Variable selection for the prediction of C[0,1]-valued AR processes using RKHS

Beatriz Bueno-Larraz[*][†] Johannes Klepsch[‡]

October 2017

Abstract

A model for the prediction of functional time series is introduced, where observations are assumed to be realizations of a C[0,1]-valued process. We model the dependence of the data with a non-standard autoregressive structure, motivated in terms of the Reproducing Kernel Hilbert Space (RKHS) generated by the covariance kernel of the data. The general definition has as particular case a set of finite-dimensional models based on marginal variables of the process. Thus, this approach is especially useful to find relevant points for prediction (sometimes called "impact points"). Some examples show that this model has a good amount of generality. In addition, problems like the non-invertibility of the covariance operators in function spaces can be circumvented using this methodology. A simulation study and two real data examples are presented to evaluate the performance of the proposed predictors.

Keywords: Autoregressive (AR); Continuous functions; Functional data analysis (FDA); Functional linear process; Prediction; Variable selection; RKHS

MSC 2010: Primary: 62M10, 62M15, 62M20; Secondary: 62H25, 60G25

1 Introduction

Functional data analysis (FDA) is one of the answers to the recent rise of complex data. It consists in viewing observations as entire curves instead of individual

[*] Universidad Autónoma de Madrid, Departamento de Matemáticas, Spain, email: beatriz.bueno@uam.es & beatriz.bueno.larraz@gmail.com

[†] Corresponding author

[‡] Center for Mathematical Sciences, Technische Universität München, 85748 Garching, Boltzmannstraße 3, Germany, email: j.j.klepsch@gmail.com

data points. In many applications the observed curves are recorded sequentially in time. Therefore, often there are reasons to reject the assumption of independence among the curves. For a recent overview of the literature of FDA we refer to Cuevas (2014). Furthermore, Bosq (2000) gives a good introduction into the field of linear processes in function spaces and introduces functional autoregressive processes in depth.

In the present paper we introduce a predictor based on a functional autoregressive (AR) model,

$$X_n(s) = \rho X_{n-1}(s) + \varepsilon_n(s), \quad s \in [0,1],$$

which will be especially suitable for variable selection purposes. It is assumed many times that $L^2[0,1]$, the space of square-integrable functions on $[0,1]$, is the function space in which this time series $(X_n)_{n\in\mathbb{Z}}$ takes its values. This is a sensible choice, as $L^2[0,1]$ is a separable Hilbert space and offers desirable geometric properties through the definition of the natural scalar product. However, considering our variable selection purpose, the main drawback of $L^2[0,1]$ is that, strictly speaking, it consists of equivalence classes of functions. That is, two functions represent the same L^2-function if the set where they differ has measure zero. In applications this is usually not an issue but, for any particular point $s \in [0,1]$, the value $f(s)$ is not well-defined for f an L^2-function.

In order to carry out the variable selection, pointwise definitions are required. Therefore, the space of continuous functions on $[0,1]$, $C[0,1]$, which is a Banach space with the supremum-norm, is a more natural space to work in. In addition, the subsequent change of norm allows us to obtain uniform convergence results. The problem of estimating ρ in functional AR models in $C[0,1]$ has usually been solved by projecting onto a finite dimensional subspace of $C[0,1]$, spanned by some eigenfunctions of the covariance operator of $(X_n)_{n\in\mathbb{Z}}$. For instance, Pumo (1998) presented a direct extension of the methodology derived in Bosq (2000). Therefore, his method inherits the problems of this methodology, stemming from the non-invertibility of the covariance operators in infinite-dimensional spaces. Some limitations of this principal component approach for Hilbert space–valued processes have been discussed extensively in the literature. By this way, as pointed out in Kargin and Onatski (2008) and Hörmann et al. (2015), the resulting space is shown to be optimal in order to represent the variability of the process, but the dependence might be lost by the dimension reduction. Likewise, Bernard (1997) indicates the sensitivity of the proposal to small errors in the estimation of small eigenvalues.

This paper presents a different approach that appears natural in the context of $C[0,1]$–valued processes. Here the projection on a finite dimensional space is replaced by the choice of the p most relevant points of $\{X_n(s), s \in [0,1]\}$ for the prediction of $\{X_{n+1}(s), s \in [0,1]\}$, under a suitable optimality criterion (in a similar sense as in Kargin and Onatski (2008) and Mokhtari and Mourid (2003)). In order to do this, we adapt the methodology introduced by Berrendero et al. (2017a) for the problem of scalar regression for independent data. Specifically, we define a new functional autoregressive model based on the Reproducing Kernel Hilbert Space (RKHS) generated by the covariance kernel of $(X_n)_{n\in\mathbb{Z}}$. It is shown that this new model falls into the class of Banach space–valued processes (ARB(q)) introduced in Bosq (2000), which directly gives us sufficient conditions for the existence of a unique stationary solution. Furthermore, the proposed model includes as particular cases the models in which the response function is given only by finitely many points. Whenever this is the case, we are able to prove, under some standard conditions, almost sure convergence of the estimated points and also of the estimated functions X_{n+1}, both uniformly and in $L^2[0,1]$. In this finite setting as well, our predictor coincides with the optimal one, in the sense that it is the best probabilistic predictor, as stated in Mokhtari and Mourid (2003). In addition, we develop a consistent estimator for the number p of relevant variables to select. Although we end up with a finite dimensional vector, the problem is still fully functional, since the definition of the optimality criterion is based on the whole process.

The advantages of predicting autoregressive processes with this new approach are numerous, besides the ones already mentioned. For instance, the use of this RKHS based model avoids the theoretical need of inverting the covariance operator since this is carried out, in some sense, by the inner product of the space. Additionally, the proposed method is flexible concerning the structure of the data: whether the data is observed on a grid or available as continuous functions - the methodology remains similar with slight technical differences. Nevertheless, for theoretical considerations the data is assumed to be given in a fully functional fashion. In addition, we will see that this technique is more computationally efficient. Concerning variable selection, its main advantage in comparison with other dimension reduction techniques is the interpretability in terms of the original data, which is usually desired in real data applications.

In order to show the practical relevance of the method, a simulation study is conducted. The new proposal is compared to the prediction methods of Aue et al. (2015) and Kokoszka and Reimherr (2013). To evaluate the performance in the real world, highway traffic volume and particle concentration data sets are studied.

In general, the literature in the field of functional time series analysis is developing quickly. Recent publications include time-domain methods like Hörmann and Kokoszka (2010), where a weak dependence concept is introduced, Aue et al. (2015), Klepsch and Klüppelberg (2017) and Klepsch et al. (2017), where prediction methodologies based on linear models are developed, and Aue and Klepsch (2017), where an estimator of functional linear processes based on moving average model fitting is derived. Besides, another examples of statistical papers taking advantage of the usefulness of Reproducing Kernel Hilbert Spaces are, among others, Hsing and Eubank (2015), Berrendero et al. (2017b), Berrendero et al. (2017a) and Kadri et al. (2015).

The paper is organised as follows: after introducing the notation and some background on RKHS theory, we define in Section 3 the new functional autoregressive model. Variable selection and the estimators are presented in Section 4, whose asymptotic properties are shown in Section 5. In this section the estimator for the number p of relevant variables is presented. Section 6 includes the experimental study along with some practical consideration for the implementation. Some proofs, which are mainly based on the theory developed in Berrendero et al.

(2017a), are included in Section 8.

2 Methodology

Before facing the problem described in the introduction, we give some notation and theoretical background addressing.

2.1 Notation

We will work with stochastic processes $X = X_n$, $n \in \mathbb{Z}$, taking values in the space of continuous functions over $[0, 1]$, $C[0, 1]$, whose marginal variables $X(s)$, $s \in [0, 1]$, are defined in a probability space $(\Omega, \mathcal{F}, P)$. For the sake of clarity in the equations, we will make the following abuse of notation: we will understand that we are working always with the centered process $X - \mathbb{E}X$, that will be denoted simply as X. We will denote as $\| \cdot \|$ the supremum norm in $C[0, 1]$. A standard assumption will be that the processes belongs to the space $L^2_C(\Omega)$ whose norm is given by $(\mathbb{E}\|X\|^2)^{1/2}$. In this case, each $X(s)$ will belong to $L^2_{\mathbb{R}}(\Omega)$, the space of square integrable random variables with norm $(\mathbb{E}|X(s)|^2)^{1/2}$. The norms in these spaces are denoted as $\| \cdot \|_{L^2_C(\Omega)}$ and $\| \cdot \|_{L^2_{\mathbb{R}}(\Omega)}$ respectively.

Given a stationary process X_n, we will define its lagged covariance function $c_n(s, t)$ for $s, t \in [0, 1]$, $n = 0, 1, \ldots$, as $\mathrm{Cov}(X_n(s), X_0(t))$.

In the sections devoted to variable selection, we will denote the vectors of points as $T_p \in [0, 1]^p$, where the subindex indicates the dimension. The covariance matrix of the marginal variables $X(t)$ indexed by T_p will be Σ_{T_p}. Moreover, for a general function $f : [0, 1] \to \mathbb{R}$, the evaluation $f(T_p)$ will be understood to be the column vector whose entries are $f(t_j)$, $t_j \in T_p$. Similarly for functions in several variables.

As usual in the literature, we will use a hat to denote the estimations derived from the samples. In addition, an asterisk will indicate the optimal quantities under some criterion.

2.2 Some background on Reproducing Kernel Hilbert Spaces

The prediction method proposed below is based on some properties of the Reproducing Kernel Hilbert Space (RKHS) associated with the covariance kernel of the process. Therefore, some basic background on these spaces is needed. We include here just a couple of basic definitions and properties that we use later on. We refer to Berlinet and Thomas-Agnan (2004) and Appendix F of Janson (1997) for a more in-depth theory.

Let $X(\cdot)$ be a centered stochastic process in $[0, 1]$ such that $X(s) \in L^2_{\mathbb{R}}(\Omega)$ for all $s \in [0, 1]$. We denote by $c_0(s, t)$ the covariance function (or covariance kernel) of this process. Denoting as $L^2[0, 1]$ the space of square integrable functions on $[0, 1]$, we can define the pre-Hilbert space

$$\mathcal{H}_0(X) := \{f \in L^2[0, 1] \ : \ f(\cdot) = \sum_{i=1}^{n} a_i c_0(t_i, \cdot), \ a_i \in \mathbb{R}, \ t_i \in [0, 1], \ n \in \mathbb{N}\} \quad (1)$$

with inner product $\langle f, g\rangle_{\mathcal{H}_0} = \sum_{i,j} \alpha_i \beta_j c_0(t_i, s_j)$, where $f(\cdot) = \sum_i \alpha_i c_0(t_i, \cdot)$ and $g(\cdot) = \sum_j \beta_j c_0(s_j, \cdot)$. It is not difficult to see that the representation of the elements of this space is unique whenever the covariance kernel is strictly positive definite. The RKHS associated with c_0, $\mathcal{H}(X)$, is the completion of this space. In other words, $\mathcal{H}(X)$ is the set of functions from $[0,1]$ to $\mathbb{R}$ that are pointwise limits of Cauchy sequences in $\mathcal{H}_0(X)$ (Berlinet and Thomas-Agnan (2004, p. 18)).

Reproducing Kernel Hilbert Spaces appear occasionally in the literature of functional data, since they are useful to impose smoothness conditions and help to reduce noise and irrelevant information. However we use RKHS with a different goal, since our interest lies in variable selection. Therefore, the so-called *reproducing property* of these spaces will be particularly useful. It states that, for all $f \in \mathcal{H}(X)$ and $s \in [0,1]$, $\langle f, c_0(s,\cdot)\rangle_{\mathcal{H}} = f(s)$. That is, the inner product with the covariance kernel behaves in some sense like Dirac's delta. We are interested in selecting variables on the trajectories drawn from the process X. However, in general, the realizations of the process do not belong to $\mathcal{H}(X)$ with probability one (for instance, Theorem 11 of Pillai et al. (2007)). Thus, this reproducing property can not be applied directly taking the trajectories as the function f in this last equation.

In order to circumvent this problem, we will use another Hilbert space which is also closely related to the process. We can derive a similar definition of a pre-Hilbert space using the marginal variables of the process $X(s)$, instead of the kernel functions $c_0(s,\cdot)$,

$$\mathcal{L}_0(X) \;=\; \{U \in L^2_{\mathbb{R}}(\Omega) \;:\; U = \sum_{i=1}^{n} a_i X(t_i),\; a_i \in \mathbb{R},\; t_i \in [0,1],\; n \in \mathbb{N}\}$$

with the same inner product as $L^2_{\mathbb{R}}(\Omega)$. We denote the closure of this space as $\mathcal{L}(X) \subset L^2_{\mathbb{R}}(\Omega)$.

Remark 1. *By definition of both spaces, the finite sums* $\sum_{i=1}^{n} a_i c_0(t_i,\cdot)$ *are dense in* $\mathcal{H}(X)$ *and the finite sums* $\sum_{i=1}^{n} a_i X(t_i)$ *are dense in* $\mathcal{L}(X)$*, with their corresponding norms.*

These two spaces $\mathcal{L}(X)$ and $\mathcal{H}(X)$ can be connected using the following congruence (bijective transformation preserving the inner product), named *Loève's isometry* (Berlinet and Thomas-Agnan (2004, Theorem 35) and Lukić and Beder (2001, Lemma 1.1))

$$\begin{array}{rccl} \Psi_X : & \mathcal{L}(X) & \to & \mathcal{H}(X) \\ & U & \mapsto & \mathbb{E}[UX(\cdot)] \;=\; \int_\Omega U(\omega)(X(\omega))(\cdot)\mathrm{d}P(\omega). \end{array} \tag{2}$$

That is, $\mathcal{L}(X)$ and $\mathcal{H}(X)$ are isometric spaces. If the random variable U is an element of $\mathcal{L}_0(X)$, i.e. $U = \sum_{i=1}^{n} a_i X(t_i)$, its image by the isometry is given by

$$\Psi_X(U) \;=\; \mathbb{E}\left[X(\cdot)\sum_{i=1}^{n} a_i X(t_i)\right] \;=\; \sum_{i=1}^{n} a_i \mathbb{E}[X(\cdot)X(t_i)] \;=\; \sum_{i=1}^{n} a_i c_0(t_i,\cdot)$$

That is, using the inverse of Ψ_X we can recover the Dirac's delta behaviour that we had before in $\mathcal{H}(X)$ with the reproducing property. In the following sections we will see how we can use this isometry to perform variable selection in AR processes.

3 Model definition

For a time series $(X_n)_{n\in\mathbb{Z}}$ taking values in some function space, the standard autoregressive model is of the form (see Chapter 6 Bosq (2000))

$$X_n = \rho X_{n-1} + \varepsilon_n, \qquad n \in \mathbb{Z}, \tag{3}$$

for some bounded linear operator ρ and some white noise process $(\varepsilon_n)_{n\in\mathbb{Z}}$ taking values in the same function space. In this section we propose a different functional autoregressive model "customized" to give a theoretical framework for variable selection. We also expose sufficient conditions for the model to have a unique stationary solution.

We will work with $(X_n)_{n\in\mathbb{Z}} \in L^2_C(\Omega)$ a centered stationary process taking values in $C[0,1]$. In this context we propose to replace (3) with

$$X_n(s) = \langle \phi(s,\cdot), X_{n-1}(\cdot)\rangle_{\mathcal{H}} + \varepsilon_n(s), \qquad n \in \mathbb{Z}, \tag{4}$$

where $\langle\ ,\ \rangle_{\mathcal{H}}$ denotes the scalar product of the RKHS generated by the covariance kernel of $(X_n)_{n\in\mathbb{Z}}$ and for some appropriate kernel ϕ such that $\phi(s,\cdot) \in \mathcal{H}(X)$ for all $s \in [0,1]$. Note that, since the process is stationary, its covariance structure remains invariant and then the space $\mathcal{H}(X)$, whose inner product is used, does not depend on n. As mentioned above, the trajectories of the process do not belong to $\mathcal{H}(X)$, thus the inner product of the previous equation has to be appropriately interpreted. As suggested in Parzen (1961), we will understand $\langle \phi(s,\cdot), X_{n-1}(\cdot)\rangle_{\mathcal{H}}$ as $\Psi^{-1}_{X_{n-1}}(\phi(s,\cdot))$, where $\Psi_{X_{n-1}}$ denotes the Loève's isometry defined in Equation (2). Aiming for clarity, we still write $\langle \phi(s,\cdot), X_{n-1}(\cdot)\rangle_{\mathcal{H}}$ when useful.

Equation (4) is simply the pointwise definition of a fully functional model, for each $s \in [0,1]$. This definition can be applied for any process $X_n \in L^2_C(\Omega)$ for $n \in \mathbb{Z}$, since then $X_n(s) \in L^2_{\mathbb{R}}(\Omega)$ and the Loève's isometry can be applied. Moreover, using the definition of $\Psi_{X_{n-1}}$,

$$\phi(s,\cdot) \;=\; \Psi_{X_{n-1}}\left(X_n(s) - \varepsilon_n(s)\right) \;=\; \mathbb{E}\left[\left(X_n(s) - \varepsilon_n(s)\right) X_{n-1}(\cdot)\right] \;=\; c_1(s,\cdot), \tag{5}$$

and $\|c_1(s,\cdot)\|_{\mathcal{H}} = \|X_n(s) - \varepsilon_n(s)\|_{L^2_{\mathbb{R}}(\Omega)} < \infty$. That is, the pointwise evaluations $X_n(s)$ of the process given in Equation (4) can be written as $\Psi^{-1}_{X_{n-1}}(c_1(s,\cdot)) + \varepsilon_n(s)$, which is always well-defined. However, it has to be carefully analysed whether or not the model (4) can be understood as a fully functional model in $L^2_C(\Omega)$. From this last equation we also see that, when changing the working space from $L^2[0,1]$ to $\mathcal{H}(X)$, the solution of the model does not require to invert the covariance operator. It could be understood as if the "inversion" was intrinsically carried out by the inner product of $\mathcal{H}(X)$.

In view of the previous discussion, we propose the following general definition.

Definition 1. *A sequence $X_n \in L^2_{C[0,1]}$, $n \in \mathbb{Z}$, with strictly positive covariance kernel is called Functional Continuous Autoregressive process of order 1 (FCAR(1)) if it is stationary and such that*

$$X_n(\cdot) \;=\; \langle \phi(\cdot,\sim), X_{n-1}(\sim)\rangle_{\mathcal{H}} + \varepsilon_n(\cdot) \;\equiv\; \Psi^{-1}_{X_{n-1}}(\phi(\cdot,\sim)) + \varepsilon_n(\cdot), \qquad n \in \mathbb{Z} \tag{6}$$

where $\phi(s,\cdot) \in \mathcal{H}(X)$ for every $s \in [0,1]$ and where $(\varepsilon_n)_{n\in\mathbb{Z}}$ is a strong $C[0,1]$-white noise independent of X_n.

Note that the assumption of having a strictly positive covariance kernel is equivalent to avoid processes such that the variance of some marginal variable is zero. In order to make sense of this functional definition and to be able to obtain some properties about the process, we will show that this model is in fact part of a more general family. As already mentioned, pointwise for $s \in [0,1]$ the model above is equal to $\Psi^{-1}_{X_{n-1}}(c_1(s,\cdot)) + \varepsilon_n(s)$, where the function c_1 is given by the sequence $\{X_n\}_{n\in\mathbb{Z}}$. Thus, leaving aside convergence issues, we could understand that X_n is given by $\rho X_{n-1} + \varepsilon$, where ρ is an operator on $C[0,1]$ depending on c_1 and such that it coincides with $\Psi^{-1}_{X_{n-1}}(c_1(s,\cdot))$ pointwisely. Let us see that this interpretation is well founded.

Proposition 1. *The FCAR(1) of Defintion 1 is in the model class of Banach space valued autoregressive ARB(1) processes introduced in Definition 6.1 of Bosq (2000) with $B = C[0,1]$ with equality in $L^2_C(\Omega)$, whenever its covariance function is strictly positive.*

Proof. We have to show that the series X_n given in Equation (6) can be written as $\rho(X_{n-1}) + \varepsilon_n$ with ρ a bounded linear operator acting on $C[0,1]$ and this equality holds in $L^2_C(\Omega)$. It means that $\mathbb{E}\|(X_n - \varepsilon_n) - \rho X_{n-1}\|^2$ is equal to zero.

We start showing that the pointwise definition of ρ, given just before the statement, can be extended to a functional model in $L^2_C(\Omega)$. We will do it by relying on the finite dimensional representation of the elements in $\mathcal{H}_0(X)$. We know that $\phi(s,\cdot)$ is in $\mathcal{H}(X)$ for every $s \in [0,1]$ by Equation (5). Hence $\phi(s,\cdot)$ can be approximated arbitrarily well by a finite linear combination of the form $\sum_{i=1}^{p} \alpha_i(s) c_0(t_i,\cdot)$, since these finite linear combinations are dense in $\mathcal{H}(X)$. More precisely, for every $s \in [0,1]$, the kernel $\phi(s,t)$ has a pointwise representation as $\lim_{k\to\infty} \phi_k(s,t)$, where

$$\phi_k(s,t) = \sum_{i=1}^{p_k} \alpha_i^k(s) c_0(t_i^k, t)$$

is a Cauchy sequence in $\mathcal{H}_0(X)$. By the Loève's isometry, $\Psi^{-1}_{X_{n-1}}(\phi_k(s,\cdot))$ is also a Cauchy sequence in $L^2_{\mathbb{R}}(\Omega)$ and then $\lim_{k\to\infty} \Psi^{-1}_{X_{n-1}}(\phi_k(s,\cdot)) \in L^2_{\mathbb{R}}(\Omega)$ for every $s \in [0,1]$. Moreover, since the inverse of this isometry is also an isometry, accordingly continuous in $\mathcal{H}(X)$, we can rewrite Equation (4) as

$$\begin{aligned} X_n(s) &= \Psi^{-1}_{X_{n-1}}\left(\lim_{k\to\infty} \phi_k(s,\cdot)\right) + \varepsilon_n(s) = \lim_{k\to\infty} \Psi^{-1}_{X_{n-1}}\left(\phi_k(s,\cdot)\right) + \varepsilon_n(s) \\ &= \lim_{k\to\infty} \sum_{i=1}^{p_k} \alpha_i^k(s) X_{n-1}(t_i^k) + \varepsilon_n(s). \end{aligned}$$

Thus, we want to define ρ as $\lim_{k\to\infty} \rho_k$, where $\rho_k(f)(s) = \sum_{i=1}^{p_k} \alpha_i^k(s) f(t_i^k)$, is a sequence of operators acting on $f \in C[0,1]$. Let us see now that this limit converges in $L^2_C(\Omega)$ for X_{n-1}. Note that,

$$\mathbb{E}\left\| \lim_{k\to\infty} \sum_{i=1}^{p_k} \alpha_i^k(\cdot) X_{n-1}(t_i^k)\right\|^2 = \mathbb{E}\Big[\sup_{s\in[0,1]} \big|\lim_{k\to\infty} \sum_{i=1}^{p_k} \alpha_i^k(s) X_{n-1}(t_i^k)\big|\Big]^2$$

$$\leq \mathbb{E}\|X_{n-1}\|^2 \big(\sup_{s\in[0,1]} \big| \lim_{k\to\infty} \sum_{i=1}^{p_k} \alpha_i^k(s) \big| \big)^2$$

$$= \mathbb{E}\|X_{n-1}\|^2 \| \lim_{k\to\infty} \sum_{i=1}^{p_k} \alpha_i^k(\cdot) \|^2. \tag{7}$$

We need to uniformly bound this last limit. Due to the continuity of the trajectories, c_1 is a continuous function on the compact space $[0,1]^2$. Thus, by Equation (5) we know that $\sup_{s,t\in[0,1]} |\phi(s,t)|$ is equal to $\sup_{s,t\in[0,1]} |c_1(s,t)| < \infty$. Also,

$$\|\phi(\cdot,t)\| \;=\; \| \lim_{k\to\infty} \sum_{i=1}^{p_k} \alpha_i^k(\cdot) c_0(t_i^k, t)\| \;\geq\; \| \lim_{k\to\infty} \sum_{i=1}^{p_k} \alpha_i^k(\cdot)\| \inf_{s,t\in[0,1]} c_0(s,t).$$

Assuming that $c_0(s,t) > 0$ for all $s,t \in [0,1]$, there exists a $\delta > 0$ such that $\inf_{s,t\in[0,1]} c_0(s,t) = \delta$, and then

$$\| \lim_{k\to\infty} \sum_{i=1}^{p_k} \alpha_i^k(\cdot)\| \leq \delta^{-1} \; \|\phi(\cdot,t)\| < \infty. \tag{8}$$

Now with (7), (8) and the fact that $X_{n-1} \in L^2_C(\Omega)$, we can conclude that the limit of $\sum_{i=1}^{p_k} \alpha_i^k(\cdot) X_{n-1}(t_i^k)$ converges in $L^2_C(\Omega)$ when $k \to \infty$.

It would remain to show that the limit operator $\rho = \lim_{k\to\infty} \rho_k$ is bounded. However it follows using a similar reasoning, for $f \in C[0,1]$

$$\begin{aligned} \sup_{\|f\|\leq 1} \|\rho(f)\| \;&\leq\; \big(\| \lim_{k\to\infty} \sum_{i=1}^{p_k} \alpha_i^k(\cdot)\|\big)^2 \;\leq\; \delta^{-1} \|\phi(t,\cdot)\| \\ &\leq\; \Big(\inf_{s,t\in[0,1]} c_0(s,t) \Big)^{-1} \sup_{s,t\in[0,1]} \phi(s,t) < \infty. \end{aligned}$$

□

The above proposition allows us to use results derived on ARB(1) processes. For instance, we immediately get the following corollary from Theorem 6.1 in Bosq (2000) concerning the existence of a unique stationary solution.

Corollary 1. *If there exists $j_0 \in \mathbb{N}$ such that $\|\rho^{j_0}\|_{\mathcal{L}} < 1$, where ρ is the operator that coincides pointwisely with $\Psi^{-1}_{X_{n-1}}(c_1(s,\cdot))$, then (6) has a unique strictly stationary solution given by*

$$X_n = \sum_{j=0}^{\infty} \rho^j \varepsilon_{n-j}, \quad n \in \mathbb{Z},$$

where the series converges in $L^2_C(\Omega)$.

The fact that we obtain convergence in $L^2_C[0,1]$ is intrinsically related to the definition of the model since, by the definition of the Loève's isometry, convergence in the RKHS $\mathcal{H}(X)$ is translated to convergence in $L^2_{\mathbb{R}}(\Omega)$. However, as we will clarify in next section, in some particular cases both the definition of the model and its solution converge with probability one. For convenience, we summarize the assumptions required for the existence of such a solution to (6).

Assumption 1. *$(X_n)_{n\in\mathbb{Z}}$ is a stationary sequence such that $X_n \in L^2_{C[0,1]}$ with strictly positive covariance kernel and there exists $j_0 \in \mathbb{N}$ such that $\|\rho^{j_0}\|_{\mathcal{L}} < 1$, where ρ is the operator that coincides pointwisely with $\Psi^{-1}_{X_{n-1}}(c_1(s,\cdot))$.*

An extension of model FCAR(1) defined in Equation (6) to FCAR(q) can be carried out whenever $X_n(s) = Z(s+n)$ for $s \in [0,1]$ and Z is a stationary process with continuous trajectories. In this case we can write for $Z_{n,q}(s) = Z(s+n-q+1)$, $s \in [0,q]$,

$$X_n(s) \;=\; \Psi_{Z_{n-1,q}}(\phi(s,\cdot))^{-1} + \varepsilon_n(s),\;\; s \in [0,1]$$

where now each $\phi(s,\cdot)$ belongs to the RKHS associated with $Z_{n-1,q}$ and q is the minimum for which this model holds. In this case, Equation (5) is

$$\phi(s,t) \;=\; \mathbb{E}[(X_n(s) - \varepsilon_n(s))Z_{n-1,q}(t)] \;=\; c_i(s,t) \;\text{ for } t \in (q-i, q-i+1].$$

All the theory presented in the paper remains valid in this case, with some additional assumptions, so for the sake of simplicity we will present here the case $q = 1$. Nevertheless, we will include some comments along the paper to clarify the changes due to this extension.

4 Variable Selection

The use of some dimension reduction technique is nearly mandatory when dealing with functional data problems, in order to circumvent issues like the non-invertibility of the covariance operators. We will consider here dimension reduction via variable selection. The goal of variable selection techniques when dealing with functional data is to find the p most relevant points $t_1, \ldots, t_p$ that best summarize the functions X_n, according to some optimality criterion. Whenever the number of selected points p is small, this type of dimension reduction is directly interpretable. In some real data applications this is an advantage compared with other methods based on projections, since variable selection techniques keep more in touch with the original data.

In order to perform variable selection, we need to restrict ourselves to a specific class of kernels $\phi(\cdot,\sim)$ in model (6). Specifically, we will approximate the real kernel of the model by kernels that depend only on a fixed number of points p as,

$$\phi(s,\cdot) \;=\; \sum_{j=1}^{p} \alpha_j(s) c_0(t_j,\cdot) \;\in\; \mathcal{H}(X), \tag{9}$$

for all $s \in [0,1]$. That is, all the evaluations of the kernel $\phi(s,\cdot)$ for different s depend on the same set of points $t_1, \ldots, t_p$. This restriction is not as strong as it may appear since, as pointed out in Remark 1, these finite linear combinations are dense in $\mathcal{H}(X)$. Therefore, any possible function $\phi(s,\cdot)$ in this space could be arbitrarily well approximated by just increasing the number of points p. However, in the practical examples tested in Section 6, we have seen that a small number p is usually enough to obtain a good approximation. In addition, as showed in Proposition 2 of Mokhtari and Mourid (2003), if the true kernel of the model is as in Equation (9), the best linear predictor based on the marginal variables $X_{n-1}(t_1), \ldots, X_{n-1}(t_p)$ is in fact the best probabilistic predictor of $(X_n - \varepsilon_n)$.

In view of the discussion of the previous section, a natural question is how the requirements of Corollary 1 are modified when the kernel belongs to this restricted family. By the expression of Loève's isometry for finite linear combinations of the covariance kernel, the model under (9) is given by

$$X_n(\cdot) \;=\; \sum_{j=1}^{p} \alpha_j(\cdot) X_{n-1}(t_j) + \varepsilon_n(\cdot). \tag{10}$$

Therefore, $X_n = \rho X_{n-1} + \varepsilon_n$ where, for $f \in C[0,1]$,

$$\rho(f)(\cdot) = \sum_{j=1}^{p} \alpha_j(\cdot) f(t_j).$$

Thus, the process would have a unique strictly stationary solution under (9) whenever $\|\rho^{j_0}\|_{\mathcal{L}} < 1$ for some $j_0 \in \mathbb{N}$. For instance, a sufficient condition for this to hold is that $\sum_{i=1}^{p} \|\alpha_i\| < 1$, which is an easily verifiable condition.

Remark 2. *Note that now the sums are finite, so the limits taken in the proof of Proposition 1 are not needed anymore. Therefore, in this case the model holds with probability one and the series of Corollary 1 that defines the unique strictly stationary solution also converges almost surely.*

Regarding model $FCAR(q)$ with $q > 1$, the working spaces change. Now $\mathcal{L}(Z_{n,q})$ is the space generated by the marginal variables $X_n(s), \ldots, X_{n-q}(s)$ and $\mathcal{H}(Z_{n,q})$ is the RKHS associated with this delayed process, whose reproducing kernel $c_0(s,t)$ for $s,t \in [0,q]$ is the covariance function of $Z_{n,q}$. Then if we have a sparse function as in (9), we can split it as

$$\phi(s,\cdot) \;=\; \sum_{j=1}^{p^{(1)}} \alpha_j^{(1)}(s) c_0(t_j^{(1)},\cdot) + \ldots + \sum_{j=1}^{p^{(q)}} \alpha_j^{(q)}(s) c_0(t_j^{(q)},\cdot),$$

where $t_j^{(i)} \in (q-i, q-i+1]$. Hence Equation (10) is rewritten as

$$\begin{aligned} X_n(\cdot) \;&=\; \sum_{j=1}^{p^{(1)}} \alpha_j^{(1)}(\cdot) X_{n-1}(t_j^{(1)} - q + 1) \;+\; \sum_{j=1}^{p^{(2)}} \alpha_j^{(2)}(\cdot) X_{n-2}(t_j^{(2)} - q + 2) \\ &\quad + \ldots + \sum_{j=1}^{p^{(q)}} \alpha_j^{(q)}(\cdot) X_{n-q}(t_j^{(q)}). \end{aligned}$$

There exist examples in which the sparsity assumption given in (9) holds intrinsically in the process. For instance, we rewrite here Example 6.2 of Bosq (2000), that shows that this restricted model class has by himself a good amount of generality.

Example 1. *Let Z be a continuous version of the Ornstein-Uhlenbeck process,*

$$Z(s) = \int_{-\infty}^{s} e^{-\theta(s-t)} \mathrm{d}W(t),$$

where W is a standard Wiener process. If we define $X_n(s) = Z(n+s)$ for $s \in [0,1]$, it can be shown that X_n can be rewritten as in Equation (10)*. In this setting the operator ρ is given, for $f \in C[0,1]$, by*

$$\rho(f)(s) = e^{-\theta s} f(1), \;\; s \in [0,1],$$

that is, $q = 1$, $p = 1$ and $t_1 = 1$. Then $X_n = \rho X_{n-1} + \varepsilon_n$ where now ε_n is a white noise given by

$$\varepsilon_n(s) = \int_{n}^{n+s} e^{-\theta(n+s-t)} \mathrm{d}W(t).$$

It is important to emphasize that, in order to carry out the variable selection, we are not assuming that the model belongs to this restricted class. We are just going to search the points $t_1, \ldots, t_p$ such that the sparse model (10) best approximates the real process X_n according to some criterion.

4.1 Optimality criteria

The following question to address is which should be the optimality criterion to perform the variable selection. The main objective of this regression method is to be able to predict the curve X_n in terms of X_{n-1}. Therefore, since each $X_n(s)$ for $s \in [0,1]$ is a random variable in $L^2_{\mathbb{R}}(\Omega)$, the first approach could be to minimize pointwise the distance

$$q_1(T_p \; ; \; \alpha_1, \ldots, \alpha_p)(s) \; = \; \Big\| X_n(s) - \sum_{j=1}^{p} \alpha_j(s) X_{n-1}(t_j) \Big\|_{L^2_{\mathbb{R}}(\Omega)},$$

in the same spirit as in Berrendero et al. (2017a), where the coefficients $\alpha_j(s)$ (which are just real numbers now) depend on the points $(t_1, \ldots, t_p)$. By Equation (9), all the kernel functions $\phi(s, \cdot)$ depend on the same set of points points $(t_1, \ldots, t_p)$, independently of s. Therefore, we have to find the functions $\alpha_j(\cdot)$ such that their evaluations $\alpha_j(s)$ for $s \in [0,1]$ give the best approximation of $X_n(s)$ for a given set of points T_p. Then we can integrate q_1^2 over s to define

$$Q_1(T_p) \; := \; \int_0^1 \min_{\alpha_j(s) \in \mathbb{R}} q_1(T_p \; ; \; \alpha_1, \ldots, \alpha_p)^2(s) \; \mathrm{d}s, \tag{11}$$

and select the set of points T_p that minimizes this integral. We already know from Berrendero et al. (2017a) that $q_1^2(s)$ is a convex function in $\alpha_j(s)$ for each $s \in [0,1]$. Thus, we can obtain an explicit expression of the minimizing functions, denoted by $\alpha_j^*(s)$, pointwise for each $s \in [0,1]$. We will see in the proof of Proposition 2 that these optimal functions are given by

$$(\alpha_1^*(s), \ldots, \alpha_p^*(s)) \; = \; \Sigma_{T_p}^{-1} \, c_1(s, T_p),$$

where $c_1(\cdot, T_p) = (c_1(\cdot, t_1), \ldots, c_1(\cdot, t_p))'$ is the vector of lagged-covariance functions.

There is another important issue that has to be taken into account when minimizing this function Q_1 over T_p. Up to now we have assumed that the points t_j belong to $[0,1]$. However, if we want to have identifiability of the set $T_p \in [0,1]^p$, we need to restrict the search to a compact subset of this space. This problem has been solved in different ways in the literature. The chosen solution is to work, for some $\delta > 0$, in

$$\Theta_p = \{T_p = (t_1, \ldots, t_p) \in [0,1]^p \ : \ t_{i+1} - t_i \geq \delta, \text{ for } i = 1, \ldots, p\},$$

which is the space proposed in Berrendero et al. (2017a). Another possibilities can be studied, for instance the space used by Ji and Müller (2016). The choice of a value $\delta > 0$ is mainly technical, to avoid problems with the invertibility of the covariance matrices. In addition it allows us also to obtain meaningful sets T_p, since it discards repeated points. This is also not a strong restriction in practice, since usually the data is given in a discretized fashion, and the value δ can be chosen as small as desired. However, as pointed out in Berrendero et al. (2017a), all the theory remains valid for $\delta = 0$, by simply adjusting everywhere the value p to the dimension of the vector without repeated entries.

Although the optimality criterion defined by Q_1 is theoretically sensible, it has not an easily computable expression. We will see in the following result that it can be rewritten in a more feasible way. For the sake of simplicity, we will use the following notation: if we have two sets of real valued functions $\{f_i\}$ and $\{g_i\}$ we will write, using vector notation,

$$\sum_{i=1}^{N} (f_i g_i)(\cdot) \ = \ \sum_{i=1}^{N} f_i(\cdot) g_i(\cdot) \ = \ (f_1(\cdot), \ldots, f_N(\cdot))(g_1(\cdot), \ldots, g_N(\cdot))'.$$

Proposition 2. *Let $(X_n)_{n \in \mathbb{Z}}$ be a FCAR(1) model satisfying Assumption 1 and with $\mathbb{E}\|\varepsilon_n^2\|_\infty < \infty$. Then,*

$$\arg\min_{T_p \in \Theta_p} Q_1(T_p) \ = \ \arg\max_{T_p \in \Theta_p} Q_0(T_p),$$

where $Q_0(T_p) := \int_0^1 c_1(s, T_p)' \Sigma_{T_p}^{-1} c_1(s, T_p) \mathrm{d}s$.

The proof of this result is an extension to the current setting of the proof of Proposition 1 in Berrendero et al. (2017a) and can be found in Section 8.1. As part of this proof, we have seen that these criteria are also equivalent to minimize the average distance in $\mathcal{H}(X)$ between the kernels $\phi(s, \cdot)$ and elements of $\mathcal{H}_0(X)$. This equivalence is theoretically interesting, but does not have real implications for our prediction purposes.

For the criterion of the *FCAR*(q) with $q > 1$ we have to substitute in this proposition the function $c_1(s,t)$ by the continuous picewise-defined function

$$c(s,t) \ = \ c_i(s, t - q + i) \ \text{ for } \ t \in (q - i, q - i + 1] \ \text{ and } \ s \in [0,1]. \qquad (12)$$

In order to carry the proof out, in this case we should assume that the marginal variables of $Z_{n,q}$ are all linearly independent, to ensure the invertibility of the covariance matrices Σ_{T_p} of $Z_{n,q}$, for $T_p = (t_1^{(1)}, \ldots, t_{p(q)}^{(q)}) \in [0,q]^p$. This assumption introduces some additional restrictions to the model, for instance, the functions $\alpha_j^{(i)}(s)$ should not vanish for $s \in [0,1]$ and $1 < i \leq q$.

4.2 Estimation from the sample

As mentioned, the optimality criterion defined by function Q_0 is simple to implement in practice. In this section we study the asymptotic properties of the natural estimator of Q_0. These results will be useful in the next section to show also asymptotic results on the selected points and the estimated trajectories. We will work with a sample $X_1, \ldots, X_m$ of size m drawn from a FCAR(1) process satisfying Assumption 1. Then, for a given number of points p, the natural estimator for the functions $Q_0(T_p)$ is

$$\widehat{Q}_m(T_p) = \int_0^1 \widehat{c}_1'(\cdot, T_p)\ \widehat{\Sigma}_{T_p}^{-1}\ \widehat{c}_1(\cdot, T_p) \mathrm{d}s, \tag{13}$$

where $\widehat{c}_1(\cdot, T_p) = (\widehat{c}_1(\cdot, t_1), \ldots, \widehat{c}_1(\cdot, t_p))'$ and $\widehat{c}_1$ is the usual estimator of the covariance function,

$$\widehat{c}_1(s, t_j) \ = \ \frac{1}{m-1} \sum_{i=1}^{m-1} X_{i+1}(s) X_i(t_j).$$

Equivalently for the entries of the sample covariance matrix $\widehat{c}_0(t_i, t_j)$, $t_i, t_j \in T_p$. Then, according to this criterion, we propose to select as the most relevant points

$$\widehat{T}_p \ = \ \arg\max_{T_p \in \Theta_p} \widehat{Q}_m(T_p). \tag{14}$$

In Section 5 we prove some consistence results for this estimator, under the assumption that the finite dimensional model defined by Equation (10) holds. To this end, we first need to show a couple of convergence results of the sample covariances involved in the expression of $\widehat{Q}_m$. The main one is based on a result of Pumo (1998).

Lemma 1. *Assume that $(X_n)_{n\in\mathbb{Z}}$ is a FCAR(1) model satisfying Assumptions 1 and such that:*

H1. The process $X_n(t)X_n(s)$ for $t, s \in [0,1]$ is uniformly geometrically strong mixing.

H2. (Cramer conditions) For every $t, s \in [0,1]$ there exist constants $m > 0$ and $M < \infty$ such that

- $m \ \leq \ \mathbb{E}[X_0^2(t) X_0^2(s)] \ \leq \ M$ *and*
- $\mathbb{E}|X_0(t)X_0(s)|^k \ \leq \ M^{k-2} k!\ \mathbb{E}[X_0^2(t)X_0^2(s)]$ *for* $k \geq 3$.

Then,

$$\sup_{t,s\in[0,1]} |\widehat{c}_0(s,t) - c_0(s,t)| \overset{\text{a.s.}}{\to} 0 \qquad \textit{and} \tag{15}$$

$$\sup_{t,s\in[0,1]} |\widehat{c}_1(s,t) - c_1(s,t)| \overset{\text{a.s.}}{\to} 0. \tag{16}$$

Proof. By Lemma 1 of Pumo (1998) we know that for some positive constants A_1, A_2, A_3,

$$\mathbb{P}\left(\sup_{t,s\in[0,1]} |\widehat{c}_0(s,t) - c_0(s,t)| \geq \varepsilon\right) \leq (2\sqrt{m} + A_1)\exp\left(-A_2\varepsilon^2\sqrt{m}\right) \\ +A_3\varepsilon^{\frac{2}{5}} m \exp\left(-\log(r^{-1})\sqrt{m}\right),$$

where r is given by assumption H1. By Borel-Cantelli, if the sums over m of these probabilities are finite for every $\varepsilon > 0$, we get the almost sure convergence stated in Equation (15). The sum is of order

$$\sum_{m=1}^{\infty}\mathbb{P}\left(\sup_{t,s\in[0,1]} |\widehat{c}_0(s,t) - c_0(s,t)| \geq \varepsilon\right) \sim 2\sum_{m=1}^{\infty}\frac{\sqrt{m}}{\mathrm{e}^{C_\varepsilon\sqrt{m}}} + \sum_{m=1}^{\infty}\frac{A_1}{\mathrm{e}^{C_\varepsilon\sqrt{m}}} + D_\varepsilon\sum_{m=1}^{\infty}\frac{m}{\mathrm{e}^{C_r\sqrt{m}}},$$

where $C_\varepsilon, C_r, D_\varepsilon > 0$ and these three series converge, for example by the limit comparison test with $\sum m^{-\alpha}, \alpha > 1$. Concerning (16), the same Lemma of 1 of Pumo (1998) states that the bounds for these probabilities are equivalent but with $m-1$ in place of m. □

These Cramer conditions, or some variation of them, appear quite often in the literature related with limit theorems for AR processes in Banach spaces. For instance, all bounded processes satisfy them, and also the Ornstein-Uhlenbeck process of Example 1. In this last case, $|X_n(s)X_n(t)| = e^{-k(t+s)}X_{n-1}(1)^2$, then,

$$e^{-k(t+s)}\mathbb{E}X_0(1)^{2k} \leq M^{k-2}k!e^{-2(t+s)}\mathbb{E}X_0(1)^4,$$

where $X_0(1) \sim \mathcal{N}(0, 0.5)$. Using the expression for the moments of a Gaussian variable,

$$e^{-k(t+s)}\frac{(2k)!}{2^{2k}k!} \leq M^{k-2}e^{-2(t+s)}\frac{3k!}{4},$$

which is satisfied, for instance, for $M \geq 5e^{-2}/12$.

On the basis of the previous result, we can show the uniform convergence of the sample criterion function to its population counterpart and that both functions are continuous on Θ_p.

Lemma 2. *Assume that X_n satisfies the same hypotheses as in Lemma 1. Let $p \geq 1$ be such that the covariance matrices Σ_{T_p} are invertible for all $T_p \in \Theta_p$. Then functions Q_0 and $\widehat{Q}_m$ are continuous on Θ_p and*

$$\sup_{T_p\in\Theta_p} |\widehat{Q}_m(T_p) - Q_0(T_p)| \overset{\text{a.s.}}{\to} 0.$$

The proof of this result can be found in Section 8.2, and it is based on the pointwise properties of the integrands of Q_0 and $\widehat{Q}_m$ shown in Berrendero et al. (2017a).

For the case of greater order $FCAR(q)$ with $q > 1$, this result, and therefore all the results of the next section, hold whenever the process $Z_{n,q}$ fulfils assumptions H1 and H2 of Lemma 1. This is equivalent to suppose that all the products $X_i(t)X_j(s)$ satisfy H1 and H2 for $0 \leq i, j \leq q-1$. That is, as expected, we should impose some extra assumptions on the lagged-covariances in order to be able to estimate them uniformly.

5 Asymptotic results in the finite dimensional setting

In the previous section we introduced the variable selection method in a general context. That is, we have presented an optimality criterion that can be used to select the p most relevant points, without imposing any additional restriction to the model that generates the data. However, if we assume that the data is generated by the restricted class of kernels like Equation (9), we can prove some asymptotic results for the estimator $\widehat{T}_p$ and also for the estimated trajectories. Therefore, in this section we are going to assume that the kernel of the model depends only on p^* points,

$$\phi(s, \cdot) \;=\; \sum_{j=1}^{p^*} \alpha_j(s) c_0(t_j^*, \cdot), \tag{17}$$

for all $s \in [0,1]$, where p^* is the minimum integer for which this expression holds. We will denote as $T^* = T^*_{p^*} \in \Theta_{p^*}$ the set of points that generate the model. Two questions arise straightway from this expression; how good is our estimator $\widehat{T}_{p^*}$ when searching the real points T^*, and how can we approximate the real number of points p^*.

5.1 Estimated points and trajectories

From the expression of Q_1 given in Equation (11), it is clear that under (17), the set T^* is a global minimum of Q_1 on Θ_{p^*}, and therefore a global maximum of Q_0. If we assume for now that the value p^* is known, under some reasonable conditions we can prove that this optimum is unique and thus the estimated points $\widehat{T}_{p^*}$ converge to the real ones T^*. This result is an extension of Theorem 1 of Berrendero et al. (2017a) and its proof is included in Section 8.3.

Theorem 2. *Under the assumptions of Lemma 2 for $p = p^*$, whenever* (17) *holds and the covariance matrix $\Sigma_{T_{p^*} \cup S_{p^*}}$ is invertible for all $T_{p^*}, S_{p^*} \in \Theta_{p^*}$ with $T_{p^*} \neq S_{p^*}$, then:*

(a) The vector $T^* \in \Theta_{p^*}$ is the only global maximum of Q_0 on this space.

(b) $\widehat{T}_{p^*} \xrightarrow{a.s.} T^*$ with the sample size, where $\widehat{T}_{p^*}$ is given in Equation (14) *using $p = p^*$.*

(c) $\widehat{T}_{p^*}$ converges to T^* in quadratic mean.

Once that we have selected the most relevant points from the sample, we want to estimate the trajectories of the process. That is, we want to approximate

$$X_{n,T^*}(\cdot) \;\equiv\; \rho X_{n-1}(\cdot) \;=\; \alpha_1(\cdot) X_{n-1}(t_1^*) + \ldots + \alpha_{p^*}(\cdot) X_{n-1}(t_{p^*}^*). \tag{18}$$

In the proof of Proposition 2 we have seen that the functions $(\alpha_1(\cdot), \ldots, \alpha_{p^*}(\cdot))'$ used to carry out this projection are given by $\Sigma_{T_{p^*}}^{-1}(c_1(\cdot, t_1^*), \ldots, c_1(\cdot, t_{p^*}^*))$. Therefore, we can construct the estimated curve as

$$\widehat{X}_{n,\widehat{T}_{p^*}}(\cdot) \;=\; \widehat{\alpha}_1(\cdot) X_{n-1}(\widehat{t}_1) + \ldots + \widehat{\alpha}_{p^*}(\cdot) X_{n-1}(\widehat{t}_{p^*}),$$

where now the functions $(\widehat{\alpha}_1(\cdot), \dots, \widehat{\alpha}_{p^*}(\cdot))'$ are computed using the sample version of the covariances as $\widehat{\Sigma}^{-1}_{\widehat{T}_{p^*}}(\widehat{c_1}(\cdot, \widehat{t}_1), \dots, \widehat{c_1}(\cdot, \widehat{t}_{p^*}))$. Thus our proposed estimator for X_{n,T^*} is

$$\widehat{X}_{n,\widehat{T}_{p^*}}(\cdot) \;=\; \widehat{c}_{\widehat{T}_{p^*}}(\cdot)'\widehat{\Sigma}^{-1}_{\widehat{T}_{p^*}} X(\widehat{T}_{p^*}). \tag{19}$$

Under the same conditions of the previous theorem, we can see that this estimator converges to X_{n,T^*} a.s. in $C[0,1]$, and also in $L^2_C(\Omega)$ imposing an additional assumption to the process.

Theorem 3. *Under the same assumptions of Theorem 2,*

(a) $\widehat{X}_{n,\widehat{T}_{p^*}}(\cdot)$ *converges to* $X_{n,T^*}(\cdot) = \rho X_{n-1}$ *a.s. in* $C[0,1]$ *(that is,* $\|\widehat{X}_{n,\widehat{T}_{p^*}} - X_{n,T^*}\| \overset{\text{a.s.}}{\to} 0$*).*

(b) If, in addition, there exists $\eta > 0$ *such that* $\mathbb{E}\||X_n|^{2+\eta}\| < \infty$*, then it also converges in* $L^2_C(\Omega)$ *(that is,* $\mathbb{E}\|\widehat{X}_{n,\widehat{T}_{p^*}} - X_{n,T^*}\|^2 \to 0$*).*

The proof of this theorem is based on the one of Theorem 2 of Berrendero et al. (2017a) and can be found in Section 8.4.

5.2 Number of relevant points

The remaining question is how can we decide the number points to select, since the value p^* is unknown in most of the cases. Here we propose a consistent estimator for this quantity. The idea behind of the proposal is that, in view of the expression for the optimality criterion Q_1 (Equation (11)), its minimum value is not reached if we use $p < p^*$, but by using $p > p^*$ we can not improve more. Therefore, the number p^* would be the smallest p such that the minimum value of $Q_1(T_p)$ (or the maximum of Q_0) keeps unchanging when adding more points.

The sample version of this idea would be as follows: if we define

$$\Delta = \min_{p<p^*}(Q_0(T^*_{p+1}) - Q_0(T^*_p)) > 0$$

and we are able to fix some $0 < \varepsilon < \Delta$, we define

$$\widehat{p} \;=\; \min\{p \;:\; \widehat{Q}^{max}_m(p+1) - \widehat{Q}^{max}_m(p) < \varepsilon\}, \tag{20}$$

where $\widehat{Q}^{max}_m(p) = \max_{T_p \in \Theta_p} \widehat{Q}_m(T_p)$. This estimator is a.s. consistent for the real number of relevant variables.

Theorem 4. *Suppose that assumptions of Lemma 2 hold for* $p \le p^*$ *and that* p^* *is the smallest integer such that Equation* (17) *is satisfied. Then the estimator given by Equation*(20) *fulfils* $\widehat{p} \overset{\text{a.s.}}{\to} p^*$*.*

Proof. We can prove similar results as the ones given in Lemma 4 of Berrendero et al. (2017a) using the same reasoning, with the only difference that now $Q_0(T^*) = \int_0^1 \|\phi(s,\cdot)\|^2_{\mathcal{H}} \mathrm{d}s$ (as in the proof of Proposition 2). □

When this estimator $\widehat{p}$ is used in practice, there exists still the problem of fixing the value ε. In the following section we will comment a couple of possible approaches to solve it, that seem to perform well in the tested examples.

6 Experiments

In this section we present the results of the experimental study developed in order to check the performance of the proposal, both using real and simulated data. We compare our proposal with other two methods of the recent literature. Before introducing the experimental setting, we make a couple of theoretical comments that ease the implementation of the method besides providing the pseudo-code.

6.1 Practical considerations

In order to obtain the most relevant points in practice, we should maximize the expression of $\widehat{Q}_m$ (given in Equation (13)) in the p-dimensional space Θ_p. However, due to computational limitations, this optimization is not feasible even for relatively small values of p. Therefore, some kind of greedy approximation should be carried out. Under the assumption that the process is non-degenerate, we can decompose the function Q_0 in a way that directly suggests an iterative approximation to this optimization problem. If the vector $T_{p+1} \in \theta_{p+1}$ is such that it contains all the entries of T_p plus a new one $t_{p+1} \in [0,1]$, using Equation (16) of Berrendero et al. (2017a) we can write,

$$\begin{aligned} Q_0(T_{p+1}) &= \int_0^1 c_1'(s, T_{p+1})\ \Sigma_{T_{p+1}}^{-1}\ c_1(s, T_{p+1})\ \mathrm{d}s \\ &= Q_0(T_p)\ +\ \frac{\int_0^1 \operatorname{cov}(X_n(s) - X_{n,T_p}(s), X_{n-1}(t_{p+1}))^2\ \mathrm{d}s}{\operatorname{var}(X_{n-1}(t_{p+1}), X_{n-1,T_p}(t_{p+1}))}, \end{aligned}$$

where X_{n,T_p} and X_{n-1,T_p} are the same kind of projections defined in Equation (18). This derivation can be also done using the sample counterpart of Q_0,

$$\widehat{Q}_m(T_{p+1}) = \widehat{Q}_m(T_p) + \frac{\int_0^1 \widehat{\operatorname{Cov}}(X_n(s) - X_{n,T_p}(s), X_{n-1}(t_{p+1}))^2\ \mathrm{d}s}{\widehat{\operatorname{Var}}(X_{n-1}(t_{p+1}), X_{n-1,T_p}(t_{p+1}))}.$$

Then the proposed algorithm selects at each step the point t_{p+1} that maximizes this quotient. The starting point would be the one that maximizes $\widehat{Q}_m(t) = \widehat{c}_0(t,t)^{-1} \int_0^1 \widehat{c}_1(s,t)^2 \mathrm{d}s$. As usual when dealing with greedy algorithms, this approximation does not guaranty that the global maximum of $\widehat{Q}_m$ is reached. However, as shown below, it performs well in practice.

In order to compute it we can use a similar reasoning as in the proof of Proposition 2 of Berrendero et al. (2017a) and rewrite the previous equation as,

$$\widehat{Q}_m(T_{p+1}) = \widehat{Q}_m(T_p)\ +\ \frac{\int_0^1 (\widehat{c}_1(s,T_p)'\ \widehat{\Sigma}_{T_p}^{-1}\ \widehat{c}_0(t_{p+1},T_p) - \widehat{c}_1(s,t_{p+1}))^2\ \mathrm{d}s}{\widehat{c}_0(t_{p+1},t_{p+1}) - \widehat{c}_0(t_{p+1},T_p)'\ \widehat{\Sigma}_{T_p}^{-1}\ \widehat{c}_0(t_{p+1},T_p)}, \qquad (21)$$

where $\widehat{c}_1(s,T_p)$ is the vector whose entries are given by the sample covariances $\widehat{\operatorname{Cov}}(X_n(s), X_{n-1}(t_j))$, and equivalently for $\widehat{c}_0$. Also due to computational limitations, the search of the most relevant points should be done on a grid of $[0,1]$. If the data is given in a discretized fashion, then the grid is directly given by the data. However, if it is fully functional, the grid can be defined arbitrarily fine. Under the assumption that all the covariance matrices Σ_{T_p} are invertible for T_p in this grid, the quotient of Equation (21) is easy to compute. In addition, depending on the nature of the data, the estimations of the covariances $\widehat{c}_1(\cdot, T_p)$ can be made fully functional or using the values on the fixed grid. Both possibilities are implemented for the numerical study.

Taking these things into account, the pseudo-code of the proposal for the variable selection, given a fixed value p, would be the one provided in Algorithm 1. As explained before, if we want to use the model for order grater than one, we should substitute the sample lagged covariance function $\widehat{c}_1(s,t)$ by the sample version of the picewise-defined function of Equation (12).

Algorithm 1 Variable selection for a given p

```
procedure RKHS VARIABLE SELECTION
First point:
    $Quot(grid) \leftarrow \widehat{c}_0(t,t)^{-1} \int_0^1 \widehat{c}_1(s,t)^2 \mathrm{d}s,\ \forall t \in grid$
    $M(1) \leftarrow max(Quot)$
    $pts(1) \leftarrow$ which $Quot == M(1)$
Rest of the points:
    for i in 2 to p do
        $Quot(grid \backslash pts) \leftarrow$ Quotient of Eq. (21) $\forall t_{p+1} \in \{grid$ and not in $pts\}$
        $M(i) \leftarrow max(Quot)$
        $pts(i) \leftarrow$ which $Quot == M(i)$
Return pts (points) and M
```

As mentioned in the previous section, other sensitive point in practice is how many points should we keep. In order to apply the estimator of Equation (20), some value for ε must be fixed. In other words, it should be determined for which value of p the quotient of Equation (21) has converged to zero. The standard approach to this problem is to fix the parameter by cross-validation. However, we try also in the experiments the proposal given in Berrendero et al. (2017a); to apply the usual k-means with $k = 2$ to the logarithms of the values of the quotient in Equation (21). If we denote has $L_m(p)$ the values of these logarithms, $\widehat{p}$ would be the minimum p such that all the $L_m(p)$ for $p > \widehat{p}$ do not belong to the same cluster as $L_m(1)$. The pseudo-code for this clustering approximation is given in Algorithm 2.

Algorithm 2 Cluster approximation to estimate $\widehat{p}$

```
procedure NUMBER OF POINTS
    $M, points \leftarrow$ RKHS method for P points
    $L_m \leftarrow log(M)$
k-means procedure with $k = 2$:
    $clusters \leftarrow kmeans(L_m)$
    $cl1 \leftarrow clusters(1)$
    $\widehat{p} \leftarrow$ tail of $clusters == cl1$
```

6.2 Methodology

As mentioned, we compare the efficiency of the proposal with two other recent methods. Both of them carry out the dimension reduction using functional principal components. We also compare with two "base" methods that do not reduce dimension. These methods allow us to contextualize the errors, since they provide us something comparable to bounds of these errors. We indicate in brackets the name used in the tables for each method.

- The method proposed in this paper (RKHS) has been implemented in four different ways. As mentioned in the previous section, we use two approaches to select the number of relevant variables; doing clustering on the maximum values of the $\widehat{Q}_m$ functions (CL) and by cross-validation (CV). In addition, the points can be selected by using covariance vectors on a grid or computing the purely functional lagged-covariance functions. We use one or another depending on the nature of the data.

- Method proposed in Aue et al. (2015) (fFPE). This proposal uses a dimension reduction method based on functional principal components analysis to find a finite dimensional space on which the prediction is performed using a vector autoregressive model. The model order and dimension of the finite dimensional space are chosen by the fFPE criterion. For details, see Aue et al. (2015), where the good empirical properties of the approach are demonstrated in depth.

- Method proposed in Bosq (2000) and Kokoszka and Reimherr (2013) (KR). This prediction method by Bosq is the one known as the standard prediction method for functional autoregressive processes. To determine the order of the functional autoregressive model to be fitted, we use the multiple testing procedure of Kokoszka and Reimherr (2013).

- Exact and Naive methods are implemented in order to provide some bounds on the errors. These methods are also used, for instance, in Horváth and Kokoszka (2012). The exact prediction consists in "predicting" the response directly as ρX_{n-1}. Therefore, it can be only applied for simulated data, since the operator is unknown for real data sets. It is not really a prediction method but gives us an idea of the minimum error that we can make. The Naive approach simply predicts $\widehat{X}_n$ as X_{n-1}.

Both, the maximum numbers of points to select and the number of principal components, are always limited to 10. For the simulated data, all the methods are tested using a sample size $n = 115$, where 100 realizations are used for training and the remaining 15 for test. Different sample sizes have been tested, but no significant differences have been detected. Each experiment has been replicated 100 times. For the real data sets we use a window moving approach with five blocks to obtain several measures of the errors. The size of the windows is adjusted depending on the sample size of each set. The order of the process is always limited to 3 for all the methods. However, for our implementations we have to set it to order one for most of the simulated data, since sometimes the curves can not be interpreted as $X_n(s) = Z(s+n)$ being Z a continuous trajectory.

Usually the functional data sets are given in a discretized fashion. Some of the tested methods require to transform previously the data to truly functional. However, our discrete proposal can deal also with discretized data. In addition, when the data is irregular, some information could be lost when transforming the data to functional. This complicates the comparison between the different methods. Therefore, for this kind of discretized data sets we measure two different types of errors.

- Discrete errors: The error is measured using the original discretized data. The discrete version of the proposal (the one that uses covariance vectors) is tested. The predictions returned by the methods that use truly functional data are evaluated into the same grid given by the data.
- Functional errors: The data are transformed to functions using a Bsplines base before applying the methods. We have found that using Bsplines is more suitable in this setting, since the standard Fourier basis introduce a periodicity in the data that complicate the inversion of the covariance matrices. For most of the data sets, 10 functions of the basis are used for the displayed results, but different numbers has been tested without significant changes. This transformed set is the one used to measure the errors. The functional version of the proposal is tested now, estimating the whole lagged-covariance functions in a purely functional way.

As we will see in the following section, one of the simulated sets is purely functional. In this case only the functional errors are measured. Two different norms are used to measure the error: the typically used $L^2[0,1]$ norm and the supremum norm of $C[0,1]$ that has been used along the paper. Each of these norms measure different characteristics of the predictions. We also measure two different relative errors,

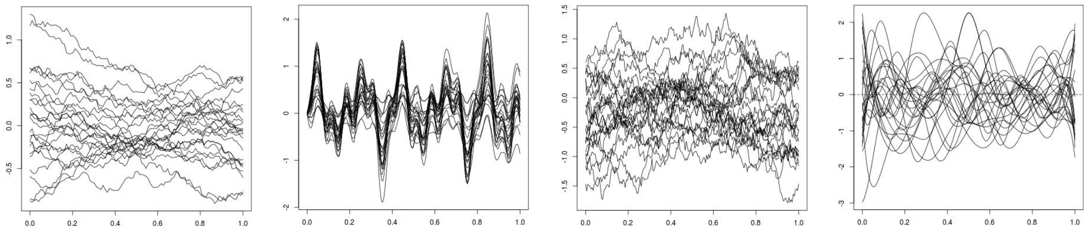
$$\varepsilon_1 = \sum_{i=1}^{n} \frac{\|X_i - \widehat{X}_i\|}{\|X_i\|}, \tag{22}$$

$$\varepsilon_2 = \frac{\sum_{i=1}^{n} \|X_i - \widehat{X}_i\|}{\sum_{i=1}^{n} \|X_i\|}. \tag{23}$$

The first one gives the same importance to all curves regardless of their norm, while the second one place more importance to the errors in the curves of biggest norms, since it is just a scaling of the absolute error.

6.3 Simulated data

We test the different methods using simulated sets that fulfil the sparsity assumption of Equation (17) as well as some which not. Most of them are inspired by other data sets used in the literature. Some realizations of these processes can be found in Figure 1.



(a) Sparse with logs. (b) Sparse with sins. (c) O.U. (d) FAR

Figure 1: 25 trajectories of each of the simulated data sets.

- Two data sets satisfying the sparsity assumption with standard Brownian innovations. The real points are $T^* = (0.3, 0.5, 0.9)$ with two different sets of functions α_j. The first ones are logarithms, $\log((1+s)j^{-1})$ for $j = 1, 2, 3$ and $s \in [0, 1]$, similar to the function used for the simulated data in Berrendero et al. (2017a). The second set of functions is $sin(30\pi j^{-1}s)$, since we also want to incorporate a data set with high variation. When transforming this last data set to purely functional, 30 Bspline functions are used instead of 10, to be able to capture most of the variation.

- Ornstein-Uhlenbeck process exposed in Example 1, which corresponds to Example 6.2 of Bosq (2000). This is the only simulated set for which $X_n(s) = Z(s+n)$, so that we can use the model $FCAR(3)$.

- FAR with linearly decaying eigenvalues of the covariance operator ($s = 1 : 15$). Following the simulation example used in Aue et al. (2015), this set consists of spanning a D-dimensional space by the first D Fourier basis functions, and to then generate random $D \times D$ parameter matrices and a D-dimensional noise process, where the construction ensures a linear decay of the eigenvalues of the covariance operator. The slow decay of the eigenvalues of the covariance operator makes sure that problems with PCA based methods due to non-invertibility of the covariance operator are avoided. In this example only the functional errors are measured since it is purely functional by construction.

Table 1 summarizes the different measurements for these data sets. Regarding our two proposals, there is not a method that uniformly outperform the other one. In general, our proposals and the FPCA approach with the fFPE criterion are mainly the victors (marked with bold numbers). The code used to run these simulations is provided as ancillary file.

6.4 Real data sets

We test also a couple of real data sets already used in other recent papers.

- Particulate matter concentrations (PM10). This data set is used, for instance, in Aue et al. (2015) and consists on 175 samples. It contains the $\mu g m^{-1}$ concentration in air of a particular substance with an aerodynamic diameter of less than 10 μm. The measures were taken each half hour from

				RKHS+cl	RKHS+CV	fFPE	KR	Exact	Naive
Sparse with log.	ε_1 error	Disc.	L2	**0.64**	0.71	0.65	0.66	0.55	3.24
			sup	**0.68**	0.71	**0.68**	0.82	0.65	1.64
		Funct.	L2	0.65	**0.64**	0.67	0.67	0.56	3.32
			sup	**0.65**	**0.65**	**0.65**	0.83	0.62	1.67
	ε_2 error	Disc.	L2	**0.16**	0.18	0.17	0.18	0.14	1.32
			sup	**0.30**	0.32	**0.30**	0.39	0.29	0.80
		Funct.	L2	**0.32**	**0.32**	0.33	0.34	0.28	2.63
			sup	**0.55**	**0.55**	0.56	0.77	0.53	1.57
Sparse with sins	ε_1 error	Disc.	L2	**0.72**	0.74	0.83	0.94	0.60	2.42
			sup	**0.71**	0.73	0.84	0.95	0.67	1.50
		Funct.	L2	**0.78**	0.81	0.81	0.94	0.65	2.60
			sup	**0.77**	**0.77**	**0.77**	0.95	0.69	1.53
	ε_2 error	Disc.	L2	**0.38**	**0.38**	0.42	0.48	0.33	1.10
			sup	**0.36**	0.37	0.42	0.49	0.34	0.75
		Funct.	L2	**0.78**	0.79	0.80	0.91	0.69	2.19
			sup	**0.75**	**0.75**	**0.75**	0.94	0.68	1.47
O.U.	ε_1 error	Disc.	L2	1.07	1.01	**1.00**	1.15	0.83	2.33
			sup	**0.88**	**0.88**	0.93	0.98	0.88	1.35
		Funct.	L2	**1.00**	**1.00**	1.05	1.20	0.85	2.49
			sup	**0.91**	**0.91**	0.92	0.98	0.86	1.34
	ε_2 error	Disc.	L2	**0.31**	**0.31**	0.35	0.42	0.30	0.65
			sup	**0.44**	**0.44**	0.47	0.50	0.44	0.65
		Funct.	L2	**0.66**	**0.66**	0.68	0.83	0.59	1.26
			sup	**0.85**	**0.85**	0.86	0.94	0.81	1.21
FAR	ε_1 error		L2	**1.00**	1.01	1.01	1.13	0.85	2.20
			sup	**1.00**	**1.00**	**1.00**	1.04	0.91	1.45
	ε_2 error		L2	**0.96**	**0.96**	0.97	1.08	0.81	1.96
			sup	**0.98**	**0.98**	**0.98**	1.02	0.89	1.39

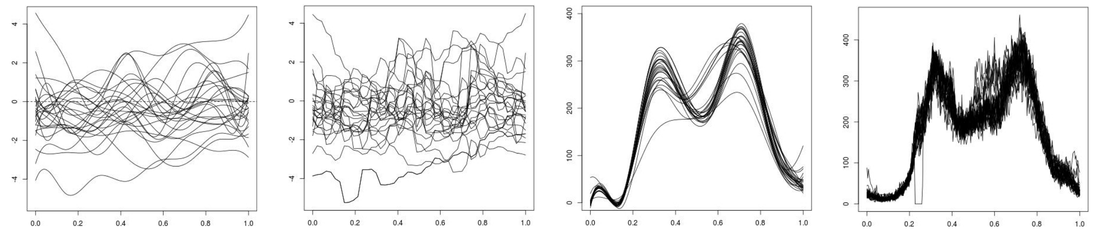
Table 1: Errors for the simulated data sets (ε_1 of Eq. (22) and ε_2 of Eq. (23))

from October 1, 2010 to March 31, 2011 in Ganz, Austria. The data is preprocessed in the same way as suggested in Aue et al. (2015). For the five windows we take blocks of 115 observations, 100 for training and 15 for test.

- Vehicle traffic data presented in Aue and Klepsch (2017). The original data set was provided by the Autobahndirektion Südbayern. This data set contains the amount of vehicles traveling each minute on the highway A92 in Southern Bavaria, Germany, from January 1 to June 30, 2014. Retaining only working days, we work with 119 samples divided into 5 windows of size 99; 94 for train and 5 for test.

Some curves of these data sets are included in Figure 2. As usual, the original data are given as discrete measurements, so we show both fashions; the original discrete sets and their functional approximations.

In Table 2 we summarize the different error measurements for both data sets. Taking these results into account, it is even less clear which implementation of our proposal, the cross-validation one or the cluster one, is the best choice. In both examples it seems that the FPCA approach with fFPE slightly outperforms the other methods. However, the differences between it and our proposals are in



(a) Functional PM10. (b) Original PM10. (c) Functional traffic. (d) Original traffic.

Figure 2: 25 trajectories of the real data sets, both discrete and functional.

general small, even achieving the same error, or improving it, in about half of the measures.

In addition, Table 3 shows the mean execution times of each of the methods for the PM10 set. Here both the functional (f) and the discrete (d) implementations of our proposal are measured. It seems that working with the transformed functional data is slower in general, and then our two discrete implementations are considerably faster than the other methods.

				RKHS+cl	RKHS+CV	fFPE	KR	Naive
PM10	ε_1 error	Disc.	L2	0.97	**0.74**	0.82	1.48	1.65
			sup	0.92	**0.86**	**0.86**	1.06	1.15
		Func.	L2	**0.73**	0.81	0.82	1.59	1.68
			sup	0.86	0.88	**0.85**	1.08	1.11
	ε_2 error	Disc.	L2	0.56	**0.47**	0.50	0.86	0.80
			sup	0.85	0.81	**0.80**	0.98	1.02
		Func.	L2	0.52	0.51	**0.48**	0.85	0.76
			sup	0.82	0.82	**0.78**	0.99	0.97
Traffic	ε_1 error	Disc.	L2	**1.00**	1.02	**1.00**	1.04	67.48
			sup	1.02	**1.01**	**1.01**	**1.01**	4.70
		Func.	L2	1.52	**1.42**	1.49	**1.42**	240.57
			sup	1.08	1.11	**1.05**	1.11	10.22
	ε_2 error	Disc.	L2	0.90	0.89	**0.83**	0.95	40.57
			sup	0.99	**0.98**	**0.98**	0.99	4.26
		Func.	L2	0.84	0.93	**0.75**	0.92	62.07
			sup	0.92	0.99	**0.90**	1.00	6.82

Table 2: Errors for real data sets (ε_1 of Eq. (22) and ε_2 of Eq. (23))

	RKHS+cl (d)	RKHS+CV (d)	RKHS+cl (f)	RKHS+CV (f)	fFPE	KR
1	0.10	0.09	0.60	0.76	0.66	2.72

Table 3: Execution times in seconds

7 Conclusions

Along the present paper we have fundamentally extended the theory developed for regression with scalar response and independent data, introduced by Berrendero et al. (2017a), to the setting of prediction of functional time series, whose dependence is modeled using an autoregressive structure. That is, a variable selection technique for prediction is developed, based on the theory of Reproducing Kernel Hilbert Spaces. This variable selection approach helps to overcome some of the usual problems coming from the use of other dimension reduction techniques. Additionally, our model is general enough to incorporate the finite dimensional representations of the process as a particular case. In this setting, the change of environment from the standard $L^2[0,1]$ to $C[0,1]$ allows us to prove the uniform almost sure convergence of the estimations. We also provide an a.s. consistent estimator for the number of relevant variables involved in the model, which has been shown to be computationally efficient.

When compared with other prediction methods of the literature, our proposal seems to be quite competitive. According to our experiments, the proposed method seems to perform better than the others whenever our theoretical assumptions are satisfied. In the couple of real data sets tested, it seems that it slightly outperformed. However, if we take also into account the execution times, we find its performance reasonably. That is, our proposals, specifically the discrete approaches, would be more suitable for large data sets. Besides, the proposed estimators do not depend on the fashion in which the data is given; the implementation can be directly adapted to discrete or fully functional data sets.

8 Proofs

In this section we include the proofs that are mainly based on the pointwise results of Berrendero et al. (2017a).

8.1 Proposition 2

This proof is inspired by the one of Proposition 1 of Berrendero et al. (2017a). Pointwise for each $s \in [0,1]$, using Loève's isometry we have that

$$\Big\|X_n(s) - \sum_{j=1}^p \alpha_j(s) X_{n-1}(t_j)\Big\|^2_{L^2_{\mathbb{R}}(\Omega)} \;=\; \Big\|\phi(s,\cdot) - \sum_{j=1}^p \alpha_j(s) c_0(t_j,\cdot)\Big\|^2_{\mathcal{H}} + \sigma(s),$$

where $\sigma(s) = \operatorname{var}(\varepsilon_n(s)) \leq \|\mathbb{E}\varepsilon_n^2\| \leq \mathbb{E}\|\varepsilon_n^2\| < \infty$, so the minimizing values $\alpha_j(s)$ are the same for both sides of the equality. Again pointwise for each $s \in [0,1]$, using the reproducing property of $\mathcal{H}(X)$,

$$\begin{aligned}\Big\|\phi(s,\cdot) - \sum_{j=1}^p \alpha_j(s) c_0(t_j,\cdot)\Big\|^2_{\mathcal{H}} \;=\; & \|\phi(s,\cdot)\|_{\mathcal{H}} + \sum_{i,j=1}^p \alpha_i(s)\alpha_j(s) c_0(t_i,t_j) \\ & - 2\sum_{j=1}^p \alpha_j(s)\phi(s,t_j).\end{aligned}$$

Since c_0 is a positive-definite function, this last function is convex in $\alpha_j(s)$ for each $s \in [0,1]$. Therefore we can compute its minimum pointwisely, which is achieved at $(\alpha_1^*(\cdot),\ldots,\alpha_p^*(\cdot))' = \Sigma_{T_p}^{-1}c_1(\cdot,T_p)$, since $c_1(s,t) = \phi(s,t)$ for each s (Equation (5)). Then if we substitute this optimum in the previous equation we get

$$\min_{\alpha_j(s)\in\mathbb{R}} \left\|\phi(s,\cdot) - \sum_{j=1}^{p} \alpha_j(s)c_0(t_j,\cdot)\right\|_{\mathcal{H}}^2 \;=\; \|\phi(s,\cdot)\|_{\mathcal{H}}^2 - c_1(s,T_p)'\Sigma_{T_p}^{-1}c_1(s,T_p).$$

Hence, integrating over $s \in [0,1]$,

$$Q_1(T_p) = \int_0^1 \sigma(s)\mathrm{d}s + \int_0^1 \|\phi(s,\cdot)\|_{\mathcal{H}}^2\mathrm{d}s - Q_0(T_p) = C - Q_0(T_p).$$

This constant C is finite since the integral of $\sigma(s)$ is bounded by $\|\mathbb{E}\varepsilon_n^2\| < \infty$ and $\int_0^1 \|\phi(s,\cdot)\|_{\mathcal{H}}^2\mathrm{d}s \le \sup_{s\in[0,1]} \|\phi(s,\cdot)\|_{\mathcal{H}}^2$ being $\|\phi(s,\cdot)\|_{\mathcal{H}}^2$ a continuous function on $[0,1]$ (it is the composition of two continuous functions, $s \mapsto \phi(s,\cdot) = c_1(s,\cdot)$ and $f \mapsto \|f\|_{\mathcal{H}}$).

8.2 Proof of Lemma 2

Denoting

$$Q_0(T_p) \;=\; \int_0^1 c_1(s,T_p)'\Sigma_{T_p}^{-1}c_1(s,T_p)\;\mathrm{d}s \;=\; \int_0^1 q_0(T_p;s)\mathrm{d}s,$$

$$\widehat{Q}_m(T_p) \;=\; \int_0^1 \widehat{c}_1(s,T_p)'\widehat{\Sigma}_{T_p}^{-1}\widehat{c}_1(s,T_p)\;\mathrm{d}s \;=\; \int_0^1 \widehat{q}_m(T_p;s)\mathrm{d}s,$$

we can extend the proofs of Lemmas 2 and 3 of Berrendero et al. (2017a) to our setting.

For the continuity of the functions, it can be shown that $q_0(T_p;s)$ and $\widehat{q}_m(T_p;s)$ are continuous in T_p for each $s \in [0,1]$, using the same reasoning as in the proof of Lemma 2 of Berrendero et al. (2017a) but using now Lemma 1. Then if $\|T_p - S_p\|_2 < \delta$, where $\|\cdot\|_2$ is the usual vector norm,

$$|Q_0(T_p) - Q_0(S_p)| \;=\; \left|\int_0^1 (q_0(T_p;s) - q_0(S_p;s))\mathrm{d}s\right| \;\le\; \int_0^1 |q_0(T_p;s) - q_0(S_p;s)|\mathrm{d}s \;<\; \varepsilon.$$

And equivalently to see that $\widehat{Q}_m$ is continuous with probability one.

To see the uniform convergence of the functions we will use, by the same reasoning as in Lemma 3 of Berrendero et al. (2017a) with Lemma 1, that $\sup_{T_p\in\Theta_p} |\widehat{q}_m(T_p;s) - q_0(T_p;s)| = \|\widehat{q}_m(\cdot;s) - q_0(\cdot;s)\|_\infty$ goes to 0 a.s. for each $s \in [0,1]$. Then, since $\|\cdot\|_\infty$ is a convex function (by Minkowski's inequality), we get

$$\|\widehat{Q}_m(\cdot) - Q_0(\cdot)\| \;=\; \left\|\int_0^1 (\widehat{q}_m(\cdot;s) - q_0(\cdot;s))\mathrm{d}s\right\| \;\le\; \int_0^1 \|\widehat{q}_m(\cdot;s) - q_0(\cdot;s)\|\;\mathrm{d}s \overset{a.s}{\to} 0,$$

by Jensen's inequality.

8.3 Proof of Theorem 2

(a) Because of the equivalence of the criteria proved in Proposition 2, it is enough to see that T^* is the only global minimum of $Q_1(T_{p^*})$ in Θ_{p^*}. From Equation (11) it is clear that T^* minimizes Q_1 since

$$\begin{aligned} Q_1(T_{p^*}) &= \int_0^1 \|X_n(s) - X_{n,T_{p^*}}(s)\|^2_{L^2_{\mathbb{R}}(\Omega)} \, \mathrm{d}s \\ &= \int_0^1 \|X_{n,T^*}(s) - X_{n,T_{p^*}}(s)\|^2_{L^2_{\mathbb{R}}(\Omega)} \, \mathrm{d}s + \|\mathrm{var}(\varepsilon_n)\|_{L^2[0,1]}, \end{aligned}$$

where $L^2[0,1]$ is the space of square integrable functions over $[0,1]$, and therefore its minimum value is $\|\mathrm{Var}(\varepsilon_n)\|_{L^2[0,1]} = \|\sigma\|_{L^2[0,1]}$. If there exists another vector $S^* \neq T^*$ that it also achieves this value, we must have $\|X_{n,T^*}(s) - X_{n,S^*}(s)\|^2_{L^2(\Omega)} = 0$ for almost every $s \in [0,1]$ (except on a set of measure zero with respect the Lebesgue measure). However, it is enough to have one point s_0 in which this equality holds. For this point we have that $X_{n,T^*}(s_0) = X_{n,S^*}(s_0)$ a.s., which contradicts the assumption that the covariance matrix $\Sigma_{T^* \cup S^*}$ is invertible, and then $T^* = S^*$.

(b) The result follows from the fact that $\widehat{Q}_m$ and Q_0 are continuous functions such that $\widehat{Q}_m$ tends uniformly a.s. to Q_0 (Lemma 2) and Q_0 has a unique maximum in Θ_{p^*} (part (a)).

(c) The same argument as in the proof of Theorem 1.c of Berrendero et al. (2017a).

8.4 Proof of Theorem 3

(a) By a proof analogous to the one of Theorem 2 of Berrendero et al. (2017a) it can be shown that $\widehat{X}_{n,\widehat{T}_{p^*}}(s) \overset{a.s.}{\to} X_{n,T^*}(s)$ for each $s \in [0,1]$. By Ascoli-Arzela theorem, since the functions $\widehat{X}_{n,\widehat{T}_{p^*}}(\cdot)$ are defined on the compact space $[0,1]$, if they are uniformly bounded and equicontinuos, then they converge uniformly to X_{n,T^*}.

Using Lemma 1 it can be shown that the sample version of the inverse of the covariance matrix, $\widehat{\Sigma}^{-1}_{T_{p^*}}$, as a function on $T_{p^*} \in \Theta_{p^*}$, converges uniformly with probability one to its population counterpart $\Sigma^{-1}_{T_{p^*}}$ (using a similar reasoning as in the proof of Lemma 3 of Berrendero et al. (2017a)). Then by Theorem 2(b), $\widehat{\Sigma}^{-1}_{\widehat{T}_{p^*}}$ also converges uniformly a.s. on Θ_{p^*} to $\Sigma^{-1}_{T_{p^*}}$. Therefore, in view of Equation (19), in order to check the uniformly boundedness, it is enough to have that the functions $\widehat{c}_1(s, \widehat{t}_j)$ are bounded for $s \in [0,1]$. The remaining terms of (19) converge a.s. to the real ones and do not depend on s. It follows straightforward from Equation 16 that,

$$\widehat{c}_1(s, \widehat{t}_j) \;\leq\; \sup_{s,t \in [0,1]} |\widehat{c}_1(s,t)| \;\leq\; C + \sup_{s,t \in [0,1]} |c_1(s,t)| \;<\; \infty$$

since it is a continuous function on a compact.

In order to prove equicontinuity, as the support is compact, it is enough to see that the functions are equicontinuous at every point $s_0 \in [0,1]$. If we denote as $\|\cdot\|_2$ the usual norm of a vector,

$$\begin{aligned}
|\widehat{X}_{n,\widehat{T}_{p^*}}(s_0) - \widehat{X}_{n,\widehat{T}_{p^*}}(s)| &= |(\widehat{c}_1(s_0,\widehat{T}_{p^*}) - \widehat{c}_1(s,\widehat{T}_{p^*}))'\widehat{\Sigma}^{-1}_{\widehat{T}_{p^*}} X(\widehat{T}_{p^*})| \\
&\leq \sup_{T_{p^*}\in\Theta_{p^*}} \|\widehat{c}_1(s_0,T_{p^*}) - \widehat{c}_1(s,T_{p^*})\|_2 \|\widehat{\Sigma}^{-1}_{T_{p^*}} X(T_{p^*})\|_2 \\
&\leq \sup_{T_{p^*}\in\Theta_{p^*}} \|\widehat{c}_1(s_0,T_{p^*}) - \widehat{c}_1(s,T_{p^*})\|_2 \left(\|\Sigma^{-1}_{T_{p^*}} X(T_{p^*})\|_2 + C\right).
\end{aligned}$$

Then we have to see that the sample covariance functions $\widehat{c}_1(\cdot,t_j)$ are pointwise equicontinuous for every $t_j \in [0,1]$. This follows from the a.s. convergence; these functions converge uniformly to $c_1(\cdot,t_j)$ by Equation 16 and then, for $s_0 \in [0,1]$,

$$\begin{aligned}
|\widehat{c}_1(s_0,t_j) \;-\; \widehat{c}_1(s,t_j)\;| \;&=\; |\;\widehat{\mathrm{cov}}(X_n(s_0) - X_n(s), X_{n-1}(t_j))\;| \\
&\leq\; |\;\widehat{\mathrm{cov}}(X_n(s_0) - X_n(s), X_{n-1}(t_j)) - \mathrm{cov}(X_n(s_0) - X_n(s), X_{n-1}(t_j))\;| \\
&\quad + |\;\mathrm{cov}(X_n(s_0) - X_n(s), X_{n-1}(t_j))\;| \\
&\leq\; 2\sup_{s\in[0,1]} |\;\widehat{c}_1(s,t_j) - c_1(s,t_j)\;| + \varepsilon' \;<\; \varepsilon'' + \varepsilon' \;<\; \varepsilon,
\end{aligned}$$

for n large enough and whenever $|s - s_0| < \delta$.

(b) The statement is equivalent to see that the real valued random variables $Z_n = \left\|\widehat{X}_{n,\widehat{T}_{p^*}} - X_{n,T^*}\right\|^2$ converge to 0 in $L^1_{\mathbb{R}}(\Omega)$. From part a) we know that they converge a.s. to zero, so it only remains to check that the sequence Z_n is uniformly integrable (Vitali's convergence theorem). Since

$$0 \;\leq\; \left\|\widehat{X}_{n,\widehat{T}_{p^*}} - X_{n,T^*}\right\|^2 \;\leq\; \left(\left\|\widehat{X}_{n,\widehat{T}_{p^*}}\right\| + \left\|X_{n,T^*}\right\|\right)^2 \;\leq\; \left(\left\|\widehat{X}_{n,\widehat{T}_{p^*}}\right\| + C\right)^2$$

with $C < \infty$, it is equivalent to prove that both $\|\widehat{X}_{n,\widehat{T}_{p^*}}\|$ and $\|\widehat{X}_{n,\widehat{T}_{p^*}}\|^2$ are uniformly integrable sequences. In order to check this, it is enough to see that $\sup_m \mathbb{E}\|\widehat{X}_{n,\widehat{T}_{p^*}}\|^\eta$, $\sup_m \mathbb{E}\|\widehat{X}_{n,\widehat{T}_{p^*}}\|^{2\eta} < \infty$, where m is the sample size, for some $\eta > 1$ (de la Vallée-Poussin's theorem). We have seen, using Lemma 1, that the supremum of $\|\widehat{c}_1(s,T_{p^*})'\widehat{\Sigma}^{-1}_{T_{p^*}} - c_1(s,T_{p^*})'\Sigma^{-1}_{T_{p^*}}\|_2$ for $(s,T_{p^*}) \in [0,1]\times\Theta_{p^*}$ goes a.s to zero, then we can bound these norms as

$$\begin{aligned}
\|\widehat{X}_{n,\widehat{T}_{p^*}}\|^\eta \;&=\; \left(\sup_{s\in[0,1]} |\widehat{c}_1(s,\widehat{T}_{p^*})'\widehat{\Sigma}^{-1}_{\widehat{T}_{p^*}} X_{n-1}(\widehat{T}_{p^*})|\right)^\eta \\
&\leq\; \|X_{n-1}(\widehat{T}_{p^*})\|_2^\eta \left(\sup_{s\in[0,1]} \|\widehat{c}_1(s,\widehat{T}_{p^*})'\widehat{\Sigma}^{-1}_{\widehat{T}_{p^*}}\|_2\right)^\eta \\
&\leq\; \|X_{n-1}(\widehat{T}_{p^*})\|_2^\eta \left(\sup_{s\in[0,1],T_{p^*}\in\Theta_{p^*}} \|c_1(s,T_{p^*})'\Sigma^{-1}_{T_{p^*}}\|_2 + \varepsilon\right)^\eta
\end{aligned}$$

$$\leq \quad C\|X_{n-1}(\widehat{T}_{p^*})\|_2^{\eta},$$

where $C < \infty$, since the function involved in the supremum is a continuous function on the compact space $[0, 1] \times \Theta_{p^*}$. Equivalently for $\|\widehat{X}_{n,\widehat{T}_{p^*}}\|^{2\eta}$. To conclude the result we can use the same reasoning as in the proof of Theorem 2 of Berrendero et al. (2017a).

Acknowledgements

We are very grateful to Prof. Dr. Claudia Klüppelberg, Antonio Cuevas, José Ramón Berrendero and Alexander Aue for their help, suggestions and criticisms. We furthermore thank Siegfried Hörmann and the Autobahndirektion Süedbayern for providing the datasets. This work has been partially supported by Spanish Grant MTM2016-78751-P and the European Social Fund ("Ayudas para contratos predoctorales para la formación de doctores 2015" and "Ayudas a la movilidad predoctoral para la realización de estancias breves en centros de I+D 2016"). The paper was partly written during the visit of Beatriz Bueno-Larraz as an academic guest researcher at the Technical University of Munich.

References

A. Aue and J. Klepsch. Estimating functional time series by moving average model fitting. *preprint at arXiv:1701.00770[ME]*, 2017.

A. Aue, D. Dubart Norinho, and S. Hörmann. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110 (509):378–392, 2015.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, Boston, 2004. ISBN 1-4020-7679-7. URL `http://opac.inria.fr/record=b1128469`.

P. Bernard. Analyse de signaux physiologiques. *Mémoire Université Catholique Angers*, 1997.

J. R. Berrendero, B. Bueno-Larraz, and A. Cuevas. An RKHS model for variable selection in functional regression. *arXiv preprint arXiv:1701.02512v2*, 2017a.

J. R. Berrendero, A. Cuevas, and J. L. Torrecilla. On the use of reproducing kernel Hilbert spaces in functional classification. *Journal of the American Statistical Association*, page (to appear), 2017b. doi: 10.1080/01621459.2017.1320287. URL `http://dx.doi.org/10.1080/01621459.2017.1320287`.

D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Springer, New York, 2000.

A. Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2014.

S. Hörmann and P. Kokoszka. Weakly dependent functional data. *Annals of Statistics*, 38(3):1845–1884, 2010.

S. Hörmann, L. Kidziński, and M. Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B*, 77(2):319–348, 2015.

L. Horváth and P. Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators.* Wiley, West Sussex, UK, 2015.

S. Janson. *Gaussian hilbert spaces*, volume 129. Cambridge university press, 1997.

H. Ji and H.-G. Müller. Optimal designs for longitudinal and functional data. *J. Roy. Statist. Soc., B, to appear*, 2016.

H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 16:1–54, 2015.

V. Kargin and A. Onatski. Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, 99:2508–2526, 2008.

J. Klepsch and C. Klüppelberg. An Innovations Algorithm for the prediction of functional linear processes. *Journal of Multivariate Analysis*, 155:252–271, 2017.

J. Klepsch, C. Klüppelberg, and T. Wei. Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics*, 1:128–149, 2017.

P. Kokoszka and M. Reimherr. Determining the order of the functional autoregressive model. *Journal of Time Series Analysis*, 34:116–129, 2013.

M. N. Lukić and J. H. Beder. Stochastic processes with sample paths in reproducing kernel hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):pp. 3945–3969, 2001. ISSN 00029947. URL `http://www.jstor.org/stable/2693779`.

F. Mokhtari and T. Mourid. Prediction of continuous time autoregressive processes via the reproducing kernel spaces. *Statistical Inference for Stochastic Processes*, 6(3):247–266, 2003.

E. Parzen. Regression analysis of continuous parameter time series. In *Proceedings of the Fourth Berkeley Symposion on Mathematical Statistics and Probability*, volume 1, pages 469–489, 1961.

N. S. Pillai, Q. Wu, F. Liang, S. Mukherjee, and R. L. Wolpert. Characterizing the function space for bayesian kernel models. *Journal of Machine Learning Research*, 8(Aug):1769–1797, 2007.

B. Pumo. Prediction of continuous time processes by $C[0,1]$–valued autoregressive process. *Statistical Inference for Stochastic Processes*, 1(3):297–309, 1998. doi: 10.1023/A:1009951104780.