for fruitful discussions.

Funding

Prof. Vantini and Dr. Fontana acknowledge the financial support from Accordo Quadro ASI-POLIMI "Attività di Ricerca e Innovazione" n. 2018-5-HH.0, collaboration agreement between the Italian Space Agency and

Politecnico di Milano.

References

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

Berg, J., E. Oikarinen, M. Järvisalo, and K. Puolamäki (2017). Minimum-width confidence bands via constraint optimization. In J. C. Beck (Ed.), *Principles and Practice of Constraint Programming*, Cham, pp. 443–459. Springer International Publishing.

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

Cuevas, A., M. Febrero, and R. Fraiman (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics Data Analysis 51*(2), 1063–1074.

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

Fontana, M., G. Zeni, and S. Vantini (2023). Conformal prediction: A unified review of theory

and new challenges. *Bernoulli 29*(1), 1 – 23.

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

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

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

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

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

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

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

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

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis* (Second edition ed.).

Springer series in statistics. New York, NY: Springer. OCLC: 249216329.

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

Schüssler, R. and M. Trede (2016). Constructing minimum-width confidence bands. *Economics Letters 145*, 182–185.

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

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

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

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

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

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

The Importance of Being a Band: Finite-Sample Exact Distribution-Free Prediction Sets for Functional Data

Jacopo Diquigiovanni[1,4], Matteo Fontana[2,3], Simone Vantini[2]

[1]Department of Statistical Sciences, University of Padova, Italy.

[2]MOX - Department of Mathematics, Politecnico di Milano, Italy

[3]now at European Commission, Joint Research Centre (JRC), Ispra (VA), Italy

[4]Corresponding Author: jacopo.diquigiovanni@phd.unipd.it

S1 Technical Proofs

S1.1 Proofs of Section 2.2

Proof of Theorem 1.

Since $\mathcal{C}_{n,1-\alpha} := \{y \in \mathcal{Y}(\mathcal{T}) : \delta_y > \alpha\}$, then $\mathcal{C}_{n,1-\alpha} := \{y \in \mathcal{Y}(\mathcal{T}) : (l+1)\delta_y > (l+1)\alpha\}$. Under the hypothesis of the theorem, $(l+1)\delta_Y \sim U\{1, 2, \dots, l+1\}$ holds. As a consequence:

$$\begin{aligned}\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}\right) &= \mathbb{P}\left((l+1)\delta_Y > (l+1)\alpha\right) \\ &= 1 - \mathbb{P}\left((l+1)\delta_Y \leq (l+1)\alpha\right) \\ &= 1 - \frac{\lfloor (l+1)\alpha \rfloor}{l+1}.\end{aligned}$$

The present manuscript is a working paper, the final version has been currently submitted to a journal

In addition, since

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

then $\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}\right) \geq 1-\alpha$, i.e. $\mathcal{C}_{n,1-\alpha}$ is valid. Finally, since

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

then $\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}\right) < 1-\alpha+\frac{1}{l+1}$.

Proof that smoothed split conformal prediction sets are exact.

Let us consider the hypothesis of Theorem 1. Let us notice that

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

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

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

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

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

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

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

Then

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

S1.2 Proofs of Section 2.3

Proof that the concatenation of pointwise prediction intervals leads to a prediction band that is a subset of the simultaneous prediction band (2.4).

Let $\mathcal{U}_{n,1-\alpha}$ be the pointwise prediction set. Let us define $\tilde{R}_j(t) := |y_j(t) - g_{\mathcal{I}_1}(t)| \ \forall t \in \mathcal{T}, j \in \mathcal{I}_2$, $\tilde{R}_{n+1}(t) := |y(t) - g_{\mathcal{I}_1}(t)|$ 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 $\{\tilde{R}_h(t) : h \in \mathcal{I}_2\}$. By construction $R_j = \operatorname{ess\,sup}_{t \in \mathcal{T}} \tilde{R}_j(t)$, and so $R_j \geq \tilde{R}_j(t) \ \forall t \in \mathcal{T}, j \in \mathcal{I}_2$ 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 $\mathcal{U}_{n,1-\alpha} \subseteq \mathcal{C}_{n,1-\alpha}$.

S1.3 Proofs of Section 2.4

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

For a given $y \in \mathcal{Y}(\mathcal{T})$, let us define

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

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 $\{R_h^s : h \in \mathcal{I}_2\}$. 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

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

Proof of Remark 7.

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

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

Let us also define

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

The split conformal prediction set is defined as $\mathcal{C}^{\lambda\cdot s}_{n,1-\alpha} := \left\{y \in \mathcal{Y}(\mathcal{T}) : \delta^{\lambda\cdot s}_y > \alpha\right\}$. 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 $\{R^{\lambda\cdot s}_h : h \in \mathcal{I}_2\}$. In addition, since $R^{\lambda\cdot s}_j = R^s_j/\lambda \ \forall j \in \mathcal{I}_2$, 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 $\max_{j\in\mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)| = 0$ for at least one value t but the condition $\int_{\mathcal{T}} \max_{j\in\mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)| dt \neq 0$ 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 $\int_{\mathcal{T}} \max_{j\in\mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)|dt = 0$ is addressed only when $y_j(t) = g_{\mathcal{I}_1}(t)$ $\forall j \in \mathcal{H}_2$ and almost every $t \in \mathcal{T}$ and it represents a case of no practical interest.

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

Proof of Theorem 2.

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 $x_i := \operatorname{ess\,sup}_{t\in\mathcal{T}} |y_i(t) - g_{\mathcal{I}_1}(t)|$ $\forall$ $i \in \mathcal{I}_1$. 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 $\text{Var}[g_{\mathcal{I}_1}(t)] \to 0$. 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 $m \to +\infty$:

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

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 $\{\max_{j\in\mathcal{H}_1} |y_j(t) - g_{\mathcal{I}_1}(t)|\}_m$ 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. In view of this and since $\mathcal{C}^{\bar{s}}_{n,1-\alpha}$ and $\mathcal{C}^{\bar{s}^c}_{n,1-\alpha}$ are defined as

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

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

then $\lim_{n\to+\infty} \mathcal{C}^{\bar{s}}_{n,1-\alpha} = \lim_{n\to+\infty} \mathcal{C}^{\bar{s}^c}_{n,1-\alpha}$.

Proof of Theorem 3.

The proof consists of two steps. At the first step we show that $k^{\bar{s}^c} = \int_{\mathcal{T}} \max_{j\in\mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)|dt$, 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 $\{R_h : h \in \mathcal{I}_2\}$ made in Section 2.2 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 $\forall t \in \mathcal{T}$:

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

  and then

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

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

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

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

  If t_i^* 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 $|y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)| > \max_{j \in \mathcal{H}_2} |y_j(t_i^*) - g_{\mathcal{I}_1}(t_i^*)|$

and so the following relationship holds:

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

As a consequence,

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

Since:

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

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

If $|\mathcal{H}_2|>\lceil(l+1)(1-\alpha)\rceil$, then $R_i^{\bar{s}^c}=\int_{\mathcal{T}}\max_{j\in\mathcal{H}_2}|y_j(t)-g_{\mathcal{I}_1}(t)|dt$ 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 $k^{\bar{s}^c}=\int_{\mathcal{T}}\max_{j\in\mathcal{H}_2}|y_j(t)-g_{\mathcal{I}_1}(t)|dt$.

II step

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

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

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

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

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

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

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

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

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

Proof of Theorem 4.

We have already shown at the first step of the previous proof that $k^{\bar{s}^c} = \int_{\mathcal{T}} \max_{j \in \mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)| dt$. Since by assumption $s^d_{\mathcal{I}_1}(t^*_i) \leq \bar{s}^c_{\mathcal{I}_1}(t^*_i)\ \forall i \in \mathcal{CH}_2$ 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 $s^d_{\mathcal{I}_1}(t^*_i) = \bar{s}^c_{\mathcal{I}_1}(t^*_i) - a_i$.

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

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

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

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 $s^d_{\mathcal{I}_1}(t_\downarrow) < \bar{s}^c_{\mathcal{I}_1}(t_\downarrow)$ and $s^d_{\mathcal{I}_1}(t_\uparrow) > \bar{s}^c_{\mathcal{I}_1}(t_\uparrow)$ since otherwise $s^d_{\mathcal{I}_1}(t) = \bar{s}^c_{\mathcal{I}_1}(t)\ \forall t \in \mathcal{T}^*$. Let us define $a_\downarrow > 0$ the value such that $s^d_{\mathcal{I}_1}(t_\downarrow) = \bar{s}^c_{\mathcal{I}_1}(t_\downarrow) - a_\downarrow$. Therefore $\exists\ \underline{i} \in \mathcal{H}_2$ such that

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

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

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

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

- $\forall i \in \mathcal{CH}_2$, by definition of $\mathcal{H}_2$ we obtain that $|y_i(t_i^*) - g_{\mathcal{I}_1}(t_i^*)| > \max_{j \in \mathcal{H}_2} |y_j(t_i^*) - g_{\mathcal{I}_1}(t_i^*)|$

  and so the following relationship holds:

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

  As a consequence,

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

Since:

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

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

Proof that Theorem 4 does not imply Theorem 3.

Theorem 4 does not imply Theorem 3 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, $\forall i \in \mathcal{CH}_2$:

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

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

S1.4 Proofs about Smoothed Conformal Predictor

Proof of the smoothed conformal prediction set

By considering the notation of Section 2, first of all let us notice that, by definition, $\mathcal{C}_{n,1-\alpha,1} = \mathcal{C}_{n,1-\alpha}$.

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 $\{R_h : h \in \mathcal{I}_2\}$, and r_n (v_n respectively) the number of elements in the set $\{R_h : h \in \mathcal{I}_2\}$ 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 $\{R_h : h \in \mathcal{I}_2\}$ made in Section 2.2 $r_n = v_n = 0$ 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 non-randomized 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

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

Also the introduction of the modulation function presented in Section 2.4 can be easily generalized in the smoothed conformal context. Let us define for a given $y \in \mathcal{Y}(\mathcal{T})$

$$\delta^s_{y,\tau_{n+1}} := \frac{\left|\left\{j \in \mathcal{I}_2 : R^s_j > R^s_{n+1}\right\}\right| + \tau_{n+1}\left|\left\{j \in \mathcal{I}_2 \cup \{n+1\} : R^s_j = R^s_{n+1}\right\}\right|}{l+1}$$

$$\mathcal{C}^s_{n,1-\alpha,\tau_{n+1}} := \left\{y \in \mathcal{Y}(\mathcal{T}) : \delta^s_{y,\tau_{n+1}} > \alpha\right\}.$$

By reconsidering the previous computations and by substituting $\delta_{y,\tau_{n+1}}$ with $\delta^s_{y,\tau_{n+1}}$, w with w^s, R_h with R^s_h, 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 10.

The functions $\bar{s}^c_{\mathcal{I}_1}$ and $\bar{s}_{\mathcal{I}_1}$ are defined as in Section 2.4 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 2.4 still hold by substituting $\lceil (l+1)(1-\alpha) \rceil, \lceil (m+1)(1-\alpha) \rceil$ with $\lceil l + \tau_{n+1} - (l+1)\alpha \rceil, \lceil m + \tau_{n+1} - (m+1)\alpha \rceil$.

S2 Simulation Study

S2.1 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, Extremal depth and Bootstrap. In Section S2.2 the empirical coverage is evaluated for each approach in three different scenarios, whereas in Section S2.3 the prediction bands obtained by the methods that guarantee a proper coverage are compared in terms of efficiency. The simulation study has been mainly performed in the R programming language using the `conformalInference.fd` package (Diquigiovanni et al. 2022). The code to reproduce the simulations and the analyses of the test case is available upon request to the authors. 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). The data generating processes of the three scenarios are:

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

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

with $\mathcal{T} = [0, 1]$, $(x_{11}, x_{12}, x_{13})^T, \dots, (x_{n1}, x_{n2}, x_{n3})^T$ 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 $u_1, \dots, u_n$ i.i.d. realizations of

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

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

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

with $\mathcal{T} = [0, 1]$, $B_j^{\omega}(t)$ the b-spline basis system of order 4 with interior knots $\omega = (0.1, 0.2, \dots, 0.9)$ and $(c_{1,1}, \dots, c_{1,13})^T, \dots, (c_{n,1}, \dots, c_{n,13})^T$ i.i.d. realizations of $C = (C_1, \dots, C_{13}) \sim N_{13}(\mathbf{0}, \Sigma)$ 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 $i, j = 1, \dots, 13$, $i \neq j$.

- Scenario 3. The scenario is the previous one after contamination with outliers. Formally, $(c_{1,1}, \dots, c_{1,13})^T, \dots, (c_{n,1}, \dots, c_{n,13})^T$ 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.

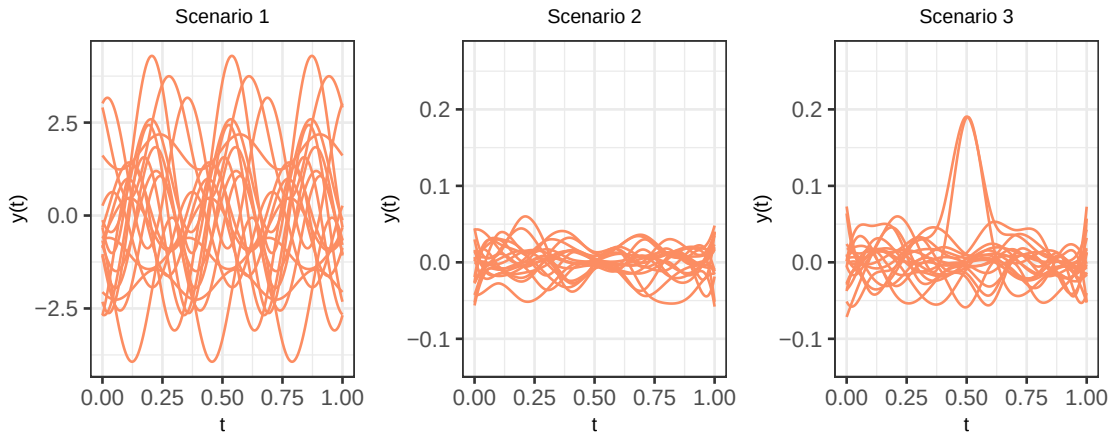

Figure S2.1: Graphical representation of the scenarios. The sample size is $n = 18$.

A graphical representation of a replication for each scenario with $n = 18$ is provided in Figure S2.1. The Conformal approach presented in Section 2 is evaluated in the non-smoothed framework and considering three different modulation functions: s^0, the normalised pointwise 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 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 & Genton 2011); *Extremal* which output the sample $(1-\alpha)$ central region induced by the

extremal depth (Narisetty & Nair 2016); *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). The simulations are achieved by using the R Programming Language (R Core Team 2018) 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.

S2.2 Coverage

In this section we focus on 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 (see Table S2.1). 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 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

		Conformal Method			**Alternative Methods**				
		s^0	$s^{\sigma}_{\mathcal{I}_1}$	$\bar{s}_{\mathcal{I}_1}$	Naive	MBD	BD	Ext. Depth	Boot.
$n = 18$	Scenario 1	0.902	0.900	0.900	0.409	0.504	0.547	0.498	0.875
		(0.088)	(0.085)	(0.087)	(0.092)	(0.109)	(0.111)	(0.107)	(0.064)
	Scenario 2	0.901	0.910	0.909	0.048	0.123	0.145	0.119	0.922
		(0.089)	(0.081)	(0.083)	(0.021)	(0.044)	(0.051)	(0.042)	(0.042)
	Scenario 3	0.904	0.904	0.907	0.049	0.124	0.148	0.122	0.932
		(0.084)	(0.089)	(0.085)	(0.023)	(0.049)	(0.055)	(0.048)	(0.061)
$n = 198$	Scenario 1	0.901	0.902	0.901	0.625	0.861	0.900	0.826	0.865
		(0.029)	(0.030)	(0.031)	(0.031)	(0.028)	(0.028)	(0.028)	(0.019)
	Scenario 2	0.901	0.899	0.900	0.189	0.733	0.788	0.678	0.897
		(0.029)	(0.031)	(0.029)	(0.019)	(0.036)	(0.032)	(0.033)	(0.015)
	Scenario 3	0.897	0.900	0.899	0.197	0.742	0.798	0.688	0.892
		(0.031)	(0.030)	(0.031)	(0.020)	(0.034)	(0.030)	(0.033)	(0.020)
$n = 1998$	Scenario 1	0.900	0.899	0.900	0.666	0.942	0.918	0.885	0.866
		(0.010)	(0.010)	(0.010)	(0.011)	(0.006)	(0.008)	(0.008)	(0.008)
	Scenario 2	0.900	0.900	0.899	0.233	0.958	0.971	0.858	0.899
		(0.009)	(0.010)	(0.010)	(0.007)	(0.006)	(0.005)	(0.008)	(0.008)
	Scenario 3	0.900	0.899	0.900	0.240	0.959	0.973	0.859	0.884
		(0.010)	(0.010)	(0.010)	(0.008)	(0.006)	(0.005)	(0.008)	(0.007)

Table S2.1: 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.

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.

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 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 S2.3 only the efficiency of the Conformal methods is evaluated and compared.

S2.3 Efficiency

In this section the sample mean and the standard deviation of the size defined as in (2.7) of the prediction bands computed in the previous section are evaluated for each combination of modulation function, sample size and scenario (see Table S2.2). 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

		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 S2.2: Size of the prediction bands.

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 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 S2.2 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

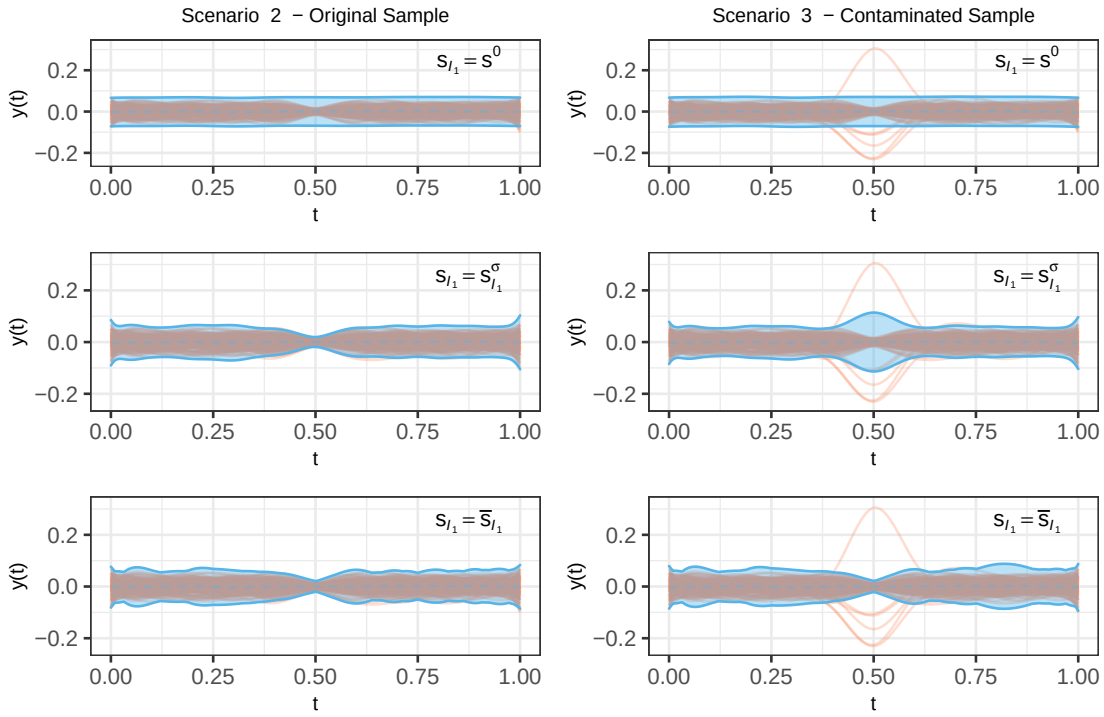

Figure S2.2: The prediction bands obtained considering a combination of modulation functions (s^0 at the top, $s^\sigma_{\mathcal{I}_1}$ in the middle, $\bar{s}_{\mathcal{I}_1}$ at the bottom) and sample (the original one on the left, the contaminated one on the right). In all cases, the dashed line represents $g_{\mathcal{I}_1}$.

us consider a sample $y_1, \dots, y_{198}$ generated as in Scenario 2 that, after being created, is exposed to a contamination process in which each function $y_i, i = 1, \dots, 198$, becomes an outlier as described in Scenario 3 with probability $\beta = 0.06$. Figure S2.2 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 (2.3)) 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 S2.2 - 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.

As an additional step, and to further evaluate the robustness of the proposed prediction method with respect to the use of different point forecasting methods, we inspect the sample mean and the standard deviation of the size, defined as in (2.7), of the prediction bands computed using different point predictors. Namely, we propose a table similar to S2.2, with the same declination in different sample sizes and different scenarios, but we explicitly explore different

point prediction methods (S2.3. The explored methods are a baseline case, represented by the sample mean (stylised "*Mean*", already used as a candidate in all the previous simulations. The baseline is accompanied by two less standard cases, represented by a functional median case (stylised "*Median*"), where the point predictor is represented by the deepest curve of the sample, according to MBD. The third case is instead represented by a trimmed mean, computed excluding the 10% of the shallowest curves in the sample, again according to MBD. In all the simulations the modulation function selected is $\bar{s}_{\mathcal{I}_1}$.

In our specific simulation scenario, the use of more complex methods does not seem to be justified by a statistically significant increase in prediction performance, nevertheless deeper explorations of this important and relatively overlooked topic are in order.

		Mean		Median		Trimmed Mean 90%	
		$Mean$	$st.dev$	$Mean$	$st.dev$	$Mean$	$st.dev$
$n = 18$	Scenario 1	11.749	(4.458)	11.849	(4.269)	11.749	(4.458)
	Scenario 2	0.183	(0.046)	0.195	(0.050)	0.183	(0.046)
	Scenario 3	0.491	(0.605)	0.506	(0.604)	0.491	(0.605)
$n = 198$	Scenario 1	7.509	(0.648)	7.582	(0.663)	7.522	(0.660)
	Scenario 2	0.120	(0.006)	0.130	(0.007)	0.120	(0.006)
	Scenario 3	0.138	(0.023)	0.149	(0.023)	0.138	(0.026)
$n = 1998$	Scenario 1	7.134	(0.186)	7.144	(0.188)	7.133	(0.188)
	Scenario 2	0.117	(0.002)	0.128	(0.006)	0.117	(0.002)
	Scenario 3	0.131	(0.003)	0.143	(0.007)	0.131	(0.003)

Table S2.3: Size of the prediction bands for different point predictors, modulated using $\bar{s}_{\mathcal{I}_1}$

Bibliography

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

Diquigiovanni, J., Fontana, M., Solari, A., Vantini, S. & Vergottini, P. (2022), *conformalInference.fd: Tools for Conformal Inference for Regression in Multivariate Functional Setting*. R package version 1.1.1.

URL: *https://CRAN.R-project.org/package=conformalInference.fd*

Narisetty, N. N. & Nair, V. N. (2016), 'Extremal depth for functional data and applications', *Journal of the American Statistical Association* **111**(516), 1705–1714.

URL: *https://doi.org/10.1080/01621459.2015.1110033*

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

URL: *https://www.R-project.org/*

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

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

URL: *https://CRAN.R-project.org/package=roahd*

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

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

06/2021 Negrini, G.; Parolini, N.; Verani, M.
A diffuse interface box method for elliptic problems

04/2021 Orlando, V.; Rea, F.; Savaré, L.; Guarino, I; Mucherino, S.; Perrella, A.; Trama, U.; Coscioni,
Development and validation of a clinical risk score to predict the risk of SARS-CoV-2 infection from administrative data: A population-based cohort study from Italy

05/2021 Antonietti, P.F.; Mazzieri, I.; Migliorini, F.
A discontinuous Galerkin time integration scheme for second order differential equations with applications to seismic wave propagation problems

02/2021 Calissano, A.; Feragen, A; Vantini, S.
Graph-Valued Regression: Prediction of unlabelled networks in a Non-Euclidean Graph-Space

03/2021 Torti, A.; Marika, A.; Azzone, G.; Secchi, P.; Vantini S.
Bridge closure in the road network of Lombardy: a spatio-temporal analysis of the socio-economic impacts

01/2021 Pegoraro, M.; Beraha, M.
Projected Statistical Methods for Distributional Data on the Real Line with the Wasserstein Metric

Franco, N.r.; Massi, M.c.; Ieva, F.; Manzoni, A.; Paganoni, A.m.; Zunino, P.; Et al.
Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity

85/2020 Cavinato, L.; Sollini, M.; Kirienko, M.; Biroli, M.; Ricci, F.; Calderoni, L.; Tabacchi, E.; Nann
PET radiomics-based lesions representation in Hodgkin lymphoma patients

84/2020 Vergara, C.; Stella, S.; Maines, M.; Catanzariti, D.; Demattè, C.; Centonze, M.; Nobile, F.; Qua
Computational electrophysiology to support the mapping of coronary sinus branches for cardiac resynchronization therapy

83/2020 Hron, K.; Machalova, J.; Menafoglio, A.
Bivariate densities in Bayes spaces: orthogonal decomposition and spline representation