Computational Statistics and Data Analysis 187 (2023) 107821

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Conformal prediction bands for two-dimensional functional time series

Niccolò Ajroldi [a], Jacopo Diquigiovanni [b], Matteo Fontana [c,*], Simone Vantini [a]

[a] MOX - Department of Mathematics, Politecnico di Milano, Milan (MI), Italy
[b] Department of Statistical Sciences, University of Padova, Padova (PD), Italy
[c] European Commission, Joint Research Centre (JRC), Ispra (VA), Italy

ARTICLE INFO

Article history:
Received 28 July 2022
Received in revised form 7 July 2023
Accepted 7 July 2023
Available online 17 July 2023

Keywords:
Conformal prediction
Forecasting
Functional autoregressive process
Functional time series
Two-dimensional functional data
Uncertainty quantification

ABSTRACT

Time evolving surfaces can be modeled as two-dimensional Functional time series, exploiting the tools of Functional data analysis. Leveraging this approach, a forecasting framework for such complex data is developed. The main focus revolves around Conformal Prediction, a versatile nonparametric paradigm used to quantify uncertainty in prediction problems. Building upon recent variations of Conformal Prediction for Functional time series, a probabilistic forecasting scheme for two-dimensional functional time series is presented, while providing an extension of Functional Autoregressive Processes of order one to this setting. Estimation techniques for the latter process are introduced, and their performance are compared in terms of the resulting prediction regions. Finally, the proposed forecasting procedure and the uncertainty quantification technique are applied to a real dataset, collecting daily observations of Sea Level Anomalies of the Black Sea.

1. Introduction

Data observed on a two-dimensional domain arise naturally across several disciplines, motivating an increasing demand for dedicated analysis techniques. Functional data analysis (FDA) (Ramsay and Silverman 2005) is naturally apt to represent and model this kind of data, as it allows preserving their continuous nature, and provides a rigorous mathematical framework. Among the others, Zhou and Pan 2014 analyzed temperature surfaces, presenting two approaches for Functional Principal Component Analysis (FPCA) of functions defined on a non-rectangular domain, Porro-Muñoz et al. 2014 focuses on image processing using FDA, while a novel regularization technique for Gaussian random fields on a rectangular domain has been proposed by Rakêt 2010 and applied to 2D electrophoresis images. Another bivariate smoothing approach in a penalized regression framework has been introduced by Ivanescu 2013, allowing for the estimation of functional parameters of two-dimensional functional data. As shown by Gervini 2010, even mortality rates can be interpreted as two-dimensional functional data.

Whereas in all the reviewed works functions are assumed to be realization of *iid* or at least *exchangeable* random objects, to the best of our knowledge there is no literature focusing on forecasting time-dependent two-dimensional functional data. In this work, we focus on time series of surfaces, representing them as two-dimensional Functional Time Series (FTS).

A two-dimensional Functional Time Series is an ordered sequence $Y_1, \dots, Y_T$ of random variables with values in a functional Hilbert space $\mathbb{H}$, characterized by some sort of temporal dependency. More formally, we consider a probability

* Corresponding author.
E-mail address: matteo.fontana@ec.europa.eu (M. Fontana).

https://doi.org/10.1016/j.csda.2023.107821

space $(\Omega, \mathcal{F}, \mathbb{P})$, and define a random function at time t as $Y_t : \Omega \to \mathbb{H}$, measurable with respect to the Borel σ-algebra $\mathcal{B}(\mathbb{H})$. In the rest of the article we consider functions belonging to $\mathbb{H} = \mathcal{L}^2([c,d] \times [e,f])$, with $c, d, e, f \in \mathbb{R}$, $c < d$, $e < f$ We stress the fact that, from a theoretical point of view, our methodology can be applied to functions defined on a generic subset of $\mathbb{R}^2$, however, for simplicity and without loss of generality, we will only consider rectangular domains.

Given a realization of the stochastic process $\{Y_t\}_{t=\dots,N}$, we aim to forecast the next surface and quantify the uncertainty around the predicted function. Whereas uncertainty quantification in the context of *univariate* FTS forecasting has received great attention in the statistical community in recent decades, no attempts have been made to extend them to functions defined on a bidimensional domain.

Most of the research tackling univariate FTS forecasting has focused on adaption of the Bootstrap to the functional setting (see e.g. Hyndman and Shang 2009, Rossini and Canale 2018 and Hernández et al. 2021). However, Bootstrap is a very computationally intensive procedure, especially in the infinite-dimensional context of functional data. In this work, we instead focus on Conformal Prediction (CP), a versatile nonparametric approach to prediction. The first appearance of such technique dates back to Gammerman et al. 1998, and it has been later presented in greater detail in Vovk et al. 2022 and Balasubramanian et al. 2006. An extensive and unified review of the theory of Conformal inference can be found in Fontana et al. 2023. The attractiveness of Conformal Prediction relies on its great versatility, which allows to couple it with any predictive algorithm, in order to obtain distribution free prediction sets. Throughout this work, we resort to Inductive Conformal Prediction, also known as Split Conformal Prediction (Papadopoulos et al. 2002). Such modification of the original Transductive Conformal method is not only computationally efficient, but also necessary in high-dimensional frameworks like the functional data one. It should be noted that the main drawback of the Full Conformal approach is the need of retraining the prediction algorithm for every possible candidate realization y. In practice, in multivariate problems, where y lies in $\mathbb{R}^p$, one runs the above routine for several candidates y over a p-dimensional regular grid. While such approach is prohibitive for high-dimensional spaces, since computational times grow exponentially with p, it becomes unfeasible in a functional setting, in which y lies in an infinite-dimensional space. Employing Split Conformal inference along with nonconformity scores tailored to functional data (Diquigiovanni et al. 2022) allows to obtain prediction sets in closed form.

Whereas CP theory has been originally developed under the assumption of exchangeability, Chernozhukov et al. 2018 reframed the CP framework in the context of randomization inference, proving approximate validity of the resulting prediction sets under weak assumptions on the conformity scores and on the ergodicity of the time series. Later, Diquigiovanni et al. 2021 adapted such methodology to allow for Functional Time Series in a Split Conformal setting. We extend such method to two-dimensional functional data.

Since we want to quantify prediction uncertainty, we also need to provide a *forecasting technique* for 2D Functional Time Series. We start by taking into account the literature on unidimensional FTS forecasting (Bosq 2000, Antoniadis et al. 2006, Horváth and Kokoszka 2012, Aue et al. 2012, Jiao et al. 2021), extending the theory of Functional Autoregressive Processes (FAR) to the bivariate setting, proposing estimation techniques for the FAR(1), and comparing them in an order to assess how forecasting performances influence the amplitude of prediction bands.

The rest of this paper is as follows: we illustrate Conformal Prediction for two-dimensional functional data in Section 2, providing theoretical guarantees of the resulting prediction bands. In Section 3, we introduce forecasting algorithms for two-dimensional Functional Time Series, proposing an extension of the FAR(1) for two-dimensional functional data. Such forecasting algorithms are then compared by means of the resulting prediction bands in a simulation study in Section 4. Finally, in Section 5 we employ the developed techniques to obtain forecasts and prediction bands of real data, predicting day by day the Black Sea level. Section 6 concludes.

2. Uncertainty quantification: conformal prediction for 2D functional time series

Consider a time series $Z_1, \dots, Z_T$ of regression pairs $Z_t = (X_t, Y_t)$, with $t = 1, \dots, T$. Let Y_t be a random variable with values in $\mathbb{H}$, while X_t is a set of random covariates at time t belonging to a measurable space. Notice that X_t is a generic set of regressors, which may contain both exogenous and endogenous variables. Later in the manuscript, we will consider X_t to contain only the lagged version of the function Y_t, namely Y_{t-1}. Given a significance level $\alpha \in [0,1]$, we aim to design a procedure that outputs a prediction set $C_{T,1-\alpha}(X_{T+1})$ for Y_{T+1} based on $Z_1, \dots, Z_T$ and X_{T+1}, with unconditional coverage probability close to $1-\alpha$. More formally, we define $C_{T,1-\alpha}(X_{T+1})$ to be a *valid* prediction set if:

$$\mathbb{P}(Y_{T+1} \in C_{T,1-\alpha}(X_{T+1})) \geq 1 - \alpha. \tag{1}$$

We would like to construct a specific type of prediction sets, commonly known as *prediction bands*, formally defined as:

$$\{y \in \mathbb{H} : y(u,v) \in B_n(u,v) \quad \forall (u,v) \in [c,d] \times [e,f]\}, \tag{2}$$

where $B_n(u,v) \subseteq \mathbb{R}$ is an interval for each $(u,v) \in [c,d] \times [e,f]$. The convenience of such type of prediction sets in applications is extensively motivated in the literature (see e.g. López-Pintado and Romo 2009, Lei et al. 2015 and Diquigiovanni et al. 2022), since a prediction set of this kind can be visualized easily, a property that is instead not guaranteed if the prediction region is a generic subset of $\mathbb{H}$.

Let $z_1, \dots, z_T$ be realizations of $Z_1, \dots, Z_T$. As the name suggests, Split Conformal inference is based on a random split of data into two disjoint sets: let I_1, I_2 be a random partition of $\{1, \dots, T\}$, such that $|I_1| = m$, $|I_2| = l$, $m, l \in \mathbb{N}$ $m, l > 0$,

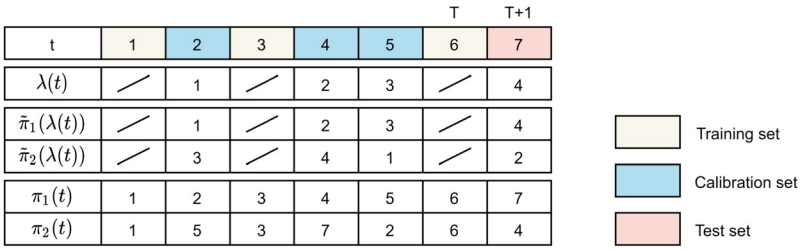


Fig. 1. Example of permutation families $\tilde{\Pi}$ and Π, with sample size $T = 6$, training set $I_1 = \{1, 3, 6\}$, calibration set $I_2 = \{2, 4, 5\}$, $l = m = 3$, size of blocking scheme $b = 2$. In this case $\lambda : \{I_2, T+1\} \equiv \{2, 4, 5, 7\} \to \{1, 2, 3, 4\}$, $\tilde{\Pi} = \{\tilde{\pi}_1, \tilde{\pi}_2\}$ and $\Pi = \{\pi_1, \pi_2\}$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

$m + l = T$. Historical observations $z_1, \dots, z_T$ are divided into a *training set* $\{z_h, h \in I_1\}$, used for model estimation, and a *calibration set* $\{z_h, h \in I_2\}$, used in an out-of-sample context to measure the nonconformity of a new candidate function. The choice of the split ratio and the type of split is non-trivial and has motivated discussion in the statistical community. We fix the training-calibration ratio equal to 1 and perform a random split, and refer to Appendix A for a more extensive discussion on the topic.

We then introduce a *nonconformity measure* $\mathcal{A}$, which is a measurable function with values in $\mathbb{R} \cup \{+\infty\}$. The role of $\mathcal{A}(\{z_h, h \in I_1\}, z)$ is to quantify the nonconformity of a new datum z with respect to the training set $\{z_h, h \in I_1\}$. The choice of the nonconformity measure is crucial to find prediction bands (2) in closed form. We employ the following nonconformity score, introduced by Diquigiovanni et al. 2022, extended here to two-dimensional functional data:

$$\mathcal{A}(\{z_h : h \in I_1\}, z) = \operatorname*{ess\,sup}_{(u,v)\in[c,d]\times[e,f]} \frac{|y(u,v) - g_{I_1}(u,v;x_{T+1})|}{s_{I_1}(u,v)}, \tag{3}$$

where $z = (x_{T+1}, y)$, g_{I_1} is a point predictor built from the training set I_1, and s_{I_1} is a *modulation function*, which is a positive function depending on I_1 that allows for prediction bands with non-constant width along the domain. Section 4 and Appendix B discuss the estimation of g_{I_1}. The functional standard deviation is employed as modulation function s_{I_1}, allowing for wider bands in the parts of the domain where data show high variability and narrower and more informative prediction bands in those parts characterized by low variability. For an extensive discussion on the optimal choice of modulation function, we refer to Diquigiovanni et al. 2022.

Consider now a candidate function $y \in \mathbb{H}$ and define the augmented dataset as $Z_{(y)} = \{Z_t\}_{t=1}^{T+1}$, where:

$$Z_t = \begin{cases} (X_t, Y_t), & \text{if } 1 \le t \le T, \\ (X_{T+1}, y), & \text{if } t = T+1. \end{cases} \tag{4}$$

The key idea of the methodology proposed by Chernozhukov et al. 2018 and extended by Diquigiovanni et al. 2021 is to generate several randomized versions of $Z_{(y)}$ through a specifically tailored permutation scheme, and compute nonconformity scores on each of them. We then decide whether to include y in the prediction region, by comparing the nonconformity score of $Z_{(y)}$ with that of its permuted replicas.

In order to obtain such replicas, we aim to define a family Π of index permutations $\pi : \{1, \dots, T+1\} \to \{1, \dots, T+1\}$, that keeps unchanged the training set indices I_1, and modifies only $\{I_2, T+1\}$, namely the indices of the calibration set and the next time step. We first introduce a function $\lambda : \{I_2, T+1\} \to \{1, \dots, l+1\}$ such that $\lambda(t)$ returns the t-th element of the ordered set $\{I_2, T+1\}$. Fix now a positive integer $b \in \{1, \dots l+1\}$ such that $\frac{l+1}{b} \in \mathbb{N}$ and define a family $\tilde{\Pi}$ of index permutations that acts on the set $\{1, \dots, l+1\}$. Each $\tilde{\pi}_i \in \tilde{\Pi}$ is required to be a bijection $\tilde{\pi}_i : \{1, \dots, l+1\} \to \{1, \dots, l+1\}$, for $i = 1, \dots, \frac{l+1}{b}$. We consider the non-overlapping blocking permutation scheme proposed by Chernozhukov et al. 2018, dividing data in blocks of size b, in such a way that each permutation is unique in $\tilde{\Pi}$:

$$\tilde{\pi}_i(j) = \begin{cases} j + (i-1)b & \text{if } 1 \le j \le l - (i-1)b + 1, \\ j + (i-1)b - l - 1 & \text{if } l - (i-1)b + 2 \le j \le l + 1. \end{cases} \tag{5}$$

By definition, we have that $|\tilde{\Pi}| = \frac{l+1}{b}$, and $\tilde{\Pi}$ forms an algebraic group, containing the identity transformation $\tilde{\pi}_1$. It is then straightforward to introduce the family Π of index permutations acting on $\{1, \dots, T+1\}$. Each $\pi_i \in \Pi$, with $i = 1, \dots, \frac{l+1}{b}$ is defined as:

$$\pi_i(t) = \begin{cases} t & \text{if } t \in I_1, \\ \lambda^{-1}(\tilde{\pi}_i(\lambda(t))) & \text{if } t \in I_2 \cup \{T+1\}. \end{cases} \tag{6}$$

Fig. 1 reports an example of the families of permutation Π and $\tilde{\Pi}$. We refer to $Z^{\pi}_{(y)} = \{Z_{\pi(t)}\}_{t=1}^{T+1}$ as the randomized version of $Z_{(y)} = \{Z_t\}_{t=1}^{T+1}$, and define the randomization p-value as:

$$p(y) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbf{1}(S(Z^{\pi}_{(y)}) \geq S(Z_{(y)})), \tag{7}$$

where nonconformity scores $S(Z_{(y)})$ and $S(Z^{\pi}_{(y)})$ are defined as:

$$S(Z_{(y)}) = \mathcal{A}(\{Z_h : h \in I_1\}, Z_{T+1}), \tag{8}$$

$$S(Z^{\pi}_{(y)}) = \mathcal{A}(\{Z_h : h \in I_1\}, Z_{\pi(T+1)}). \tag{9}$$

The idea is to apply permutations, modifying the order of observations in the calibration set, while at the same time preserving the dependence between them, thanks to the block structure of Π. For each π, we compute the nonconformity score of $Z^{\pi}_{(y)}$. The p-value (7) of a test candidate value y is then determined as the proportion of randomized versions $Z^{\pi}_{(y)}$ with a higher or equal nonconformity score than the one of the original augmented dataset $Z_{(y)}$. Notice that $p(y)$ is a measure of the *conformity* of the candidate function y with respect to the permutation family Π. It is then natural to include in the prediction set only functions y with an "high" conformity level. Given a significance level $\alpha \in [b/(l+1), 1]$, the prediction region is hence obtained by test inversion:

$$C_{T,1-\alpha}(X_{T+1}) := \{y \in \mathbb{H} : p(y) > \alpha\}. \tag{10}$$

As a remark, it should be noted that if $\alpha \in (0, b/(l+1))$ the resulting prediction set coincides with the entire space $\mathbb{H}$ (Diquigiovanni et al. 2022).

The advantage of using the Split Conformal method along with the conformity measure (3) relies on the possibility to find the prediction set in closed form. By defining k^s as the $\lceil (|\Pi|+1)(1-\alpha) \rceil$th smallest value of $\{S(Z^{\pi}_{(y)}), \pi \in \Pi \setminus \pi_1\}$, we derive:

$$\begin{aligned}
&y \in C_{T,1-\alpha}(X_{T+1}) \iff p(y) > \alpha \iff S(Z_{(y)}) \leq k^s \\
&\iff \operatorname*{ess\,sup}_{(u,v) \in [c,d] \times [e,f]} \frac{|y(u,v) - g_{I_1}(u,v;x_{T+1})|}{s_{I_1}(u,v)} \leq k^s \\
&\iff |y(u,v) - g_{I_1}(u,v;x_{T+1})| \leq k^s s_{I_1}(u,v) \qquad \forall (u,v) \in [c,d] \times [e,f] \\
&\iff y(u,v) \in \left[g_{I_1}(u,v;x_{T+1}) \pm k^s s_{I_1}(u,v) \right] \qquad \forall (u,v) \in [c,d] \times [e,f].
\end{aligned}$$

The prediction band is therefore:

$$C_{T,1-\alpha}(X_{T+1}) := \left\{ y \in \mathbb{H} : y(u,v) \in \left[g_{I_1}(u,v;x_{T+1}) \pm k^s s_{I_1}(u,v) \right] \forall (u,v) \in [c,d] \times [e,f] \right\}. \tag{11}$$

If regression pairs are exchangeable, the proposed method retains exact, model-free validity (Chernozhukov et al. 2018, Theorem 1). When such assumption is not met, one can instead guarantee approximate validity of the proposed approach under weak assumptions on the nonconformity score and the ergodicity of the time series. This result is illustrated in great detail by Theorem 2 of Chernozhukov et al. 2018 and Theorem 1 of Diquigiovanni et al. 2021. We report here the latter, with a slightly modified notation.

Let $Z = Z_{(Y_{T+1})}$, where the candidate function y is now substituted by the random function Y_{T+1}. Let $\mathcal{A}^*$ be an oracle nonconformity measure, inducing oracle nonconformity score S^*. Define F to be the cumulative (unconditional) distribution function of the oracle nonconformity scores, namely $F(x) = \mathbb{P}(S^*(Z^{\pi}_{(y)}) < x)$ and $\hat{F}$ the empirical counterpart, obtained by applying permutations $\pi \in \Pi$: $\hat{F}(x) := \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbb{1}\{S^*(Z^{\pi}_{(y)}) < x\}$. Let $\{\delta_{1\bar{l}}, \delta_{2\bar{m}}, \gamma_{1\bar{l}}, \gamma_{2\bar{m}}\}$ be sequences of numbers converging to zero.

Theorem 1. *If the following conditions hold:*

- $\sup_{a \in \mathbb{R}} |\hat{F}(a) - F(a)| \leq \delta_{1\bar{l}}$ *with probability* $1 - \gamma_{1\bar{l}}$
- $\frac{1}{|\Pi|} \sum_{\pi \in \Pi} \left[S(Z^{\pi}) - S^*(Z^{\pi}_{(y)}) \right]^2 \leq \delta^2_{2\bar{m}}$ *with probability* $1 - \gamma_{2\bar{m}}$
- $|S(Z^{\pi}) - S^*(Z^{\pi})| \leq \delta_{2\bar{m}}$ *with probability* $1 - \gamma_{2\bar{m}}$
- *With probability* $1 - \gamma_{2\bar{m}}$ *the pdf of* $S^*(Z^{\pi})$ *is bounded above by a constant* D

then the Conformal confidence set has approximate coverage α*:*

$$|\mathbb{P}(Y_{T+1} \in C_{T,1-\alpha}(X_{T+1})) - (1-\alpha)| \leq 6\delta_{1\bar{l}} + 2\delta_{2\bar{m}} + 2D\left(\delta_{2\bar{m}} + 2\sqrt{\delta_{2\bar{m}}}\right) + \gamma_{1\bar{l}} + \gamma_{2\bar{m}}. \tag{12}$$

The first condition concerns the approximate ergodicity of $\hat{F}$, a condition which holds for strongly mixing time series using blocking permutation Π defined in (6) (Chernozhukov et al. 2018). The other conditions are requirements for the quality of approximation of $S^*(Z^{\pi})$ with $S(Z^{\pi})$. Intuitively, $\delta^2_{2\bar{m}}$ bounds the discrepancy between the nonconformity scores and their oracle counterparts. Such condition is related to the quality of the point prediction and to the choice of the employed nonconformity measure.

3. Point prediction: functional autoregressive process of order one

In order to obtain CP band with empirical coverage close to the nominal one, the choice of an accurate point predictor is important. As mentioned before, whereas in the typical i.i.d. case finite-sample unconditional coverage still holds when the model is heavily misspecified (Diquigiovanni et al. 2022), in the time series context a strong model misspecification may compromise the coverage guarantees and not only the efficiency of the resulting prediction bands (Chernozhukov et al. 2018, Diquigiovanni et al. 2021). For this reason, it is important to consider models that are consistent with the functional nature of the observations and that can adequately deal with their infinite dimensionality. We build on top of the literature on functional autoregressive processes in Hilbert spaces, extending them for the first time to temporarily evolving surfaces. We narrow the forecasting methodology to the FAR(p), with $p = 1$ because of its wide success in the literature (Hernández et al. 2021, Papadopoulos et al. 2002 and Aue et al. 2012). Whereas in the scalar context it is often beneficial to consider lags p greater than one, given the intrinsic high dimensionality of functional data, we would rather fit a biased but simpler model than an unbiased but more complicated model. This issue is enhanced in the two-dimensional context, because of the extra dimension in the domain of the function, and for this reason, we consider only the case $p = 1$. We introduce the Functional Autoregressive model of order 1 in Section 3.1 and propose estimation techniques in Section 3.2.

3.1. FAR(1) model

The most popular statistical model used to capture temporal dependence between functional observations is the functional autoregressive process. The theory of functional autoregressive processes in Hilbert spaces is developed in the pioneering monograph of Bosq 2000 and a comprehensive collection of statistical advancements for the FAR model can be found in Horváth and Kokoszka 2012.

A sequence of mean zero random functions $\{Y_t\}_{t=1}^{T} \subset \mathbb{H}$ follows a non-concurrent Functional Autoregressive Process of order 1 if:

$$Y_t = \Psi Y_{t-1} + \varepsilon_t \qquad t = 2, \dots, T, \tag{13}$$

where $\{\varepsilon_t\}_{t\in\mathbb{N}}$ is a sequence of iid mean-zero innovation errors with values in $\mathbb{H}$ satisfying $\mathbb{E}[||\varepsilon_t||^2] < +\infty$ and Ψ is a linear bounded operator from $\mathbb{H}$ to $\mathbb{H}$. We consider Ψ to be a Hilbert-Schmidt operator with kernel ψ, in such a way that:

$$(\Psi x)(u, v) = \int_c^d \int_e^f \psi(u, v; w, z) x(w, z) dw dz \quad a.e., \forall (u, v) \in [c, d] \times [e, f]. \tag{14}$$

In order to ensure existence of a stationary solution of (13), one has to require that $\exists j_0 \in \mathbb{N}$ such that $||\Psi||^{j_0} < 1$ (Bosq 2000, Lemma 3.1).

3.2. FAR(1) estimation

Proceeding similarly to Horváth and Kokoszka 2012, and adopting an approach akin to the Yule-Walker estimation in the scalar setting, we propose the following estimator of Ψ:

$$\hat{\Psi}_M x = \frac{1}{T-1} \sum_{i,j=1}^{M} \sum_{t=1}^{T} \hat{\lambda}_j \langle x, \hat{\xi}_j \rangle \langle Y_{t-1}, \hat{\xi}_j \rangle \hat{\xi}_i \qquad a.e., \tag{15}$$

where $\xi_1, \dots, \xi_M$ are the first M normalized functional principal components (FPC's), $\lambda_1, \dots, \lambda_M$ are the corresponding eigenvalues, and $\langle x, \xi_1 \rangle, \dots, \langle x, \xi_M \rangle$ are the scores of x along the FPC's. Appendix C illustrates two different estimation techniques for ξ_i and λ_i, one based on a discretization of functions on a fine grid and the other designed starting from an expansion of data on a finite basis system. We further refer to Appendix B for more details on the derivation of estimator (15) and for an extensive discussion on how to adapt it to the Conformal Prediction setting, where Ψ is estimated from the training set I_1 only.

Another forecasting procedure based on FPC's has been proposed by Aue et al. 2012 for one-dimensional functional data and is here extended to the two-dimensional setting. Calling once again $\xi_1, \dots, \xi_M$ the first M functional principal components, we decompose the Functional Time Series as follows:

$$Y_t(u, v) = \sum_{j=1}^{M} \langle Y_t, \xi_j \rangle \xi_j(u, v) + e_t(u, v) = \boldsymbol{Y}_t^T \boldsymbol{\xi}(u, v) + e_t(u, v), \tag{16}$$

where $\boldsymbol{Y}_t = [\langle Y_t, \xi_1 \rangle, \dots, \langle Y_t, \xi_M \rangle]^T$ contains the projection scores, $\boldsymbol{\xi}(u, v) = [\xi_1(u, v), \dots, \xi_M(u, v)]^T$ collects the evaluated principal components, and $e_t(u, v)$ is the approximation error due to the expansion's truncation on the first M principal

components. Neglecting the approximation error e_t, one can prove that the vector $\boldsymbol{Y}_t$ follows a multivariate autoregressive process of order 1 (VAR(1)). Plugging in the estimated FPCs $\hat{\xi}_1, \dots, \hat{\xi}_M$, we can estimate the parameters of the resulting VAR(1) model using standard techniques of multivariate statistics and forecast $\hat{\boldsymbol{Y}}_{T+1}$ based on historical data $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_T$. The predicted function $\hat{Y}_{T+1}$ is then reconstructed as:

$$\hat{Y}_{T+1}(u, v) = \hat{\boldsymbol{Y}}_{T+1}^T \boldsymbol{\xi}(u, v). \tag{17}$$

We finally introduce a model that may appear simplistic, since it does not exploit the possible time dependence between functions' values in different points of the domain, but that in practical applications provides satisfying results. The prediction method assumes an autoregressive structure in each location (u, v) of the domain, ignoring the dependencies between different points. We call this model a *concurrent* FAR(1):

$$Y_t(u, v) = \psi_{u,v} Y_{t-1}(u, v) + \varepsilon_t(u, v) \qquad \forall (u, v) \in [c, d] \times [e, f], \tag{18}$$

where $\psi_{u,v} \in \mathbb{R}$ and $t = 2, \dots, T$. Supposing to have observed functional data $y_1, \dots, y_T$ on a common two-dimensional grid $\{(u_i, v_j), i = 1, \dots, N_1, j = 1, \dots, N_2\}$, we can estimate ψ_{u_i,v_j} for each location (u_i, v_j).

4. Simulation study

4.1. Study design

The goal of this section is twofold: we aim to assess the quality of the proposed CP bands and evaluate different point predictors in terms of the resulting prediction regions. Since this work is focused on uncertainty quantification, we compare forecasting performances by means of the resulting Conformal Prediction bands. Firstly and foremost, we estimate the unconditional coverage by computing the *empirical unconditional coverage* in order to compare it with the nominal confidence level $1 - \alpha$. In the second place, we consider the size of the prediction bands, since a small prediction region is preferable as it includes subregions of the sample space where the probability mass is highly concentrated (Lei et al. 2015) and it is typically more informative in practical applications.

We employ as a data generating process a FAR(1) model in order to evaluate the estimation routines presented in Section 3.1. In order to benchmark forecasting performances, we examine the forecasting methods against a naive one: $\hat{Y}_{T+1} = Y_T$. By including a forecasting algorithm that is not coherent with the data generating process, we can illustrate how the presented CP procedure performs when a good point predictor g_{I_1} is not available. Although as reported in Section 3 a sufficiently accurate forecasting algorithm is necessary to guarantee asymptotic validity, we notice that in the simulations CP bands remain valid even when such assumption is not met.

To obtain further insights, we include the performances obtained by assuming perfect knowledge of the operator Ψ. For ease of reference, we list here the forecasting algorithms, introducing some convenient notation.

- **FAR(1)-Concurrent** refers to the forecasting algorithm based on a concurrent FAR(1) model (18).
- **FAR(1)-EK** (Estimated Kernel) denotes the first estimation procedure presented in Section 3.1, where we explicitly compute $\hat{\Psi}_M$ as prescribed by (15) and then set $\hat{Y}_{T+1} = \hat{\Psi}_M Y_T$.
- **FAR(1)-EK+** (Estimated Kernel Improved) is a modification of the above method, where eigenvalues $\hat{\lambda}_i$ are replaced by $\hat{\lambda}_i + 1.5(\hat{\lambda}_1 + \hat{\lambda}_2)$, as recommended by Didericksen et al. 2010 and discussed in Appendix B.
- **FAR(1)-VAR** denotes the forecasting procedure (17), where we exploit the expansion on estimated functional principal components and forecast Y_{T+1} using the underlying VAR(1) model.
- **Naive**: we just set $\hat{Y}_{T+1} = Y_T$. This method does not attempt to model temporal evolution, it is only included to see how much can be gained by exploiting the autoregressive structure of data.
- **Oracle**: we set $Y_{T+1} = \Psi Y_T$, using the same exact operator Ψ from which data are simulated. This point predictor is clearly not available in practical application, but it is interesting to include it in order to see if poor predictions might be due to poor estimation of Ψ.

When it is required (namely in FAR(1)-EK, FAR(1)-EK+, FAR(1)-VAR), FPCA is performed using the discretization approach, as motivated in Appendix C, truncating the representation to the first 4 harmonics.

In Section 4.2, we fix the size b of the blocking scheme (6) equal to 1 and let the sample size T take values $19, 49, 99, 499$. Secondly, in Section 4.3, we instead fix the sample size equal to 119 and repeat the simulations with $b = 1, 3, 6$. As usually done in the time series setting, the first observation is taken into account as a covariate only and does not enter neither the training set nor the calibration set. The proportion of data in the training and in the calibration set are hence equal to one half of the remaining observations: $m = l = (T - 1)/2$. For each value of T, we repeat the procedure by considering $N = 1000$ simulations. Simulations are implemented in the R Programming Language (R Core Team 2020).

In order to simulate a sequence of functions $\{Y_t\}_{t=1,\dots,T}$ from a FAR(1), we assume that observations lie in a finite dimensional subspace of the function space $\mathbb{H}$. Without loss of generality, throughout this section we consider functions in $\mathbb{H} = \mathcal{L}^2([0, 1] \times [0, 1])$. $\mathbb{H}$ is spanned by orthonormal basis functions $\phi_1, \dots, \phi_M$, with $M \in \mathbb{N}$ representing the dimension of such subspace. Therefore, we have:

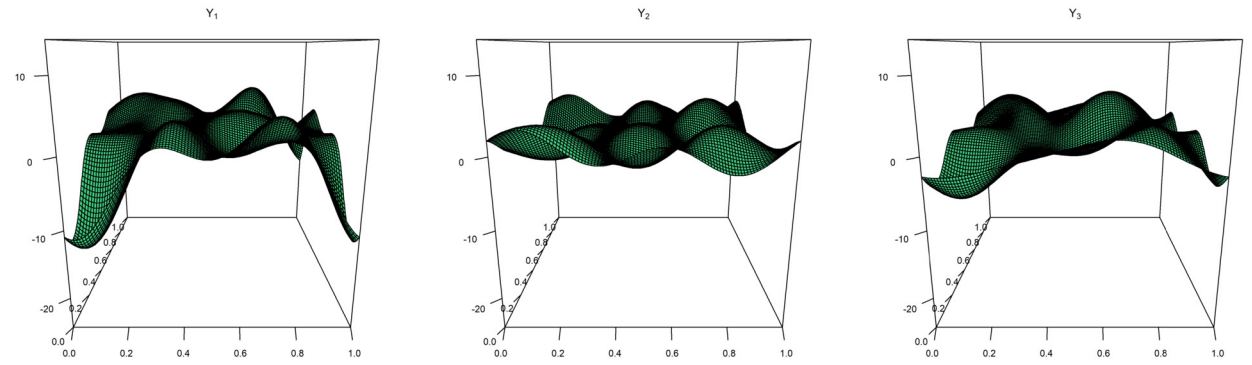


Fig. 2. Three consecutive realizations of a simulated FAR(1) represented on a grid of 10^4 points.

$$Y_t(u,v) = \boldsymbol{\phi}(u,v)^T \boldsymbol{Y}_t, \tag{19}$$

$$\varepsilon(u,v) = \boldsymbol{\phi}(u,v)^T \boldsymbol{\varepsilon}_t, \tag{20}$$

$$\psi(u,v;w,z) = \boldsymbol{\phi}(u,v)^T \boldsymbol{\Psi}\boldsymbol{W}\boldsymbol{\phi}(w,z), \tag{21}$$

where $\boldsymbol{\phi}(u,v) = [\phi_1(u,v),\dots,\phi_M(u,v)]^T \in \mathbb{R}^M$, $\forall (u,v) \in [0,1]\times[0,1]$, $\boldsymbol{Y}_t$, $\boldsymbol{\varepsilon}_t \subset \mathbb{R}^M$, $\forall t = 1,\dots,M$ and $\boldsymbol{\Psi} \in \mathbb{R}^{M\times M}$ and $\boldsymbol{W} \in \mathbb{R}^{K\times K}$, defined as $\boldsymbol{W} := \int_c^d \int_e^f \boldsymbol{\phi}(u,v)\boldsymbol{\phi}(u,v)^T du dv$. It follows that:

$$\boldsymbol{Y}_t = \boldsymbol{\Psi}\boldsymbol{Y}_{t-1}\boldsymbol{W} + \boldsymbol{\varepsilon}_t \qquad t = 1,\dots,T. \tag{22}$$

The basis system $\phi_1,\dots,\phi_M$ is constructed as the tensor product basis of two cubic B-spline systems $\{g_i\}_{i=1,\dots,M_1}$, $\{h_j\}_{j=1,\dots,M_2}$, both defined on $[0,1]$. We set $M_1 = M_2 = 5$, so that $M = 25$. For a discussion on the tensor product basis system, we refer to Appendix C.2 The matrix $\boldsymbol{\Psi}$ is defined as $\boldsymbol{\Psi} := 0.7\frac{\tilde{\boldsymbol{\Psi}}}{||\tilde{\boldsymbol{\Psi}}||_F}$, with $\tilde{\boldsymbol{\Psi}}$ having diagonal values equal to 0.8 and out-diagonal elements equal to 0.3 Innovation errors $\boldsymbol{\varepsilon}_t$ are independently sampled from a multivariate normal distribution, with mean zero and covariance matrix $\boldsymbol{\Sigma}$ having diagonal elements equal to 0.5 and out-diagonal entries equal to 0.3. Fig. 2 depicts an example of a simulated Functional Autoregressive Process of order one. A GIF of the time-evolving FAR(1) process can be found on GitHub.

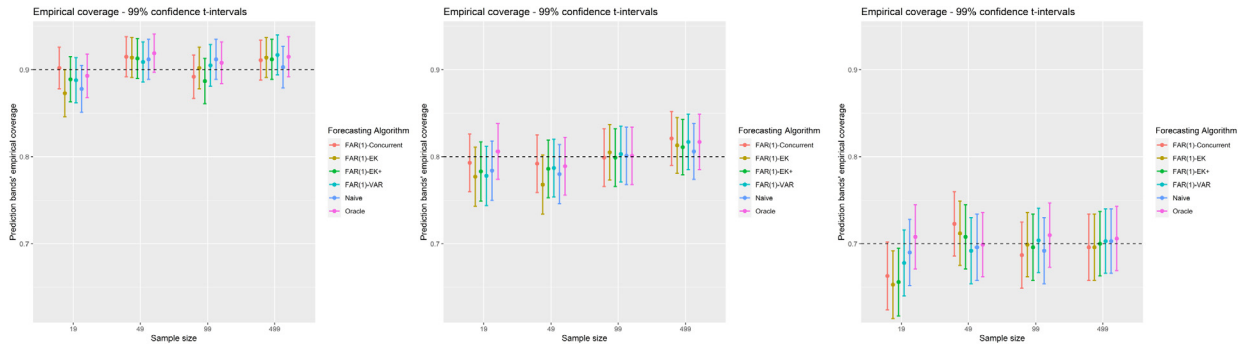
Notice that simulations have been designed in such a way to generate a *stationary process*. This condition is important to guarantee the existence of a solution to the FAR(1) equation (13), and is here guaranteed by setting $||\Psi|| < 1$, which satisfies the sufficient condition for stationary presented in Lemma 3.1 of Bosq 2000. One can indeed prove that, if relation (21) holds, then $||\Psi|| = ||\boldsymbol{\Psi}||_F$, where $||.||$ is the usual operatorial norm and $||.||_F$ denotes the Frobenius norm. In this way, FAR(1) estimation techniques are well-defined. For what concerns the theoretical assumptions of the CP scheme, proving the hypothesis of Theorem 1 is difficult, in particular in the context of functional data. To the best of our knowledge, no test for strongly mixing two-dimensional time series has been proposed, and testing the bounds on the oracle nonconformity scores is even more challenging. However, we aim to show here how the CP procedure can still be applied to obtain valid and efficient prediction bands.

4.2. Varying the sample size

We first fix the size b of the blocking scheme equal to 1 and let the sample size T take values $19, 49, 99, 499$. We replicate the experiments with different significance levels: $\alpha = 0.1$, $\alpha = 0.2$ and $\alpha = 0.3$, in order to assess how the confidence level influences the coverage and the width of the prediction bands.

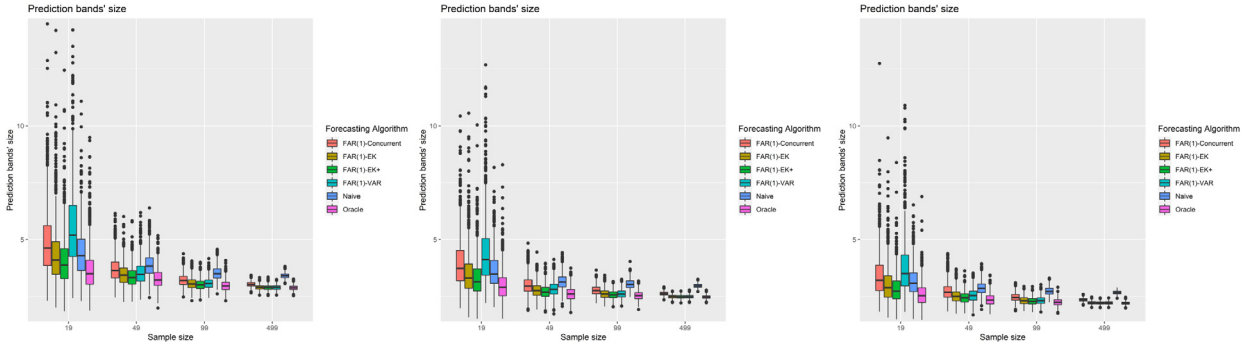
Fig. 3 shows the empirical coverage, together with the related 99% confidence interval. Empirical coverage is computed as the fraction of the $N = 1000$ replications in which y_{T+1} belongs to $C_{T,1-\alpha}(x_{T+1})$, and the confidence interval is reported in order to provide insights into the variability of the results, rather than to draw inferential conclusions about the unconditional coverage. Notice that different point predictors might intrinsically have dissimilar coverages, consequently this analysis aims to compare forecasting algorithm in terms of their predictive performances. We can appreciate that the 99% confidence interval for the empirical coverage almost always includes the nominal confidence level, regardless of the sample size at disposal. The only exception is obtained with $\alpha = 0.3$ and $T = 19$. In this case, the method produces very narrow prediction bands (Fig. 4c), that result in an empirical coverage smaller than the nominal one. This behavior however disappears as soon as the sample size T increases. It is also interesting to notice that, even when an accurate forecasting algorithm g_{I_1} is not available (namely with the Naive predictor), the proposed CP procedure still outputs prediction regions with a high unconditional coverage.

Similarly to Diquigiovanni et al. 2022, we define the size of a two-dimensional prediction band as the *volume* between the upper and the lower surfaces that define the prediction band:



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

Fig. 3. Simulation study, increasing the sample size. Empirical coverage of CP bands. Dashed line represents nominal coverage $1 - \alpha$.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

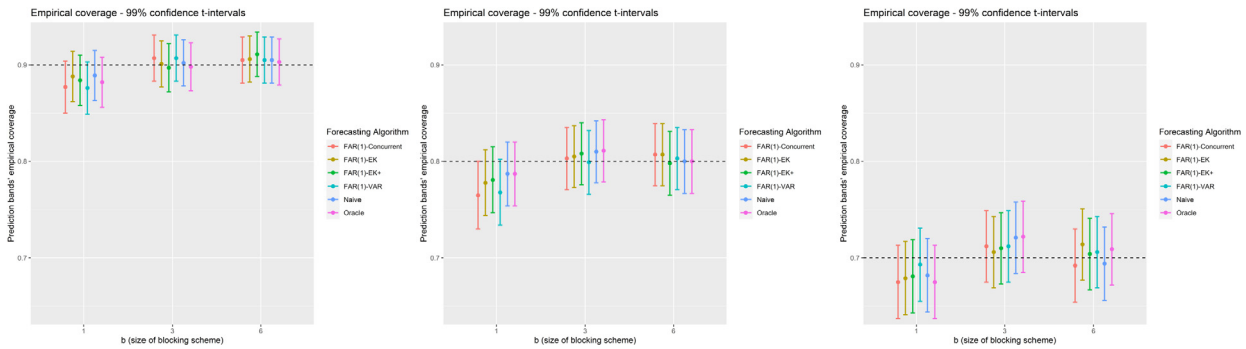
Fig. 4. Simulation study, increasing the sample size. Size of CP bands.

$$Q(s_{I_1}) := \int_0^1 \int_0^1 2k s_{I_1}(u, v) du dv = 2k. \tag{23}$$

Measuring the size of the correspondent prediction bands, we can compare the efficiency of different forecasting routines. We stress the fact that distinct point predictors may guarantee potentially different coverage levels. For this reason, it is crucial to first evaluate the empirical coverage of the resulting prediction bands and only afterward investigate their size. Fig. 4 reports boxplots with prediction bands' size for the $N = 1000$ simulations and for different values of α and T. Bands' size tends to decrease as long as the number of observations T increases, hence improving the efficiency of prediction sets. Moreover, the size tends to decrease when the confidence level $1 - \alpha$ increases. As expected, Naive predictor provides larger prediction bands, particularly in the large sample settings. On the other hand, FAR(1)-EK and FAR(1)-EK+, both based on the estimation of the autoregressive operator Ψ, provide the tightest prediction bands, not only when numerous observations are available, but also in small sample sizes scenario. Notice also that eigenvalue correction slightly improves the performances of FAR(1)-EK+ wrt FAR(1)-EK, especially when few samples are available. We acknowledge that, when $T = 19$, VAR-efpc performs remarkably worse than the other methods. We argue that this performance gap might be caused by the simultaneous OLS estimation of the underlying VAR(1) equations, which might provide biased estimates if the sample size is small. However, when the sample size increases, such forecasting algorithm performs comparably with the already mentioned FAR(1)-EK and FAR(1)-EK+. Finally, although the Conformal Prediction bands produced by the oracle predictor are obviously the most performing one, both FAR(1)-EK and FAR(1)-EK+ provide CP bands with coverage and size comparable to the theoretically perfect oracle forecasting method.

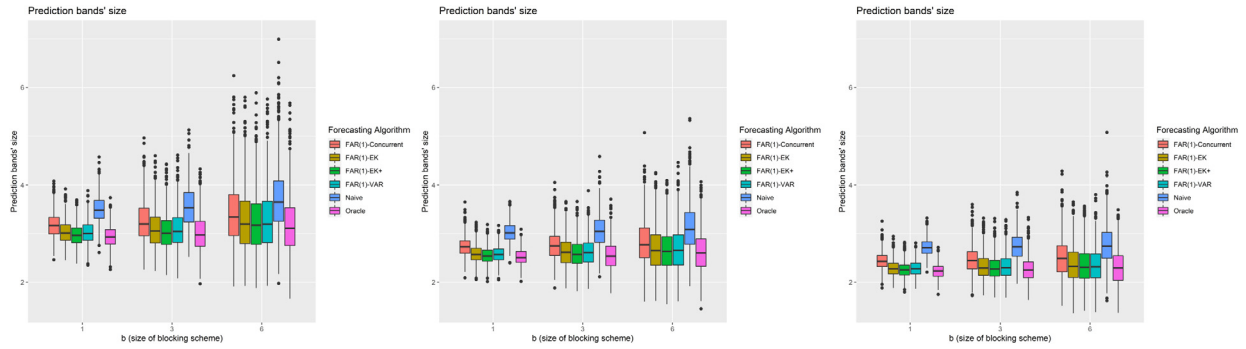
4.3. Varying the blocking scheme size

This time, we fix the sample size T and let the blocking scheme b vary, in order to determine how the validity and efficiency of the resulting prediction bands are influenced by such parameter. Once again, we repeat the experiments for $\alpha = 0.1$, $\alpha = 0.2$, $\alpha = 0.3$, whereas the value of T is fixed and equal to 119, which provides a good balance between scenarios with small and large sample sizes. Analogous results have been found by letting the value of T vary. Results are reported in Fig. 5 and Fig. 6. Once again, in all circumstances the empirical coverage is close to the nominal one, confirming validity of CP bands even for higher values of b. Moreover, one can notice that, as already pointed out by Diquigiovanni



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

Fig. 5. Simulation study, increasing the blocking scheme size. Empirical coverage of CP bands. Dashed line represents nominal coverage $1 - \alpha$.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

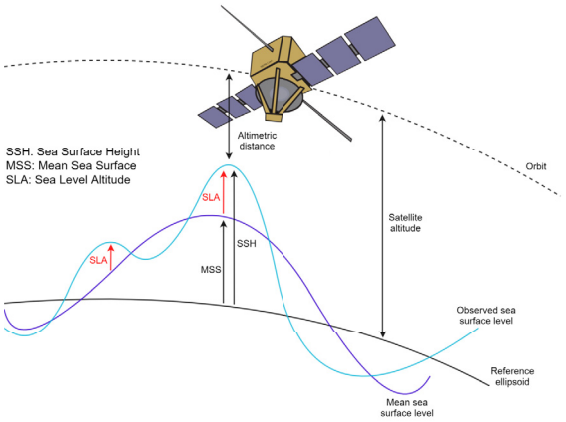
Fig. 6. Simulation study, increasing the blocking scheme size. Size of CP bands.

et al. 2021, the band size tends to decreases when b decreases, thus providing more efficient prediction regions. We argue that this behavior is related to the inverse proportionality between the blocking scheme size b and the dimension of permutation family $|\Pi|$. Finally, notice that larger prediction bands attain as expected a larger empirical coverage, and that larger values of α (smaller confidence level $1 - \alpha$) result in smaller prediction bands. A comparison of forecasting algorithms performances validates the considerations in Section 4.2.

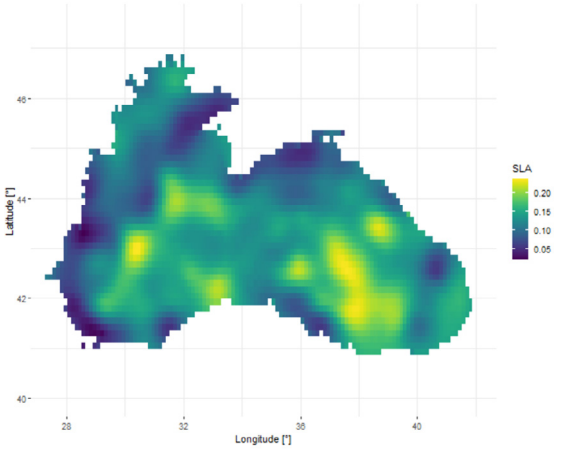
5. Case study: forecasting Black Sea level anomalies

5.1. Dataset

In this section, we aim to illustrate the application potential of the proposed methodology on a proper case study. We analyze a data set from Copernicus Climate Change Service (C3S), a project operated by the European Center for Medium-Range Weather Forecasts (ECMWF), collecting daily sea level anomalies of the Black Sea in the last twenty years (Mertz and Legeais 2018). Sea level anomalies are measured as the height of water over the mean sea surface in a given time and region. Specifically, altimetry instruments give access to the sea surface height (SSH) above the reference ellipsoid, which is calculated as the difference between the orbital altitude of the satellite and the measured altimetric distance of the satellite from the sea (see Fig. 7a). Starting from this information, Sea Level Anomaly (SLA) is defined as the anomaly of the signal around the Mean Sea Surface component (MSS), which is computed with respect to a 20-year reference period (1993-2012). Observations are collected on a spatial raster, with a resolution of 0.125° both on the longitude and on the latitude axis. Since observations are collected on a geoid, the domain lies on a two-dimensional manifold, however, because both longitude and latitude ranges are very small (14° and 7° respectively), we assume data to be observed on a bidimensional grid. The resulting lattice can hence be considered as the Cartesian product of a grid on the longitude axis made by $N_1 = 120$ points and a latitude grid of $N_2 = 56$ points. We refer to (u_i, v_j), with $i = 1, \dots, N_1$ and $j = 1, \dots, N_2$ as the (i, j)-th point of such two-dimensional mesh. Since the Black Sea does not have a rectangular shape, we model data as realization of random surfaces defined on the rectangle circumscribed to the perimeter of the sea, but identically equal to zero outside of it. As a consequence, being $\mathcal{B}$ the set of points internal to the perimeter of the Black Sea, we slightly redefine the non conformity measure (3) to become:



(a) Data collection process. The image is an author's replica of Figure 1 in Copernicus' Product User Guide and Specification v2.4.



(b) Raster representation of the Sea Level Anomaly (in meters) on 01/01/2018

Fig. 7. Copernicus Black Sea dataset.

$$\mathcal{A}(\{z_h : h \in I_1\}, z) = \operatorname*{ess\,sup}_{(u,v)\in[c,d]\times[e,f]} \mathcal{R}(u, v), \tag{24}$$

where $\mathcal{R}(u, v)$ is defined as:

$$\mathcal{R}(u, v) \begin{cases} \frac{|y(u,v) - g_{I_1}(u,v;x_{T+1})|}{s_{I_1}(u,v)}, & \text{if } (u, q) \in \mathcal{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Hereafter, we will consider the time series $\{SLA_t\}$, without making explicit the dependence on the bivariate domain.

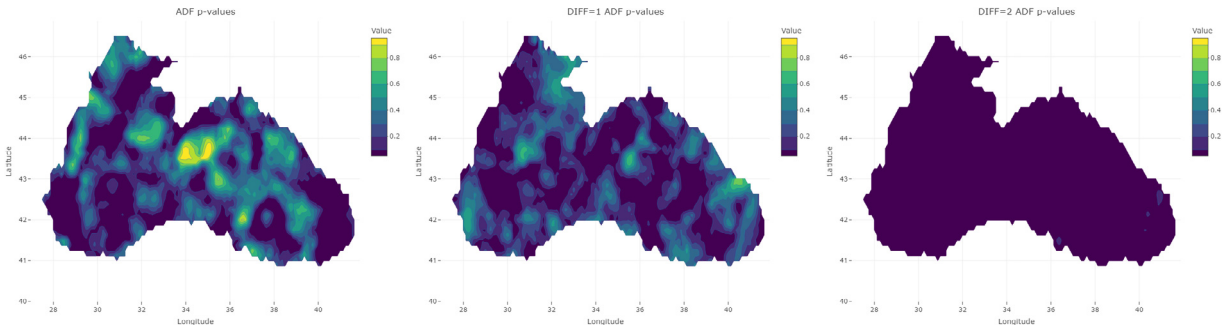
5.2. Data preprocessing

If possible, one would preferably forecast directly the time series of Sea Level Anomalies (SLA_t). However, in order to estimate the FAR(1) process, we would also like to guarantee stationarity of the time series at our disposal, and given the particular nature of the dataset, we expect observations to exhibit a strong seasonality as well as a linear trend. Indeed, both tide gauge and altimetry observations show that sea level trends in the Black Sea vary over time (Avsar et al. 2016, Cazenave et al. 2001). Tsimplis and Baker 2000 estimated a rise in the mean sea level of 2.2 mm/year from 1960 to the early 1990s, while long-track altimetry data indicate that sea level rose at a rate of 13.4 ± 0.11 mm/year over 1993–2008 (Ginzburg et al. 2011).

In order to further investigate this issue, we should proceed by testing the functional time series $\{SLA_t\}_t$ for stationarity. However, while for one-dimensional Functional Time Series one could resort to the tests proposed by Horváth et al. 2014 or Aue and van Delft 2020, to the best of our knowledge no stationarity test for two-dimensional functional time series has been developed. For such reason, and aware of the limits of this approach, we resort to analyzing stationarity of univariate time series $SLA_t(u_i, v_j)$, where each (u_i, v_j) represents a grid point in the lattice. We stress the fact that stationarity of individual time series does not guarantee stationarity of the underlying functional process, and this constitutes only a *necessary* and not sufficient condition. Therefore, the goal is not to derive inferential results on the stationarity of the process, but rather to describe the evolution of the process by means of its individual component, and to potentially obtain a better time series to work with. We test each univariate time series for stationarity, using the Augmented Dickey Fuller (ADF) test. We report in Fig. 8 a grid map of the p-values, for $\{SLA_t\}_t$, $\{\Delta SLA_t\}_t$ and $\{\Delta^2 SLA_t\}_t$, where Δ and Δ^2 denote respectively one and two differentiations. Despite the fact that we can't make inferential conclusions on the stationarity of the process based on the individual tests, we can see that the original time series exhibit a very non-stationary behavior, and after one differentiation there are many non-stationary locations. After two differentiations, all the individual time series can be confidently considered stationary. We argue that the Functional Time Series might exhibit a behavior similar to one described in terms of stationarity, and hence proceed by differentiating twice the process, defining

$$Y_t := \Delta^2 SLA_t = (SLA_t - SLA_{t-1}) - (SLA_{t-1} - SLA_{t-2}). \tag{26}$$

We forecast the differentiated time series Y_t, and obtain prediction bands for it, using the methodology presented in Section 2. However, in order to provide a better insight into the prediction problem, we need to retrieve results pertaining to the original time series SLA_t. Specifically, we apply the conformal machinery to the differentiated time series, calling $\hat{Y}_{T+1}$ the forecasted function, and obtaining the prediction band for Y_{T+1}:



(a) ADF test on $\{SLA_t\}_t$ (b) ADF test on $\{\Delta SLA_t\}_t$ (c) ADF test on $\{\Delta^2 SLA_t\}_t$

Fig. 8. ADF test's p-values on individual time series. From the left to the right: test on original data, test after one differentiation and test after two differentiations.

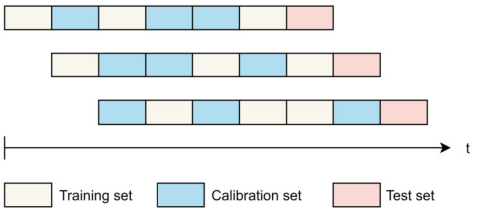


Fig. 9. Example of the training-calibration-test split in a rolling window scenario.

$$C_{T,1-\alpha} := \left\{ y \in \mathbb{H} : y(u,v) \in \left[\hat{Y}_{T+1}(u,v) \pm k^s s_{I_1}(u,v) \right] \forall (u,v) \right\}.$$

Exploiting the fact that:

$$Y_{T+1}(u,v) \in \left[\hat{Y}_{T+1}(u,v) \pm k^s s_{I_1}(u,v) \right] \quad \forall (u,v)$$
$$\iff$$
$$SLA_{T+1}(u,v) \in \left[\hat{Y}_{T+1}(u,v) + 2SLA_T(u,v) - SLA_{T-1}(u,v) \pm k^s s_{I_1}(u,v) \right] \quad \forall (u,v),$$

we define the prediction band for SLA_{T+1} as:

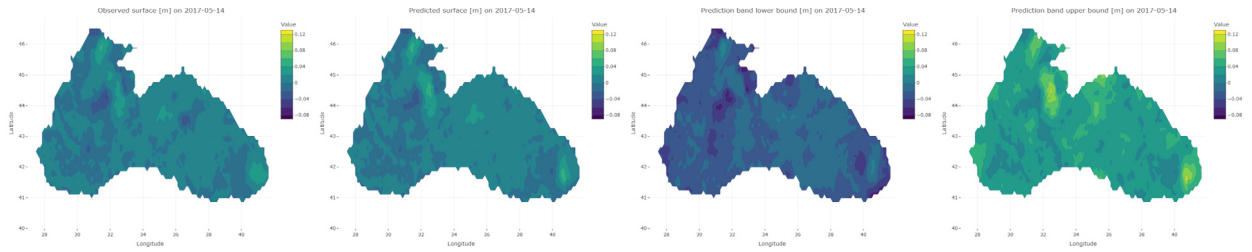
$$\tilde{C}_{T,1-\alpha} := \{ y \in \mathbb{H} : y(u,v) \in \left[\hat{Y}_{T+1}(u,v) + 2SLA_T(u,v) - SLA_{T-1}(u,v) \pm k^s s_{I_1}(u,v) \right] \forall (u,v) \}.$$

Finally, notice once again that testing the hypothesis underlying the CP procedure is not possible in this setting. We are nevertheless interested in applying the CP scheme together with the forecasting techniques in order to verify their empirical performances.

5.3. Study design

The case study employs a rolling estimation framework which recalculates the model parameters on a daily basis and consequently shifts and recomputes the entire training, calibration and test windows by 24 hours, as shown in Fig. 9. As before, we use a random split of data in the training and calibration sets, with split proportion equal to 50%. The significance level α is once again fixed equal to 0.1. For each of the 1000 days we aim to predict, we build the corresponding prediction band based on the information provided by the last 99 days, thereby fixing $T = 99$. Choosing this sample size provides accurate forecasts and thus small prediction bands, while maintaining reasonable computational times. The size of the blocking scheme is fixed to 1, since, as motivated in Section 4.3, this choice produces the narrowest prediction bands, preserving at the same time satisfactory performance in terms of empirical coverage. The rolling window is shifted 1000 times, thus iterating for almost three years the forecasting scheme. More specifically, and to allow for reproducibility of subsequent results, we consider a rolling window ranging from 01/01/2017 to 04/01/2020.

The point predictors used throughout this application are those described in Section 3. The number of Functional Principal Components is set equal to 8. For each shift of the rolling window and for each forecasting algorithm, we check if SLA_{T+1} belongs to $\tilde{C}_{T,1-\alpha}(x_{T+1})$, and save the size of the corresponding prediction band. After collection of results, we calculate the average coverage, and use it to compare performances of the different point predictors in this scenario.



(a) Observed Surface (b) Predicted Surface (c) Prediction Band Lower Bound (d) Prediction Band Upper Bound

Fig. 10. Observed surface, predicted surface, prediction band lower bound, prediction band upper bound on 14/05/2017. Results are obtained using a sample size T equal to 99, $b = 1$, $l = 48$, $\alpha = 0.1$. The employed forecasting algorithm g_{I_1} is FAR(1)-EK (15).

5.4. Results

Fig. 10 outlines observed and forecasted surfaces obtained for one of the days in the rolling window, as long as the lower and upper bounds defining the prediction band. For the sake of simplicity, we display only results obtained using the FAR(1)-EK estimator (15), since, as discussed below, it provides on average the narrowest prediction bands. In order to allow for a more insightful analysis, and to further investigate the evolution of the surfaces, we implemented a dedicated Shiny App available online where results can be explored. We report in Fig. 11a the average coverage of CP bands obtained across 1000 predictions. As in Section 4, we pair such quantity with a 99% confidence interval. Notice that in this case the confidence interval may be biased, due to the inevitable correlation between data used to construct it, however, we include it in order to assess the dispersion of the average coverage around the mean. Coherently with the results of the simulation study, we can appreciate that in all cases the prediction bands capture the observed surface y_{T+1} approximately $(1 - \alpha)\%$ of the times, regardless of the forecasting algorithm used.

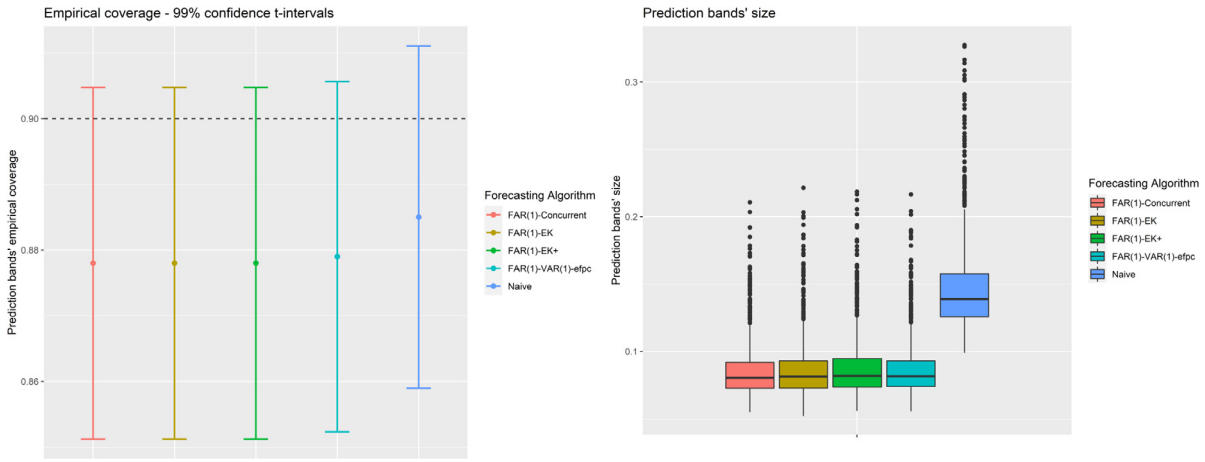
For what concerns the size of the prediction bands, the Naive predictor produces by far the widest ones (see Fig. 11b), and, this fact does not reflect in a greater coverage compared to the other methods. On the other hand, prediction bands obtained with autoregressive forecasting algorithms provide narrower prediction regions. Among these, we can see that the non-concurrent FAR(1) is the most performing one, regardless of the way in which it is estimated (namely with FAR(1)-EK, FAR(1)-EK+ or FAR(1)-VAR). Nevertheless, also the concurrent FAR(1) model provides very tight prediction bands, almost comparable with the ones produced by the non-concurrent prediction algorithm.

We are also interested in analyzing the *pointwise* properties of CP bands in this scenario. Therefore, we display in Fig. 12 a map of the pointwise coverage of the prediction bands, obtained using FAR(1)-EK. We can appreciate, as expected, a high empirical coverage across the entire domain, emphasizing once again the peculiarity of our approach, which guarantees *global* coverage of the prediction surfaces, reflected by an obvious pointwise coverage higher than the nominal one. We report in Fig. 12 the average width of CP bands, which denotes a peculiar pattern, likely caused by data collection routines. Indeed, we can see from Fig. 13b, that a similar behavior observed in the map of pointwise width can be found by plotting the standard deviation of original data. This is coherent with the employed CP framework, since the size of prediction bands depends on the amplitude of the functional standard deviation.

In conclusion, this case study confirms the validity of our procedure and proves how a FAR(1) model significantly improves the predictive efficiency even in this more complex scenario.

6. Conclusions and further developments

In this work, we introduce a mathematical framework for probabilistic forecasting of two-dimensional Functional Time Series. Leveraging the CP scheme developed by Chernozhukov et al. 2018 and Diquigiovanni et al. 2021 and adapting it to the 2D-FTS setting, we propose technique for quantifying uncertainty when predicting time evolving surfaces. In order to provide point prediction of surfaces, we model functions through a Functional autoregressive process of order one, extending the mathematical theory of autoregressive processes to allow for bivariate functions. Estimations techniques for the FAR(1) are presented and compared We test the benefits and limits of the proposed approach, first on synthetic data and then on a real novel time series dataset, collecting daily observations of sea level anomaly over the Black Sea. Empirical results proved the validity of the methodology on non-synthetic data. We acknowledge that in applying the proposed procedure to the case study, we had to introduce some simplifications due to the novelty of the subject and the limited amount of work on two-dimensional Functional Time Series. We hope that this work will encourage the development of novel analysis techniques for 2D functional data, such as stationarity tests and other forecasting tools. Finally, throughout the work, we limited the analysis to the FAR(p), with $p = 1$, as motivated in section 3. One may consider a FAR(p) model, with $p > 1$. Whereas it is not straightforward to extend the estimation of Ψ_M (15) to a FAR(p) model with $p > 1$, both the concurrent estimator (18) and the estimator based on an expansion of FPC's, may be easily adapted to the case $p > 1$. The problem can in fact be rendered as a standard functional linear regression and solved by using the many off-the-shelf methods apt to the task and present in the literature (see e.g. Chiou et al., 2016). Finally, notice that thanks to the flexibility of our approach,



(a) Empirical coverage of CP bands. Dashed line represents nominal coverage $1-\alpha$.

(b) Size of CP bands.

Fig. 11. Results of the forecasting procedure in a rolling window setting. All the methods produce prediction bands with average coverages close to the nominal one. For what concerns predictive efficiency, the naive predictor produces the widest bands, whereas the other methods produce instead more efficient bands, with sizes similar to each other.

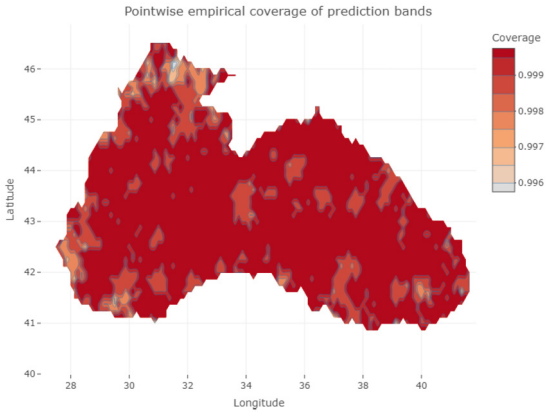
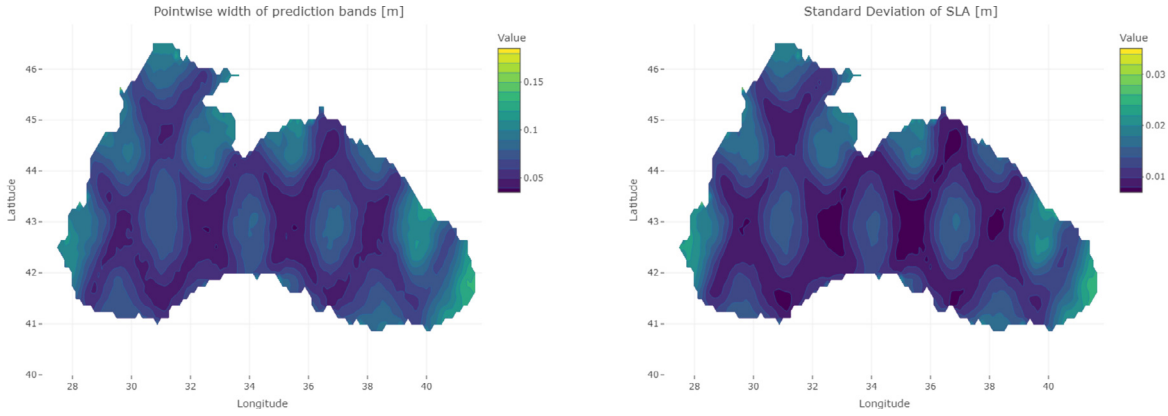


Fig. 12. Pointwise coverage of prediction bands across the domain, obtained by counting the number of times that each point falls between prediction bands, and dividing by the total number of time steps in the rolling window framework.



(a) Pointwise average width of CP bands.

(b) Functional standard deviation of original data.

Fig. 13. Average CP bands' pointwise width and functional standard deviation of original data. The figure on the right denotes a peculiar pattern in the data collection process, which is retrieved in the left panel, due to the choice of using the standard deviation as a modulation function in the CP framework.

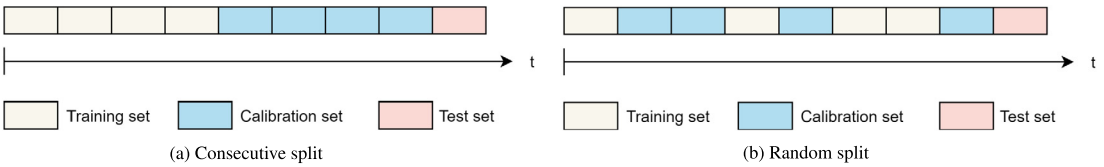


Fig. A.14. Possible types of split with $T = 8$, $m = l = 4$.

a modification of such kind can be easily achieved by changing the point predictor g_{I_1} only, while preserving the desirable properties of prediction bands.

Acknowledgements

The present research has been partially supported by MUR, grant Dipartimenti di Eccellenza 2023-2027 and by Accordo Attuativo ASI-POLIMI "Attività di Ricerca e Innovazione" n. 2018-5-HH.0, collaboration agreement between the Italian Space Agency and Politecnico di Milano. M.F. acknowledges the support of the JRC Centre of Advanced Studies CSS4P - "Computational Social Science for Policy".

Appendix A. Split ratio and type of split

The choice of the split ratio is non-trivial and has motivated discussion in the statistical community. Including more data in the training set improves the estimation of the point predictor $\hat{g}_{I_1}$. At the same time, having few data points in the calibration set produces a very rough p-value function (7), resulting in potential greater actual coverage with respect to the nominal one. This trade-off problem is enhanced in the time series context, in which one would like to have both training and calibration sets as large as possible, since asymptotic validity is guaranteed when both l and m go to infinity. Throughout this work, the training-calibration ratio is fixed equal to 50%-50%, as commonly suggested in literature. Moreover, we stress the fact that the split is random. This clearly introduce variability in the procedure since results depend on the particular division of data. We acknowledge a recent advancement in this direction, called Multi Split Conformal Prediction (Solari and Djordjilović 2021), which aggregates single split CP intervals across multiple splits.

Another interesting question regarding the type of split comes up in the time series context. Due to the lack of exchangeability, two different subdivisions are possible in this framework. A first choice could consist in a sequential division of data, where the split point is no longer random, but is a result of the training-calibration proportion (see A.14a). Wisniewski et al. 2020 applied this scheme in a rolling window fashion to forecast Market Makers' Net Positions. While this choice may seem more consistent in the presence of temporal dependence, since it does not split subsequent observations in two different sets, it may lead to very biased results if the training size m is very small or if data present a different trend or seasonal component in the training and calibration sets. The interested reader may refer to Kath and Ziel 2021 for a more comprehensive discussion on this topic. All in all, in order to make the model more robust, we preferred to split data randomly (A.14b).

Appendix B. FAR(1) estimation and adaptation to CP

Hereafter, we will assume $\{Y_t\}_{t=1}^{T} \subset \mathcal{L}^4(\Omega, \mathcal{F}, \mathbb{P})$ and consider only zero-mean stationary time series. The covariance operator $\Gamma_0 : \mathbb{H} \to \mathbb{H}$ and the lag-1 autocovariance operator Γ_1 can thus be defined as:

$$\Gamma_0 x = \mathbb{E}[\langle Y_t, x \rangle Y_t] \qquad \textit{almost everywhere } (a.e.), \tag{B.1}$$

$$\Gamma_1 x = \mathbb{E}[\langle Y_t, x \rangle Y_{t+1}] \qquad a.e. \tag{B.2}$$

Extending the work of Horváth and Kokoszka 2012 to a different functional space, we can derive the following estimators for such operators:

$$\hat{\Gamma}_0 x = \frac{1}{T} \sum_{t=1}^{T} \langle Y_t, x \rangle Y_t \qquad a.e., \tag{B.3}$$

$$\hat{\Gamma}_1 x = \frac{1}{T-1} \sum_{t=1}^{T-1} \langle Y_t, x \rangle Y_{t+1} \qquad a.e.. \tag{B.4}$$

Under rather general weak dependence assumptions these estimators are $\sqrt{n}$-consistent. One may, for example, adopt the concept of $\mathcal{L}^p$-m-approximability introduced in Hörmann and Kokoszka 2010 to prove that $\mathbb{E}[||\hat{\Gamma_0} - \Gamma_0||^2] = O(n^{-1})$, where $||.||$ is the operatorial norm: $||F|| := \sup_{||x||_{\mathbb{H}}=1} ||Fx||_{\mathbb{H}}$ for any linear bounded operator $F : \mathbb{H} \to \mathbb{H}$.

In order to derive estimator (15) of the FAR(1), we first consider the following operatorial equation:

$$\Gamma_1 = \Psi\Gamma_0. \tag{B.5}$$

A natural idea may consist in computing estimators $\hat{\Gamma}_0$, $\hat{\Gamma}_1$ from historical data and defining then $\hat{\Psi} = \hat{\Gamma}_0\hat{\Gamma}_0^{-1}$. Unfortunately, the inverse operator Γ_0^{-1} is unbounded on $\mathbb{H}$ (Horváth and Kokoszka 2012), however, thanks to Γ_0 being a symmetric, compact, positive-definite operator, one can exploit its spectral decomposition to introduce a pseudo-inverse operator $\Gamma_{0,M}^{-1}$, defined as:

$$\Gamma_{0,M}^{-1}x = \sum_{j=1}^{M} \lambda_j^{-1}\langle x, \xi_j\rangle\xi_j \qquad a.e., \tag{B.6}$$

where $\xi_1, \dots, \xi_M$ are the first M normalized functional principal components (FPC's), $\lambda_1, \dots, \lambda_M$ are the corresponding eigenvalues and $\langle x, \xi_1\rangle, \dots, \langle x, \xi_M\rangle$ are the scores of x along the FPC's. We formally define ξ_i and λ_i as eigenfunctions and eigenvalues that solve the functional equation:

$$\Gamma_0\xi_i = \lambda_i\xi \qquad i = 1, \dots, M. \tag{B.7}$$

We can now combine (B.5) and (B.6), plugging in estimated eigenfunctions and eigenvalues and calling $\hat{\Gamma}_{0,M}^{-1}$ the resulting estimator of $\Gamma_{0,M}^{-1}$, to finally derive:

$$\hat{\Psi}_M = \hat{\Gamma}_1\hat{\Gamma}_{0,M}^{-1}x, \tag{B.8}$$

$$\hat{\Psi}_M x = \frac{1}{T-1}\sum_{i,j=1}^{M}\sum_{t=1}^{T}\hat{\lambda}_j\langle x, \hat{\xi}_j\rangle\langle Y_{t-1}, \hat{\xi}_j\rangle\langle Y_t, \hat{\xi}_i\rangle\hat{\xi}_i \qquad a.e.. \tag{B.9}$$

Notice that the operator $\hat{\Gamma}_{0,M}^{-1}$ is bounded on $\mathbb{H}$ if $\hat{\lambda}_j$ are strictly greater than zero for $j = 1, \dots, M$. Nevertheless, even if such condition is met, in practice one should cautiously select the number of principal components M, because very small eigenvalues will result in very high reciprocals $\hat{\lambda}_j^{-1}$, providing in practice unbounded estimates of $\Gamma_{0,M}^{-1}$. Such observation motivated Didericksen et al. 2010 to add a positive baseline to the estimated eigenvalues $\hat{\lambda}_j$. This small modification improves the estimation of the operator Ψ, and most importantly, contributes to weaken the dependency of $\hat{\Psi}_M$ on M.

We now aim to adapt the previous estimator (B.9) to the conformal inference setting. The goal is to accommodate the forecasting algorithm in order to estimate the point predictor g_{I_1} from the training set only. As a preliminary step, given that the FAR(1) model has been presented for mean-centered data, one has to estimate the mean function $\hat{\mu}_{I_1}$ from the training set only and consequently center all the observations in the training and calibration set around $\hat{\mu}_{I_1}$. If the sample size at disposal is sufficiently large and if the stationarity assumption is fulfilled, it should make no great difference to estimate the population function μ with the sample mean $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} Y_t$ or with its restriction on the training set $\hat{\mu}_{I_1} = \frac{1}{m}\sum_{t\in I_1} Y_t$. Another fundamental step is the estimation of functional principal components. Since in the CP framework we are allowed to use only the information from the training set in order to compute $\hat{\xi}_1, \dots, \hat{\xi}_M$, it is then natural to employ only training data $\{z_h : h \in I_1\}$ in such estimation routine.

In order to obtain estimator $\hat{\Psi}_M$ (15), one has to compute $\hat{\Gamma}_1$ and $\hat{\Gamma}_{0,M}^{-1}$ from the training set. While the computation of the sample pseudo-inverse of the autocovariance estimator is straightforward:

$$\hat{\Gamma}_{0,M}^{-1}x = \frac{1}{m}\sum_{j=1}^{M}\hat{\lambda}_j\langle x, \hat{\xi}_j\rangle\hat{\xi}_j \qquad a.e., \tag{B.10}$$

the CP counterpart of $\hat{\Gamma}_1$ is more delicate and requires further discussion. Recall indeed that the classical estimator for the lag-1 autocovariance operator from $Y_1, \dots, Y_T$ is:

$$\hat{\Gamma}_1 x = \frac{1}{T-1}\sum_{t=1}^{T-1}\langle Y_t, x\rangle Y_{t+1} = \frac{1}{T-1}\sum_{t=2}^{T}\langle Y_{t-1}, x\rangle Y_t. \tag{B.11}$$

In the CP setting, however, we could define three different estimators for Γ_1:

$$\hat{\Gamma}_1 x = \frac{1}{m-1}\sum_{t\in I_1[1:m-1]}\langle Y_t, x\rangle Y_{t+1}, \tag{B.12}$$

$$\hat{\Gamma}_1 x = \frac{1}{m-1}\sum_{t\in I_1[2:m]}\langle Y_{t-1}, x\rangle Y_t, \tag{B.13}$$

$$\hat{\Gamma}_1 x = \frac{1}{m}\sum_{t\in I_1}\langle Y_{t-1}, x\rangle Y_t, \tag{B.14}$$

where $x \in \mathbb{H}$ and $I_1[i:j]$ contains the indices from the i-th to the j-th element of I_1. Notice that the three estimators differ because if $t \in I_1$, we have no assurance that $\{t-1\} \in I_1$ or even $\{t+1\} \in I_1$. We stress also the fact that the third operator is well-defined only if we reserve a burn-in set of length 1 at the front of the time series, in such a way that, if $\{2\} \in I_1$, we can still compute the estimator. Among the three options, we prefer the third one (B.14), since it averages over a larger set. One may also argue that such estimators are not coherent with the CP setting, since they are inevitably based on data from the calibration set. However, as mentioned before, we are considering the time series of *regression pairs* $Z_t = (X_t, Y_t)$, $t = 1, \dots, T$ (Chernozhukov et al. 2018). The key consideration is that, according to the FAR(1) model, we use as regressors the lagged version of the time series, namely $X_t = Y_{t-1}$ and the regression couples becomes $Z_t = (Y_{t-1}, Y_t)$, for each $t = 1, \dots, T$. From this perspective, one could rephrase the definition of the sample covariance operator by making explicit its dependence from the regressor X_t instead of Y_{t-1}:

$$\hat{\Gamma}_1 x = \frac{1}{m-1} \sum_{t \in I_1} \langle Y_{t-1}, x \rangle Y_t = \frac{1}{m} \sum_{t \in I_1} \langle X_t, x \rangle Y_t. \tag{B.15}$$

It is then straightforward to derive the estimator $\hat{\Psi}_{M,I_1}$:

$$\hat{\Psi}_{M,I_1} x = \frac{1}{m} \sum_{i,j=1}^{M} \sum_{t \in I_1} \hat{\lambda}_j \langle x, \hat{\xi}_j \rangle \langle Y_{t-1}, \hat{\xi}_j \langle Y_t, \hat{\xi}_i \rangle \hat{\xi}_i = \frac{1}{m} \sum_{i,j=1}^{M} \sum_{t \in I_1} \hat{\lambda}_j \langle x, \hat{\xi}_j \rangle \langle X_t, \hat{\xi}_j \langle Y_t, \hat{\xi}_i \rangle \hat{\xi}_i \qquad a.e.. \tag{B.16}$$

The point predictor finally becomes:

$$\hat{Y}_{T+1} = g_{I_1}(u, v; X_{T+1}) = (\hat{\Psi}_{M,I_1} X_{T+1})(u, v) = (\hat{\Psi}_{M,I_1} Y_T)(u, v) \tag{B.17}$$

In order to compute (B.16), it is first mandatory to project the calibration set onto the EPFC's in order to compute the scores $\langle X_h, \hat{\xi}_j \rangle$ for $h \in I_2$. Notice that, in the non-Conformal setting, such step comes for free when performing FPCA. In the CP context, however, EPFC's are computed from the training set only, therefore, in order to access the scores of the calibration set, one needs to explicitly add this projection step.

Appendix C. FPCA for two-dimensional functional data

A fundamental aspect in the design of many forecasting algorithms is the estimation of functional principal components (FPC's) $\{\xi_i\} i \in \mathbb{N}$. We define FPC's as functions $\xi_i \in \mathcal{L}^2([c,d] \times [e,f])$ solving the functional equation:

$$\Gamma_0 \xi = \lambda \xi. \tag{C.1}$$

In practice, we can only estimate the first $M \in \mathbb{N}$ eigenfunctions, implicitly performing dimensionality reduction. The choice of M is non-trivial and depends on the application framework. Whereas Kargin and Onatski 2005 suggested selecting it in a cross-validation setting, Aue et al. 2012 proposed a fully automatic criterion for choosing the number of principal components in terms of predictive performances. Plugging in the estimator of Γ_0, we define estimated eigenfunctions and eigenvalues as solutions of:

$$\hat{\Gamma}_0 \hat{\xi} = \hat{\lambda} \hat{\xi}. \tag{C.2}$$

On a theoretical point of view, we would like to guarantee that population eigenfunctions can be consistently estimated by empirical eigenfunctions even in the non-iid framework of Functional Time Series. We refer to Theorem 16.2 in Horváth and Kokoszka 2012, which provides asymptotic arguments for such question.

The following subsections are dedicated to the estimation of eigenfunctions and eigenvalue in the two-dimensional functional case. Extending the work of Ramsay and Silverman 2005, we present two different estimation procedures, based respectively on a discretization of the functions to a fine grid and on a linear expansion of data on a finite set of basis functions. Among the two alternatives, we would resort to the function discretization. Indeed, such choice does not require the selection of a specific type of basis and not even the number of basis to employ, which are not trivial problem-dependent questions. Moreover, notice that also the discretization procedure can be seen as a particular case of the basis expansion, using as basis system indicator functions on the grid points. Furthermore, in the subsequent, C.3 we demonstrate with a simulation study that there is no significant evidence to prefer one method against the other in terms of estimation quality.

We want to stress the fact that our methodology for FPCA is general, it works for two-dimensional functional data regardless of the presence of temporary dependence between observations. Not modeling the serial dependence structure does not invalidate the PCA procedure, but we still have to require that the dynamic is stationary in order for the covariance estimation to make sense and thus to provide meaningful estimates.

C.1. Grid discretization

Consider a grid discretization $\{u_i\}_{i=1,\dots,N_1}$ of $[c,d]$ and $\{v_j\}_{j=1,\dots,N_2}$ of $[e,f]$, let $\omega_1 = \frac{1}{N_1}$, $\omega_2 = \frac{1}{N_2}$. For any point (u_i, v_j) of the discretized grid, the lhs of the functional eigenequation (C.2) can be rewritten as:

$$\hat{\Gamma}_0\hat{\xi}(u_i,v_j) = \int_c^d \int_e^f \hat{\gamma}_0(u_i,v_j;w,z)\hat{\xi}(w,z)dwdz \approx \tag{C.3}$$

$$\approx \omega_1 \sum_{l=1}^{N_1} \int_e^f \hat{\gamma}_0(u_i,v_j;u_l,z)\hat{\xi}(u_l,z)dz \approx \tag{C.4}$$

$$\approx \omega_1\omega_2 \sum_{l=1}^{N_1}\sum_{m=1}^{N_2} \hat{\gamma}_0(u_i,v_j;u_l,v_m)\hat{\xi}(u_l,v_m). \tag{C.5}$$

By defining $N := N_1N_2$ and introducing a bijection $\zeta : \{1,\dots,N_1\} \times \{1,\dots,N_2\} \to \{1,\dots,N\}$, we can vectorize the two-dimensional grid. Therefore, we can store observed data into a bidimensional matrix, and proceed with a usual multivariate analysis. Let $\mathbb{Y} \in \mathbb{R}^{T\times N}$ be defined as $\mathbb{Y}[t,\zeta(i,j)] = y_t(u_i,v_j)$. We hence introduce the estimated variance-covariance matrix of the just defined multivariate dataset: $\hat{\mathbf{\Gamma}}_0 \in \mathbb{R}^{N\times N}$, $\hat{\mathbf{\Gamma}}_0 = \frac{1}{T}\mathbb{Y}^T\mathbb{Y}$. Notice that $\hat{\mathbf{\Gamma}}_0[\zeta(i,j),\zeta(l,m)] = \hat{\boldsymbol{\gamma}}_0(u_i,v_j;u_l;v_m)$. Let also $\hat{\boldsymbol{\xi}} \in \mathbb{R}^N$, $\boldsymbol{\xi}[\zeta(i,j)] = \xi(u_i,v_j)$. The eigenequation can thus be rewritten in the following matricial form:

$$\omega_1\omega_2\hat{\mathbf{\Gamma}}_0\hat{\boldsymbol{\xi}} = \hat{\lambda}\hat{\boldsymbol{\xi}}. \tag{C.6}$$

It is then straightforward to find the eigenvalues ρ and eigenvectors $\boldsymbol{\theta}$ of the matrix $\mathbf{\Gamma}_0$ and to derive $\hat{\lambda} = \omega_1\omega_2\rho$ and $\hat{\boldsymbol{\xi}} = \omega_1^{-1/2}\omega_2^{-1/2}\boldsymbol{\theta}$. Finally, to obtain an approximate eigenfunction $\hat{\xi}$ from discrete values $\hat{\boldsymbol{\xi}}$, we can use any convenient interpolation method.

C.2. Basis expansion

Let $\{g_i\}_{i\in\mathbb{N}}$ be a basis system for $\mathcal{L}^2([c,d])$ and $\{h_j\}_{j\in\mathbb{N}}$ a basis system for $\mathcal{L}^2([e,f])$. Consider now the tensor product basis $\{g_i \otimes h_j\}_{i,j}$, where $g_i \otimes h_j = g_ih_j$. Unfortunately, the space spanned by the tensor product basis is a proper subset of $\mathbb{H} = \mathcal{L}^2([c,d]\times[e,f])$, however, one could also prove that such subspace is *dense* in $\mathbb{H}$, thus arguing that the tensor product basis system is sufficient to model functions of $\mathbb{H}$. Therefore, we assume that each $x \in \mathbb{H}$, admits the decomposition:

$$x(u,v) = \sum_{i\in\mathbb{N}}\sum_{j\in\mathbb{N}} c_{i,j}g_i(u)h_j(v), \qquad c_{ij} = \langle x, g_i \otimes h_j\rangle. \tag{C.7}$$

Thanks to the existence of a bijection between $\mathbb{N}$ and $\mathbb{N}^2$ we can rearrange the terms of the basis system in order to obtain one depending on a single index instead of two, namely $\{\phi_k(u,v)\}_k$ instead of $\{g_i(u)\otimes h_j(v)\}_{i,j}$. We can thus rewrite (C.7) as:

$$x(u,v) = \sum_{k\in\mathbb{N}} c_k\phi_k(u,v). \tag{C.8}$$

In practical applications, one typically truncates the number of basis functions on the two univariate domains to K_1 and K_2 respectively, obtaining a total number of basis equal to $K := K_1 + K_2$. Let us now introduce the vector $\boldsymbol{\phi}(u,v) \in \mathbb{R}^K$, $\boldsymbol{\phi}(u,v) = [\phi_1(u,v),\dots,\phi_K(u,v)]^T$ and the matrix $\boldsymbol{C} \in \mathbb{R}^{T\times K}$ with elements $C[t,k] = c_{tk}$ containing the coefficients of basis projection of the random functions $Y_1,\dots,Y_T$, in such a way that:

$$Y_t(u,v) = \sum_{k=1}^{K} c_{tk}\phi_k(u,v) + \delta_t(u,v) \qquad t = \dots, T, \tag{C.9}$$

where $\delta_t(u,v)$ is a projection error, which is present due to the truncation of the basis system to the first K terms. In the remainder of this section, we will neglect the projection error and identify the observed functions with the ones reconstructed from the first M basis. Following the work of Ramsay and Silverman 2005 and exploiting representation (C.8), we aim to rephrase the eigenequation (C.2) in a matricial form. The estimated covariance function can be expressed in matrix terms:

$$\hat{\gamma}_0(u,v;w,z) = \frac{1}{T}\sum_{t=1}^{T} Y_t(u,v)Y_t(w,z) = \frac{1}{T}\sum_{t=1}^{T}\sum_{l,m=1}^{K} c_{il}\phi_l(u,v)c_{im}\phi_m(w,z) = \frac{1}{T}\boldsymbol{\phi}(u,v)^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\phi}(w,z).$$

Suppose now that an eigenfunction ξ admits the decomposition:

$$\xi(u,v) = \sum_{l=1}^{K} b_l \phi_l(u,v) + \kappa(u,v) = \boldsymbol{\phi}(u,v)^T \boldsymbol{b} + \kappa(u,v), \tag{C.10}$$

where $\boldsymbol{b} = [b_1, \ldots, b_K]^T = [\langle \xi, \phi_1 \rangle, \ldots, \langle \xi, \phi_K \rangle]^T$ contains coefficients of basis projection of ξ.

Neglecting once again the projection error κ, the goal becomes now to estimate the coefficients $\boldsymbol{b}$ and the corresponding eigenvalue λ for each eigenfunction ξ_j, for $j = \ldots, M$. Let's introduce finally $\boldsymbol{W} \in \mathbb{R}^{K \times K}$, defined as $\boldsymbol{W} := \int_c^d \int_e^f \boldsymbol{\phi}(u,v) \boldsymbol{\phi}(u,v)^T du dv$, notice that the tensor product basis is composed by two orthonormal basis systems, the resulting tensor product basis system is itself orthonormal and thus $\boldsymbol{W} = \boldsymbol{I}$, where $\boldsymbol{I}$ denotes the diagonal matrix. The lhs of (C.2) can be rewritten as:

$$\hat{\Gamma}_0 \hat{\xi}(u,v) = \int_c^d \int_e^f \frac{1}{T} \boldsymbol{\phi}(u,v)^T \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{\phi}(w,z) \boldsymbol{\phi}(w,z)^T \hat{\boldsymbol{b}} dw dz = \tag{C.11}$$

$$= \frac{1}{T} \boldsymbol{\phi}(u,v)^T \boldsymbol{C}^T \boldsymbol{C} \left(\int_c^d \int_e^f \boldsymbol{\phi}(w,z) \boldsymbol{\phi}(w,z)^T dw dz \right) \hat{\boldsymbol{b}} = \tag{C.12}$$

$$= \frac{1}{T} \boldsymbol{\phi}(u,v)^T \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{W} \hat{\boldsymbol{b}}. \tag{C.13}$$

The eigenequation thus becomes:

$$\frac{1}{T} \boldsymbol{\phi}(u,v)^T \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{W} \hat{\boldsymbol{b}} = \lambda \boldsymbol{\phi}(u,v)^T \hat{\boldsymbol{b}} \qquad \forall u, v, \tag{C.14}$$

which implies:

$$\frac{1}{T} \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{W} \hat{\boldsymbol{b}} = \lambda \hat{\boldsymbol{b}}. \tag{C.15}$$

We can hence derive the eigenvectors $\hat{\boldsymbol{b}}_j$ of $\frac{1}{T} \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{W}$ and the corresponding eigenvalues and finally reconstruct the eigenfunctions $\hat{\xi}_j$ thanks to (C.10).

C.3. Comparison of estimation methods

In this section, we aim to compare the two proposed approach for performing functional principal component analysis, namely the basis expansion and the discretization approach. Without loss of generality, we settle the study in $\mathcal{L}^2([0,1] \times [0,1])$.

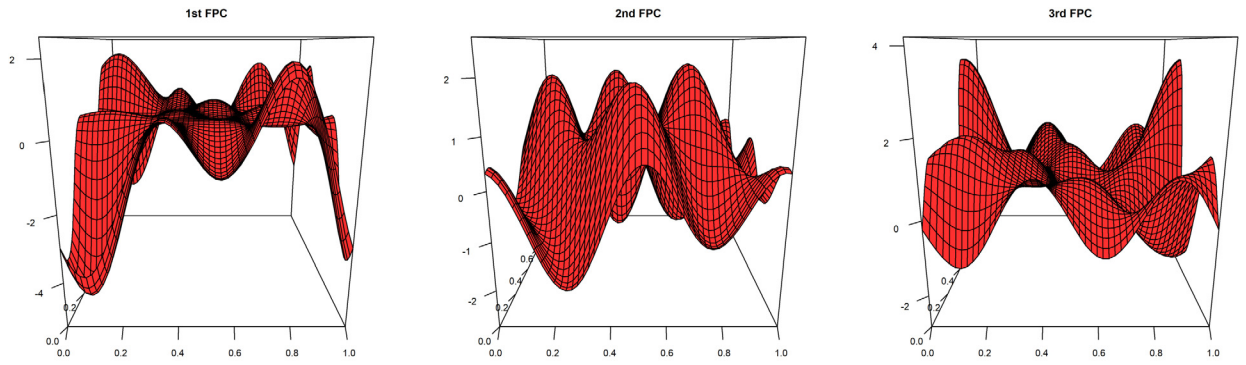
In general, given a finite basis system $\{\phi_k\}_{k=1,\ldots,K}$, one can represent a Functional Time Series $\{Y_t\}_{t=1}^T$ by means of representation (C.9), that we report here for ease of reference:

$$Y_t(u,v) = \sum_{k=1}^{K} c_{tk} \phi_k(u,v) + \delta_t(u,v) \qquad t = 1, \ldots, T. \tag{C.16}$$

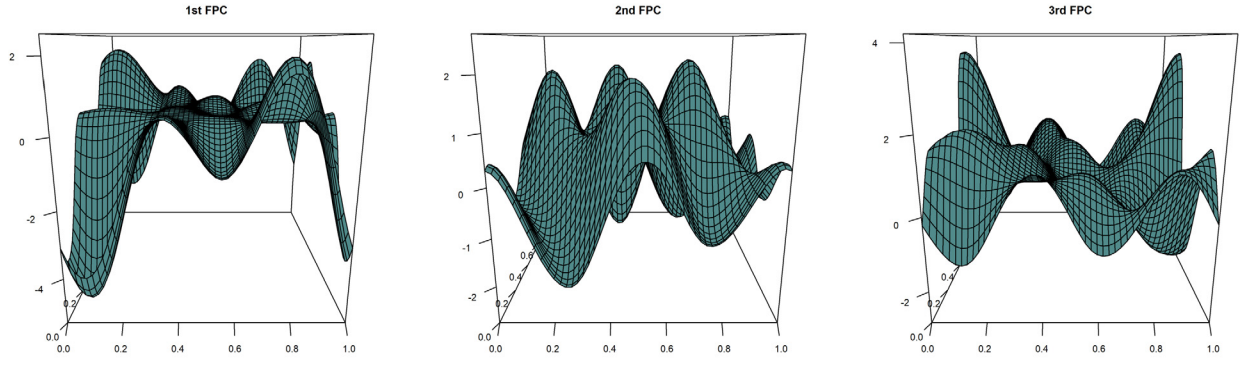
Starting from this decomposition, we have derived in Appendix C.2 an estimator of the functional principal components ξ_j, which, however, neglects the contribution of the error δ_t. For such reason, the choice of the basis system $\{\phi_k\}_{k=1}^K$ is crucial, and if a meaningful option is not available, the approximation residual δ_t would be large and this procedure would inevitably provide biased estimates of ξ_j. On the other hand, the FPCA approach based on data discretization provides good result as long as the sampling grid is sufficiently dense.

In order to prove such thesis, we simulate a time series of functions $\{Y_t\}_{t=1}^T \subset \text{span}\{\phi_1, \ldots, \phi_K\}$, from a non-concurrent FAR(1) process with Gaussian errors. Data are simulated based on a basis expansion on $\{\phi_1, \ldots, \phi_K\}$, which is constructed as the tensor product basis of two Fourier basis systems $\{g_i\}_{i=1,\ldots,K_1}$, $\{h_i\}_{i=1,\ldots,K_2}$ both defined on $[0,1]$. The sample size is chosen equal to $T = 50$, the number of basis in each of the one-dimensional systems is selected equal to 5, in order to have a total number of basis $K = 25$.

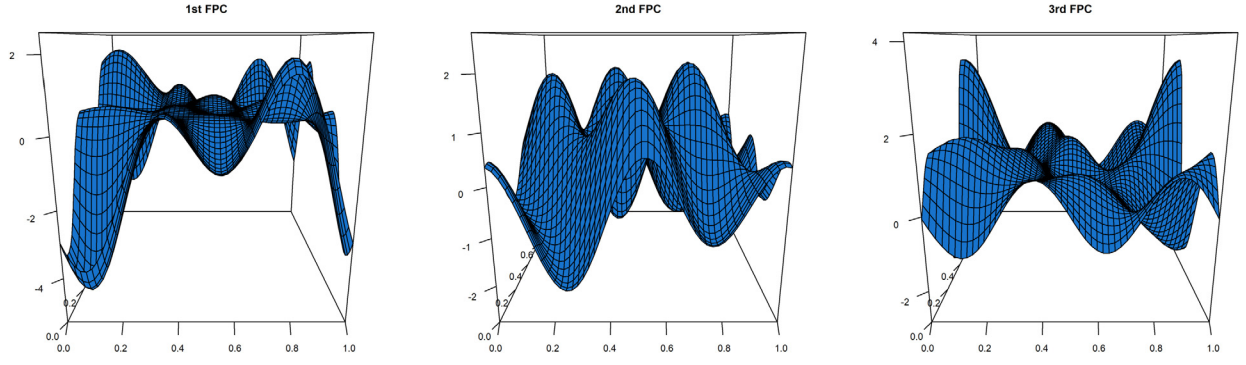
Since we know the space where the functions are embedded, we can apply the estimation procedure in Appendix C.2 using as basis system the same one used in the simulation. Notice that in this case the approximation error δ_t in (C.16) is exactly zero and one can derive optimal estimates of the functional principal components. Estimators of the first three functional principal components are represented in Fig. C.15a. We repeat the same estimation procedure, this time modeling functions with a basis built from the tensor product of two one-dimensional cubic B-Spline basis systems with 5 bases



(a) FPCA by basis expansion on the same basis system used for the simulation.



(b) FPCA by basis expansion on a different basis system.



(c) FPCA by grid discretization.

Fig. C.15. First three functional principal components, estimated with three different methods. As usual in PCA, EFPC's are unique up to a constant. For such reason, in order to compare the different approaches, estimated harmonics in the second and third row are each rescaled by means of the mean difference ratio with the ones in the first row.

each. In this case, the basis system do not coincide with the one from which functions are simulated. Nevertheless, as reported in Fig. C.15b all the scaled eigenfunctions are very close to the optimal ones estimated before. Finally, we compare the aforementioned estimators with the ones coming from the discretization approach. Each of the one-dimensional grids is discretized using a step size equal to 0.02, thus resulting in a total of 2500 points. Also in this case (see Fig. C.15c), estimated harmonics are very close to the optimal ones in Fig. C.15a.

To enable for better comparison, we report in Table C.1 the Mean Squared Error (MSE) (C.17) between the FPC's y estimated using full knowledge of the basis system from which functions are simulated and the ones estimated using other techniques ($\hat{y}$). Such quantity is computed starting from values of y and $\hat{y}$ on a two-dimensional grid $\{(u_i, v_j)\}_{i=1,\dots,N_1; j=1,\dots,N_2}$.

Table C.1
MSE between estimated FPC's on the basis system used for the simulation and FPC's estimated using another basis system or the discretization approach.

FPCA method	MSE		
	1st FPC	2nd FPC	3rd FPC
Basis on B-Spline	$2.62 \cdot 10^{-3}$	$4.45 \cdot 10^{-3}$	$2.28 \cdot 10^{-3}$
Discretization	$7.69 \cdot 10^{-4}$	$2.45 \cdot 10^{-3}$	$1.38 \cdot 10^{-3}$

$$MSE(y, \hat{y}) = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left(y(u_i, v_j) - \hat{y}(u_i, v_j)\right)^2. \tag{C.17}$$

We can appreciate very low values of MSE, regardless of the technique used for FPCA, thus suggesting that both the basis expansion and the discretization approach are valid options on a practical point of view.

References

Antoniadis, A., Paparoditis, E., Sapatinas, T., 2006. A functional wavelet-kernel approach for time series prediction. J. R. Stat. Soc. B 68, 837–857. https://doi.org/10.1111/j.1467-9868.2006.00569.x.

Aue, A., van Delft, A., 2020. Testing for stationarity of functional time series in the frequency domain. Ann. Stat. 48. https://doi.org/10.1214/19-aos1895.

Aue, A., Norinho, D., Hörmann, S., 2012. On the prediction of stationary functional time series. J. Am. Stat. Assoc. 110. https://doi.org/10.1080/01621459.2014.909317.

Avsar, N.B., Jin, S., Kutoglu, H., Gurbuz, G., 2016. Sea level change along the Black Sea coast from satellite altimetry, tide gauge and GPS observations. Geod. Geodyn. 7, 50–55. https://doi.org/10.1016/j.geog.2016.03.005. special issue: GNSS Applications in Geodesy and Geodynamics.

Balasubramanian, S., Karrh, J., Patwardhan, H., 2006. Audience response to product placements: an integrative framework and future research agenda. J. Advert. 35, 115–141. https://doi.org/10.2753/JOA0091-3367350308.

Bosq, D., 2000. Linear Processes in Function Spaces. Springer, New York. https://link.springer.com/book/10.1007/978-1-4612-1154-9.

Cazenave, A., Cabanes, C., Dominh, K., Mangiarotti, S., 2001. Recent sea level change in the Mediterranean Sea revealed by Topex/Poseidon satellite altimetry. Geophys. Res. Lett. 28, 1607–1610. https://doi.org/10.1029/2000GL012628.

Chernozhukov, V., Wüthrich, K., Yinchu, Z., 2018. Exact and robust conformal inference methods for predictive machine learning with dependent data. In: Bubeck, S., Perchet, V., Rigollet, P. (Eds.), Proceedings of the 31st Conference on Learning Theory, PMLR, pp. 732–749. https://proceedings.mlr.press/v75/chernozhukov18a.html.

Chiou, J.M., Yang, Y.F., Chen, Y.T., 2016. Multivariate functional linear regression and prediction. J. Multivar. Anal. 146, 301–312.

Didericksen, D., Kokoszka, P., Zhang, X., 2010. Empirical properties of forecasts with the functional autoregressive model. Comput. Stat. 27, 285–298. https://doi.org/10.1007/s00180-011-0256-2.

Diquigiovanni, J., Fontana, M., Vantini, S., 2021. Distribution-free prediction bands for multivariate functional time series: an application to the Italian gas market. arXiv:2107.00527.

Diquigiovanni, J., Fontana, M., Vantini, S., 2022. The importance of being a band: finite-sample exact distribution-free prediction sets for functional data. Stat. Sin. https://doi.org/10.5705/ss.202022.0087. https://www3.stat.sinica.edu.tw/ss_newpaper/SS-2022-0087_na.pdf.

Fontana, M., Zeni, G., Vantini, S., 2023. Conformal prediction: a unified review of theory and new challenges. Bernoulli 29. https://doi.org/10.3150/21-BEJ1447.

Gammerman, A., Vovk, V., Vapnik, V., 1998. Learning by transduction. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 148–155.

Gervini, D., 2010. Retracted: the functional singular value decomposition for bivariate stochastic processes. Comput. Stat. Data Anal. 54, 163–172. https://doi.org/10.1016/j.csda.2009.07.024.

Ginzburg, A.I., Kostianoy, A.G., Sheremet, N.A., Lebedev, S.A., 2011. Satellite altimetry applications in the Black Sea. In: Vignudelli, S., Kostianoy, A.G., Cipollini, P., Benveniste, J. (Eds.), Coastal Altimetry. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 367–387.

Hernández, N., Cugliari, J., Jacques, J., 2021. Simultaneous predictive bands for functional time series using minimum entropy sets. arXiv:2105.13627.

Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications. Springer Series in Statistics. Springer, New York. https://books.google.it/books?id=OVezLB__ZpYC.

Horváth, L., Kokoszka, P., Rice, G., 2014. Testing stationarity of functional time series. J. Econom. 179, 66–82. https://doi.org/10.1016/j.jeconom.2013.11.002.

Hyndman, R., Shang, H.L., 2009. Functional time series forecasting. J. Korean Stat. Soc. 38. https://doi.org/10.1016/j.jkss.2009.06.002.

Hörmann, S., Kokoszka, P., 2010. Weakly dependent functional data. Ann. Stat. 38, 1845–1884. https://doi.org/10.1214/09-AOS768.

Ivanescu, Andrada, 2013. A note on bivariate smoothing for two-dimensional functional data. Int. J. Stat. Probab. 2. https://doi.org/10.5539/ijsp.v2n2p102.

Jiao, S., Aue, A., Ombao, H., 2021. Functional time series prediction under partial observation of the future curve. J. Am. Stat. Assoc., 1–12. https://doi.org/10.1080/01621459.2021.1929248.

Kargin, V., Onatski, A., 2005. Curve forecasting by functional autoregression. J. Multivar. Anal. 99, 2508–2526. https://doi.org/10.1016/j.jmva.2008.03.001.

Kath, C., Ziel, F., 2021. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. Int. J. Forecast. 37, 777–799. https://doi.org/10.1016/j.ijforecast.2020.09.006.

Lei, J., Rinaldo, A., Wasserman, L., 2015. A conformal prediction approach to explore functional data. Ann. Math. Artif. Intell. 74, 29–43. https://doi.org/10.1007/s10472-013-9366-6.

López-Pintado, S., Romo, J., 2009. On the concept of depth for functional data. J. Am. Stat. Assoc. 104, 718–734. https://doi.org/10.1198/jasa.2009.0108.

Mertz, F., Legeais, J.F., 2018. Sea level daily gridded data from satellite observations for the Black Sea from 1993 to 2020. https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-sea-level-black-sea?tab=overview.

Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A., 2002. Inductive confidence machines for regression. In: Elomaa, T., Mannila, H., Toivonen, H. (Eds.), Machine Learning: ECML 2002. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 345–356.

Porro-Muñoz, D., Mata, F.J.S., Hernández, N., Bustamante, I.T., 2014. Functional data analysis as an alternative for the automatic biometric image recognition: iris application. Comput. Sist. 18, 111–121. https://doi.org/10.13053/CyS-18-1-2014-022.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rakêt, L.L., 2010. 2D Functional Data Analysis, with applications to image analysis. Master's thesis. Statistics Department of Mathematical Sciences, Faculty of Science, University of Copenhagen.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis. Springer Series in Statistics. Springer. https://books.google.it/books?id=mU3dop5wY_4C.

Rossini, J., Canale, A., 2018. Quantifying prediction uncertainty for functional-and-scalar to functional autoregressive models under shape constraints. J. Multivar. Anal. 170, 221–231. https://doi.org/10.1016/j.jmva.2018.10.007.

Solari, A., Djordjilović, V., 2021. Multi split conformal prediction. arXiv:2103.00627.

Tsimplis, M.N., Baker, T.F., 2000. Sea level drop in the Mediterranean Sea: an indicator of deep water salinity and temperature changes? Geophys. Res. Lett. 27, 1731–1734. https://doi.org/10.1029/1999GL007004.

Vovk, V., Gammerman, A., Shafer, G., 2022. Algorithmic Learning in a Random World, second ed. Springer, New York. https://link.springer.com/book/10.1007/b106715.

Wisniewski, W., Lindsay, D., Lindsay, S., 2020. Application of conformal prediction interval estimations to market makers' net positions. In: Gammerman, A., Vovk, V., Luo, Z., Smirnov, E., Cherubin, G. (Eds.), Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications, PMLR, pp. 285–301. https://proceedings.mlr.press/v128/wisniewski20a.html.

Zhou, L., Pan, H., 2014. Principal component analysis of two-dimensional functional data. J. Comput. Graph. Stat. 23, 779–801. https://doi.org/10.1080/10618600.2013.827986.