ESAIM: PROCEEDINGS AND SURVEYS, February 2025, Vol. 80, p. 72-87
Editors: Pierre Calka, Clément Marteau - Journées MAS 2022, Université de Rouen

SOME RECENT DEVELOPMENTS ON FUNCTIONAL DATA ANALYSIS *

Jairo Cugliari[1], Emilie Devijver[2], Anouar Meynaoui[3] and Raphaël Mignot[4]

Abstract. Recent contributions to functional data analysis are presented. Various problems are considered including the definition of the barycenter of multivariate functional data and the adaptive nonparametric estimation in the functional linear model with functional output.

Résumé. Des contributions récentes pour l'analyse de données fonctionnelles sont présentées. Les problèmes considérés incluent la définition du barycentre pour des données fonctionnelles multivariées, et l'estimation adaptative non paramètrique dans le modèle fonctionnel linéaire à sortie fonctionnelle.

Introduction

This paper presents different recent direction in the study of functional data. Functional data analysis (FDA) is a branch of statistics that deals with the analysis of continuous, often high-dimensional, data. It is an important tool for understanding and analyzing data from various fields such as engineering, biology, and economics. The goal of FDA is to model and analyze data in a way that is both mathematically and statistically rigorous, while also being practical and efficient. One of the key features of FDA is its focus on the practical point of view. This means that the methods developed in this field were traditionally easy to implement and were able to handle real-world data, which is often noisy and may have missing values. In addition, FDA methods were able to provide useful insights and predictions that were applied in a practical setting.

The results presented in the text were the subjects of a series of talks given by the authors in the session "Functional data analysis" during the Journées MAS 2022 in Rouen.

The paper is organized as follows.

Section 1 contains the contribution by Jairo Cugliari and introduces the main concepts and advantages of FDA. Section 2 contains the contribution by Raphaël Mignot and deals with a new definition of the barycenter for multivariate functional data using the signature method. Section 3 contains the contribution by Anouar Meynaoui based on his paper [8] (joint with Gaëlle Chagny and Angelina Roche) about the adaptive nonparametric estimation in the functional linear model with functional output.

1. Modeling of functional data

In this section, we provide an overview of FDA, introducing the main notions and problems. An important aspect of FDA is the modeling of functional data. This involves fitting a statistical model to the data in order to understand the underlying patterns and relationships. There are many different types of models that can

* ***PART OF THIS WORK WAS SUPPORTED BY AGENCE NATIONALE DE LA RECHERCHE PROJECT ANR-20-IADJ-0003.***

[1] Laboratoire ERIC, Université de Lyon 2, Bron, France
[2] CNRS, Université Grenoble Alpes, Grenoble INP, LIG, 38000 Grenoble, France
[3] Université Rennes and Inria, CNRS, IRMAR-UMR 6625, 35000 Rennes, France
[4] Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France

Article published online by EDP Sciences and available at https://www.esaim-proc.org or https://doi.org/10.1051/proc/202580072

be used in FDA, such as linear models, nonlinear models, and mixed models. The choice of model will depend on the specific characteristics of the data and the goals of the analysis. One of the main challenges in FDA is the high dimensionality of functional data. For example, a time series with 100 time points is already 100-dimensional, and this can quickly become infeasible to work with using traditional methods. Therefore, FDA often relies on dimension reduction techniques, such as functional principal component analysis (FPCA), to reduce the dimensionality of the data while still capturing the important features.

1.1. **Some definitions.**

A random element X that takes values in some functional space, say $\mathcal{F}$ is usually called a random function or a functional random variable. Some popular functional spaces are $\mathcal{F} = C[0,1]$, the space of continuous functions on the unit interval, or $\mathcal{F} = L_2([0,1])$ the space of (classes of) functions of finite energy over the unit interval. Then, $X = \{X(t), t \in T\}$ is a random function and $X = \{X(t), t \in T\}$ is one realization of X. A functional dataset is a collection of n random functions $X_1(t), \ldots, X_n(t)$. Although serial correlation and spatial dependence are usual, in most of the cases FDA theory assumes that the data is produced by identical and independent copies of X

A major concern is how to model random data from discrete measurement, possibly observed with noise. In practice, one only disposes a finite sampling $\mathbf{x} = \{x(t_j), j = 1, \ldots, N\}$ observed eventually with noise, from the trajectory $x(t)$ of the random function $X(t)$. Then, one wishes to approximate $x(t)$ from the discrete measurements. A popular choice is to develop $x(t)$ over the elements of a L_2 basis $\phi_1(t), \ldots, \phi_k(t), \ldots$, that is to write

$$x(t) = \sum_k \tilde{y}_k \phi_k(t) \tag{1}$$

where the coefficients $\tilde{y}_k =< x(t), \phi_k(t) >$ are the coordinates resulting of projecting the function x on each of the elements of the basis.

1.2. **Spline smoothing.**

B-spline bases are a popular choice for handling data that exhibits smooth trajectories, thanks to their favorable computational properties. B-splines constitute a basis system that is well-suited for representing splines, such as cubic splines, which are third-order polynomial piecewise functions. The smooth connections occur at specific points referred to as knots, ensuring the continuity of the second-order derivative. A compelling attribute of B-splines is the compact support exhibited by their elements. This property not only provides favorable localization characteristics but also facilitates efficient computation.

Another important property is that at each point of the domain t, the sum of the spline functions is 1. Potential boundary conditions at the ends of the support can be addressed by selecting a relatively large number of basis functions and subsequently employing a functional version of principal components.

1.3. **Functional principal component analysis.**

Similar to multivariate data analysis, Functional Principal Components Analysis (FPCA) offers a means of reducing data dimensionality while managing a controlled loss of information. Given that functional data analysis involves data of infinite dimension, it becomes essential to carefully approach the concept of dimension reduction. Specifically, our objective is to achieve a representation of functions, as shown in (1), using a relatively small set of basis functions that now depend on the data.

Furthermore, if we also seek the basis functions to constitute an orthonormal system, the solution lies in the eigen decomposition of the corresponding covariance operator (analogous to the covariance matrix in the functional context, [25]). However, the problem is that these elements are functions and so of infinite dimension. The solution is to represent themselves into a functional basis system (for instance the one presented on the precedent paragraphs). Thus, the initial curve $x(t)$ can be approximated in the eigenfunctions basis system:

$$x_i(t) = \sum_{k=1}^{p} y_{ik} \xi_k(t)$$

where the number p of eigenfunctions, expected to be relatively small, will be chosen such according to the error of approximation of the curves.

Since the representation system may be non-orthogonal then it can be shown that the inner product needed in FPCA is connected to the properties of the representational basis system. Then, the notion of dimension reduction can be understood when one compares the lower number of eigenfunctions with respect to the number of basis functions needed to represent an observation.

2. Barycenter of time series: a new approach using the signature method

Introduction

The statistical analysis of multivariate time series is a difficult task, with many applications in various domains such as in finance, environment or health. One major issue is to encode in a relevant way the temporal dependency inherent to each component and also nonlinear dependencies between components. A promising approach recently introduced aims at using the so-called signature method [11], initialized first by K-T. Chen during the 1950s [10] in the context of control theory and since then enhanced through the development of rough path theory [19].

The signature method is often coupled with usual multivariate time series analysis strategies, as a first step of encoding the inter and intra components dependencies. Note that this method can be used on time series but, more generally, on any data that have an intrinsic order $X := \{X(u_1), \dots, X(u_l)\}$ where u_i could be any type of index parameter (e.g. time, localization). This approach has shown to be truly efficient for many applications, such as the recognition of Chinese handwriting [28], bipolar disorder detection [22] and also in oceanography [27]. Here, we focus ourselves on an important task in statistical learning: the averaging of time series. We show that our approach based on the signature method has advantageous properties regarding both computational and statistical aspects.

2.1. Multivariate processes and the signature transform

2.1.1. *The signature transform*

Consider $X : [0, 1] \to \mathbb{R}^d$ a continuous multivariate process of bounded variation. Let $0 \leq t_1 < t_2 \leq 1$ and m be an integer. Denote $\otimes$ the tensor product. The iterated integrals signature, or simply the signature, of order m on $[t_1, t_2]$ of process X is

$$S^{(m)}_{[t_1,t_2]}(X) := \int_{t_1<u_1<\cdots<u_m<t_2} \cdots \int dX_{u_1} \otimes \cdots \otimes dX_{u_m}. \tag{2}$$

This object is an m-way tensor of dimension d: $S^{(m)}_{[t_1,t_2]}(X) \in (\mathbb{R}^d)^{\otimes m}$. Intuitively, each coefficient of this tensor can be thought as an association measure between m components of the considered multivariate process. The signature of multivariate process X is the infinite collection of signatures of all orders:

$$\mathbf{S}_{[t_1,t_2]}(X) := \left\{1, S^{(1)}_{[t_1,t_2]}(X), S^{(2)}_{[t_1,t_2]}(X), \dots\right\}$$

where we use the following convention $S^{(0)}_{[t_1,t_2]}(X) = 1$. Observe that this is a collection of tensors of increasing shapes: $S^{(1)}_{[t_1,t_2]}(X)$ is a vector, $S^{(2)}_{[t_1,t_2]}(X)$ is a matrix, $S^{(3)}_{[t_1,t_2]}(X)$ a 3-way tensor, etc.

Example 1: let us consider $X : [0, 1] \to \mathbb{R}^d$ a linear process. We find that for any set of indices $(i_1, \dots, i_m)$ with i_j an integer between 1 and d:

$$\left[S^{(m)}_{[0,1]}(X)\right]_{(i_1,\dots,i_m)} = \frac{1}{m!} \prod_{j=1}^{m} (X^{(i_j)}(1) - X^{(i_j)}(0)) \in \mathbb{R}.$$

Example 2: let $X : [0,1] \to \mathbb{R}$ be a univariate differentiable process. The computation of the signature is then straightforward and we obtain for any order m,

$$S^{(m)}_{[0,1]}(X) = \frac{1}{m!}(X(1) - X(0))^m.$$

The signature of a univariate process is thus not very useful since it depends only on the increment over the whole interval of definition. In practice, we will deal only with multivariate processes. In some cases when X is univariate, it might be interesting to use as input the bi-dimensional process $(t, X(t))$, which is called time augmentation in the literature.

Note that in some cases, there is a geometric interpretation of the coefficients of the signature. For instance, the signature of order 2 can be related to the so-called Lévy area of the process. The reader can find figures and details in [11].

2.1.2. *Fundamental properties of the signature transform*

The signature satisfies two properties. First, it is an intrinsic description, characterizing under some assumptions a multivariate process if we ignore shifts and time reparametrizations. Moreover, the space of signatures is a non compact nilpotent Lie group equipped with $\boxtimes$ operation which is an extension of the tensor product $\otimes$ in the following sense: $a \boxtimes b = (c_0, c_1, \dots)$ where $c_K := \sum_{k=0}^{K} a_k \otimes b_{K-k}$ for any integer K. The $\boxtimes$ operation is related to the concatenation of two processes through the so-called Chen relation. Denote $\star$ the concatenation of two processes, i.e. let $X : [t_1, t_2] \to \mathbb{R}^d$ and $Y : [t_2, t_3] \to \mathbb{R}^d$ be two continuous processes. Then the concatenation is the process such that for any $t \in [t_1, t_3]$,

$$(X \star Y)(t) := \begin{cases} X(t) & \text{if } t_1 \le t \le t_2 \\ Y(t) - Y(t_2) + X(t_2) & \text{if } t_2 \le t \le t_3 \end{cases}.$$

Then, Chen relation is

$$\mathbf{S}_{[t_1,t_3]}(X \star Y) = \mathbf{S}_{[t_1,t_2]}(X) \boxtimes \mathbf{S}_{[t_2,t_3]}(Y). \tag{3}$$

Iterated integrals are in practice computed numerically up to a chosen order. Note that the signature up to order L is of dimension

$$\sum_{k=0}^{L} d^k = \frac{d^{L+1} - 1}{d - 1}$$

if $d \neq 1$ and $L + 1$ if $d = 1$. The finite collection of signature of order less or equal to L is called the truncated signature space. It is a subspace of $T^L(\mathbb{R}^d) := \bigoplus_{k=0}^{L} (\mathbb{R}^d)^{\otimes k}$. The space $(T^L(\mathbb{R}^d), +, \cdot, \boxtimes)$ is a non commutative Lie algebra. Denote $T^L_1(\mathbb{R}^d) := \{a = (a_0, \dots, a_L) \in T^L(\mathbb{R}^d) : a_0 = 1\}$. We have that $T^L_1(\mathbb{R}^d)$ is a Lie group. We can define the exponential, logarithm and inverse transforms in the following way. Let $g \in T^L(\mathbb{R}^d)$ and $(1 + g) \in T^L_1(\mathbb{R}^d)$, then

$$\mathrm{Exp}(g) := \sum_{i=0}^{L} \frac{g^{\boxtimes i}}{i!} \quad , \quad \mathrm{Log}(1+g) := \sum_{i=1}^{L} (-1)^{i-1} \frac{g^{\boxtimes i}}{i} \quad \text{and} \quad (1+g)^{-1} := \sum_{i=0}^{L} (-1)^i g^{\boxtimes i}. \tag{4}$$

Denote the p-variation norm of $X : I \to \mathbb{R}^d$, where $p \ge 1$ and I is an interval of $\mathbb{R}$, as

$$\|X\|_{p\text{-var}} := \left(\sup_{\mathcal{D}} \sum_{t_k \in \mathcal{D}} d(X(t_k), X(t_{k-1}))^p \right)^{1/p}$$

where $\mathcal{D}$ ranges over all finite partitions of I and d is any distance on $\mathbb{R}^d$. Then, denote

$$\mathcal{C}^{p\text{-var}}([0,1], \mathbb{R}^d) := \left\{ X \in \mathcal{C}([0,1], \mathbb{R}^d) : \|X\|_{p\text{-var}} < \infty \right\}$$

and $G^L(\mathbb{R}^d)$ the following group:

$$G^L(\mathbb{R}^d) := \{ \mathbf{S}^{(\le L)}_{[0,1]}(X) : X \in \mathcal{C}^{1\text{-var}}([0,1], \mathbb{R}^d) \}$$

where we have denoted by $\mathbf{S}_{[0,1]}^{(\leq L)}(X)$ the finite collection of signatures up to order L. Any element $a \in T_1^L(\mathbb{R}^d)$ is an element of $G^L(\mathbb{R}^d)$ if and only if it satisfies a specific property called the shuffle product. The signature space is in fact a sub-Riemannian manifold and for a more comprehensive overview of the theory of iterated integrals signature, the reader can refer itself to [14, Chapter 7].

The iterated integrals (2) are defined for multivariate continuous processes. In practice, data is sampled discretely. The resulting discrete multivariate process is interpolated. Often, we use the linear interpolation. The obtained continuous process is differentiable almost everywhere and Equation (2) is a Riemann-Stieltjes integral:

$$S_{[t_1,t_2]}^{(m)}(X) = \int_{t_1<u_m<t_2} \cdots \left(\int_{t_1<u_1<u_2} \dot{X}_{u_1} du_1 \right) \otimes \cdots \otimes \dot{X}_{u_m} du_m$$

where we denote $\dot{X}_t := \frac{d}{dt} X(t)$. Note that the signature can be defined in a more general setting than processes of bounded variations ($\|X\|_{1\text{-var}} < \infty$). The signature can be defined with Itô or Stratonovich integration and applied for instance to Brownian motions.

2.2. **Barycenter of multivariate time series and the signature transform**

Let $(X_i)_{1\leq i\leq N}$ be a dataset of N multivariate time series. The goal is to define a notion of barycenter of this set of time series. As an intermediary step, we define the barycenter of the corresponding signature transforms. Let's say each time series is of dimension $d \times l$ where d is the number of components and l is the number of timestamps. From this section onwards, the signature of discrete data $X = (X(t_1), \ldots, X(t_l))$ is the signature of the linearly interpolated data, as presented in the previous subsection. For clarity purposes, we set $t_1 = 0$ and $t_l = 1$ and the interval is not denoted $\mathbf{S}(X) := \mathbf{S}_{[0,1]}(X)$. We want to define a barycenter of N signatures $\{\mathbf{S}(X_1), \ldots, \mathbf{S}(X_N)\}$ with corresponding weights $(w_i)_{1\leq i\leq N}$. One of the main features of the signature method is that time series of different lengths l can be compared: only the number of components d has to be identical. The obtained signature in both case is an element of $T^m(\mathbb{R}^d)$. We present three approaches that all have as a starting point the Lie group nature of the space of signatures. Denote G the space of signatures and $\mathfrak{g}$ its associated Lie algebra. Denote $\mathrm{Exp} : \mathfrak{g} \to G$ and $\mathrm{Log} : G \to \mathfrak{g}$ the exponential and logarithm mappings as defined in the previous section.

2.2.1. *Approach 1: naive method*

The structure of Lie group is such that the Euclidean barycenter of elements of G defined as $\bar{\mathbf{S}} := \frac{1}{N}\sum_{i=1}^N \mathbf{S}(X_i)$ is possibly not an element of G. In order to obtain a barycenter that lives in G, a first strategy is to compute the Euclidean barycenter of the elements mapped into the Lie algebra $\mathfrak{g}$ and then map them back into the Lie group G:

$$\bar{\mathbf{S}}_{\text{naive}} := \mathrm{Exp}\left(\sum_{i=1}^N w_i \,\mathrm{Log}\, \mathbf{S}(X_i) \right).$$

This approach has one main drawback: this barycenter is not invariant under right or left multiplication, i.e. it exists a process Y such that

$$\mathrm{Exp}\left(\sum_{i=1}^N w_i \,\mathrm{Log}(\mathbf{S}(X_i) \boxtimes \mathbf{S}(Y)) \right) \neq \mathrm{Exp}\left(\sum_{i=1}^N w_i \,\mathrm{Log}\, \mathbf{S}(X_i) \right) \boxtimes \mathbf{S}(Y). \tag{5}$$

As mentioned in Section 2.1.2, multiplication $\boxtimes$ in G is related to the concatenation operation in the space of processes. Using Equation (3), this invariance is illustrated in process space in Figure 1. This barycenter lead us to non intuitive phenomenons that we wish to avoid. This motivates our next approach.

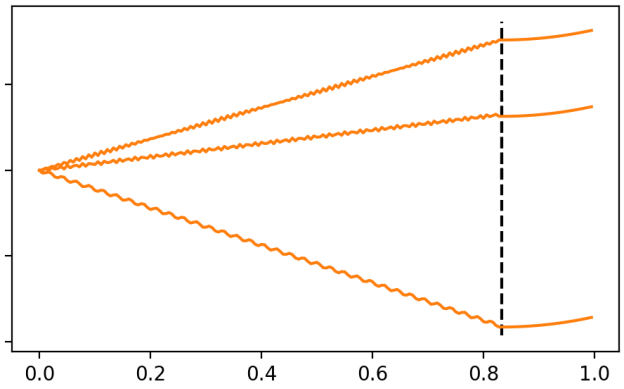

FIGURE 1. Right translation: three curves X_1, X_2, X_3 (left side of the vertical dotted line) upon which a curve Y is concatenated (right side). An averaging method $\mathbf{m}$ is right-invariant if $\mathbf{m}(\{X_i \star Y\}_{1\leq i\leq n}) = \mathbf{m}(\{X_i\}_{1\leq i\leq n}) \star Y$.

2.2.2. *Approach 2: group exponential barycenter*

This approach is an adaptation of [21, Algorithm 1] to the context of the signature method. In this paper, the goal is to approach the so-called Riemannian center of mass (or Fréchet mean) of the set of points:

$$\arg\min_{\mu} \sum_{i=1}^{N} w_i d(z_i, \mu)^{\alpha}.$$

Observe that in a Euclidean space, with $\alpha = 2$ the Fréchet mean is the usual Euclidean mean and with $\alpha = 1$ it is the median.

The barycenter is obtained iteratively. First, we initialize the value of the barycenter $\bar{\mathbf{S}}_{(0)}$. For instance, it can be set to be an observation randomly drawn. Then, the value of the barycenter at step k is updated through the following formula:

$$\bar{\mathbf{S}}_{(k+1)} := \bar{\mathbf{S}}_{(k)} \boxtimes \mathrm{Exp}\left(\sum_{i=1}^{N} w_i \,\mathrm{Log}\left(\bar{\mathbf{S}}_{(k)}^{-1} \boxtimes \mathbf{S}(X_i)\right)\right)$$

where the notation $^{-1}$ refers to the element defined in Equation (4). This algorithm terminates when the stopping criterion is reached. This criterion can be a maximum number of iterations or a threshold on the distance between two consecutive iterations $\bar{\mathbf{S}}_{(k)}$ and $\bar{\mathbf{S}}_{(k+1)}$. Under the assumptions that the signature data $(\mathbf{S}(X_i))_{1\leq i\leq N}$ is sufficiently close to the neutral element and that the initialization $\bar{\mathbf{S}}_{(0)}$ is sufficiently close to the data, this procedure converges to a solution [21, Corollary 5]. In numerical experiments on signature data, we have observed that most of the time, only around a dozen of iterations is necessary to converge.

2.2.3. *Approach 3: optimization on time series space*

In the two previous approaches, the obtained barycenter is a signature. To obtain a process X corresponding to the barycenter $\bar{\mathbf{S}}$, i.e. such that $\mathbf{S}(X) = \bar{\mathbf{S}}$, a reconstruction is necessary. This reconstruction is not exact and requires the signature up to a large order [9]. In practice, this is often not feasible, especially when d is larger than 5.

In the following approach, the obtained barycenter is a time series. Let X be a time series of dimension $d \times \tilde{l}$ where $1 \leq \tilde{l} \leq l$ and let L be a truncation order. The barycenter is the minimizer of the following objective function

$$f(X) := \sum_{i=1}^{N} \sum_{j=1}^{L} w_i d(S^{(j)}(X), S^{(j)}(X_i))^2$$

where d is a distance to be chosen, e.g. $d(g,h) := \|g-h\|_F$ with $\|.\|_F$ the Frobenius norm, or $d(g,h) := \|g^{-1} \boxtimes h\|_{\mathrm{CC}}$ where $\|.\|_{\mathrm{CC}}$ is the Carnot-Carathéodory norm. This objective function can be optimized using for instance gradient descent method. Numerically, this can easily be performed using automatic differentiation.

In order to obtain better performances and robustness, various parameters can be tuned. The parameter $\tilde{l}$ can be changed to avoid overfitting or underfitting. Multiple starting points can be tried in order to increase the likelihood of finding the global minimum.

2.2.4. *Comparison of the three approaches*

The three presented approaches have different use cases: depending on whether it is necessary to obtain a time series barycenter or not. For instance, in a clustering task it is not necessary to get a barycenter in processes space, as we will see in the next section. But a barycenter in time series space might be easier to analyze, especially when it comes to outliers, sensitivity analysis and interpretations of figures.

2.3. **Numerical experiments**

2.3.1. *K-means algorithm*

Our various notions of barycenter can be coupled with ubiquitous machine learning algorithms. For clustering tasks, the reference is the K-means method. We wish to separate N samples $(Z_i)_{1\leq i\leq N}$ into K groups. Note that samples can be points in $\mathbb{R}^d$ or time series. The procedure is the following.

(1) Initialize $\mu_1, \ldots, \mu_K$ centers of each of the K groups.
(2) While stopping criterion is not satisfied, do
- Assignation step: for any $1 \leq i \leq N$, Z_i is assigned to the group that has the nearest center according to a chosen distance.
- Update step: each center μ_j is re-computed in order to be the barycenter of the points in its group (after assignation step).

Thus, two parameters appear in this algorithm: a metric and a notion of barycenter. Through the previous steps, the following *inertia* is minimized

$$I := \arg\min_{\mathbf{C}} \sum_{k=1}^{K} \sum_{j:X_j \in C_k} d(Z_j, \mu_k)^2$$

where $\mathbf{C} := \{C_1, \ldots, C_K\}$ is the clustering and d the chosen distance.

In order to evaluate the performances of a clustering procedure, various metrics exist. A ubiquitous one is the rand index (RI). Assume that we have the *true* clustering $\mathbf{C} = \{C_1, \ldots, C_K\}$. Denote $\hat{\mathbf{C}} = \{\hat{C}_1, \ldots, \hat{C}_K\}$ the obtained clustering. Let a be the number of pair of samples (Z_i, Z_j) that belong to the same group in $\mathbf{C}$ and to the same group in $\hat{\mathbf{C}}$. Let b be the number of pair of samples that are in two separate groups in $\mathbf{C}$ and in two separate groups in $\hat{\mathbf{C}}$. The rand index is defined as

$$\mathrm{RI} := \frac{a+b}{\binom{N}{2}}.$$

Observe that if $\mathbf{C} = \hat{\mathbf{C}}$ then $\mathrm{RI} = 1$. This metric decrease to zero as the two clustering get *different*.

2.3.2. *Application with the signature transform*

We use the PenDigits [1] dataset where each sample is the recording of a person writing a digit with a pen on a digital screen. The $N = 10992$ instances are made of the two dimensional coordinates of the position of the pen resampled to only 8 datapoints, each consecutive position having a constant time step of 100ms. Thus, each time series is of dimension $(d \times l) = (2 \times 8)$. The true label associated to each sample is known. The goal is to classify the different digits. Note that the optimal number of groups might not be 10: a digit can be represented in various ways such as the digit 4 (in one stroke or in two strokes). We perform K-means with four different values of K: 8, 10, 12 and 14 groups. The stopping criterion is a fixed value of iterations: 10.

Averaging method	Distance
Euclidean mean	Euclidean
Dynamic Time Warping Barycenter Averaging [23]	Dynamic Time Warping
Naive (Sec. 2.2.1)	Euclidean
Group exponential barycenter (Sec. 2.2.2)	Euclidean
Optimization on time series space (Sec. 2.2.3)	Euclidean

TABLE 1. Parameters for the K-means algorithm. First two methods do not use the signature.

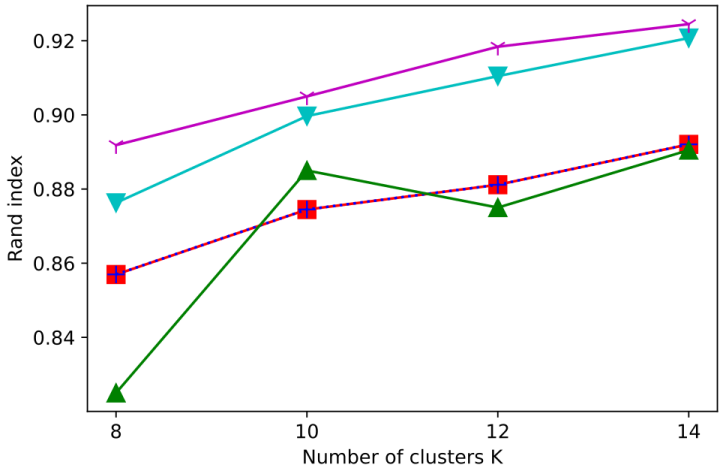

FIGURE 2. Rand Index for each of the averaging method of Table 1. Cyan triangle is Euclidean mean. Violet Y is Dynamic Time Warping Barycenter Averaging. Red square is naive. Blue cross is group exponential barycenter. Green upright triangle is optimization on time series space.

Five strategies are compared: see Table 1. The method Dynamic Time Warping Barycenter Averaging [23] is computed with `tslearn` Python package. The iterated integrals signature is computed with `signatory` [17] Python package up to the order 8.

Results are shown on Figure 2. First, the three strategies that involve the signature have slightly worse performances than the two others. Notice the scale on the y-axis: performances are very close to each other. As shown in Table 1, the chosen distance of the three signature strategies is the Euclidean distance. This might not be optimal, since it is not designed for the signature space. However, it is the most time efficient. Notice that the blue and the red curves are almost indistinguishable. This is not the case in general and actually the clusters obtained with the two corresponding methods are different.

It is known that the K-means algorithm assumes that clusters have an isotropic (circular) shape. We do not have any guarantee regarding that matter in signature space. Other clustering strategies coupled with the signature might yield better results.

Note that we have computed another metric: the Adjusted Mutual Information. Conclusions are the same than those presented here with the Rand Index.

2.4. **Conclusion**

Having defined notions of barycenter for the iterated integrals paves the way for the generalization of classical statistical learning algorithms. We have illustrated this with the K-means algorithm and other are currently explored, such as the Principal Component Analysis. Other approaches to define a barycenter using the signature transform can be investigated and are also under study.

3. Adaptive nonparametric estimation in the functional linear model with functional output

3.1. Objective of the work

Assume that we observe an i.i.d. sample $\{(X_i, Y_i), i = 1, \dots, n\}$ of a couple of functional random variables (X, Y). For simplicity, X and Y are assumed to take values in the functional space $\mathbb{H} = L^2([0,1])$ (the space of square integrable functions on $[0,1]$)[1], with its usual scalar product. The link between the variable of interest Y and the functional covariate X is linear. Meaning that, there exists an operator $S \in \mathcal{L}(\mathbb{H})$, the space of continuous linear operators on $\mathbb{H}$, such that

$$Y = SX + \varepsilon,$$

where $\varepsilon \in \mathbb{H}$ stands for an (unobserved) noise. The functional variables X and ε are supposed to be both centered, and independent. The noise ε is square integrable, $\sigma_\varepsilon^2 = \mathbb{E}\|\varepsilon\|^2 < \infty$. The *slope operator* S is supposed to be an integral operator and we denote by $\mathcal{S} \in L^2([0,1]^2)$ its kernel:

$$\begin{array}{rccl} S : \mathbb{H} & \longrightarrow & \mathbb{H} \\ f & \longmapsto & \displaystyle\int_0^1 \mathcal{S}(s,\cdot) f(s) ds. \end{array}$$

The aim is to estimate the unknown operator S (or its kernel $\mathcal{S}$) from the sample $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ and to study the theoretical properties of the estimators.

3.2. Estimation method

3.2.1. *Notations*

We introduce some notations that will be used in the following. We denote by $\mathcal{L}_2(\mathbb{H})$ the subspace of Hilbert-Schmidt operators on $\mathbb{H}$ with its usual Hilbert-Schmidt norm defined for any operator $T \in \mathcal{L}_2(\mathbb{H})$ as follows

$$\|T\|_{\mathrm{HS}} = \left(\sum_{j=1}^{\infty} \|T\phi_j\|^2 \right)^{1/2},$$

where $(\phi_j)_{j\geq 1}$ is a Hilbertian basis of $\mathbb{H}$. This definition is independent of the Hilbertian basis choice. Note also that an integral operator is Hilbert-Schmidt if and only if the associated kernel is square integrable. This means that by assumptions, our target operator S is Hilbert-Schmidt. We will also define the covariance and cross-covariance operators that play a key role in the estimation procedure. To do so, we first define the tensor product between two elements a and b of $\mathbb{H}$ as

$$\begin{array}{rccl} b \otimes a : & \mathbb{H} & \longrightarrow & \mathbb{H} \\ & f & \longmapsto & \langle a, f\rangle b. \end{array}$$

The covariance operator of X, denoted by Γ is the operator defined by

$$\begin{array}{rccl} \Gamma : & \mathbb{H} & \longrightarrow & \mathbb{H} \\ & f & \longmapsto & \mathbb{E}[X \otimes X(f)] = \mathbb{E}[\langle X, f\rangle X]. \end{array}$$

Note that the covariance operator is a natural extension of the covariance matrix, in the infinite dimensional framework. The eigenelements of Γ are denoted by $(\lambda_j, \varphi_j)_{j\geq 1}$. Furthermore, the cross-covariance operator Δ of (X, Y) is given by

$$\begin{array}{rccl} \Delta : & \mathbb{H} & \longrightarrow & \mathbb{H} \\ & f & \longmapsto & \mathbb{E}[Y \otimes X(f)] = \mathbb{E}[\langle X, f\rangle Y]. \end{array}$$

[1] In other words, for every event ω, $X(\omega)$ belongs to $\mathbb{H}$ (the same holds for Y).

Empirical counterparts of Γ and Δ, respectively denoted by Γ_n and Δ_n will be needed in the definition of our estimators. These operators are naturally defined on $\mathbb{H}$ by

$$\Gamma_n = \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i \quad \text{and} \quad \Delta_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \otimes X_i.$$

To study the behavior of our estimators, we use an optimality risk called the Mean Square Prediction Error (MSPE). The MSPE is defined for a given estimator $\widehat{S}_n$ of S as

$$\mathrm{MSPE}(\widehat{S}_n) = \mathbb{E}\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\|^2,$$

where X_{n+1} is a new observation of X, which is independent of $(X_i, \varepsilon_i), i = 1, \ldots, n$. This risk can also be written as

$$\mathrm{MSPE}(\widehat{S}_n) = \mathbb{E}\left[\|\widehat{Y}_{n+1} - \mathbb{E}\left[Y_{n+1}|X_{n+1}\right]\|^2 \,|(X_i, Y_i)_{i=1,\ldots,n}\right], \tag{6}$$

where $Y_{n+1} = SX_{n+1} + \varepsilon_{n+1}$ and $\widehat{Y}_{n+1} = \widehat{S}_n X_{n+1}$.

3.2.2. *Projection estimation*

Following the model selection device of [5], we construct the estimators by minimizing a contrast function, over finite dimensional subspaces of $\mathcal{L}_2(\mathbb{H})$. Let $(\phi_j)_{j\geq 1}$ be an orthonormal basis of $\mathbb{H}$. We introduce a collection of finite linear subspaces of $\mathcal{L}_2(\mathbb{H})$, called the models and denoted by V_{m_1,m_2} for given m_1, m_2 in $\mathbb{N}\backslash\{0\}$. They are defined as

$$V_{m_1,m_2} = \mathrm{Span}\{\phi_k \otimes \phi_j,\ (k,j) \in [\![1, m_1]\!] \times [\![1, m_2]\!]\}.$$

We also define the contrast function γ_n by

$$\begin{array}{rccl} \gamma_n : & \mathcal{L}_2(\mathbb{H}) & \to & \mathbb{R}_+ \\ & T & \mapsto & 1/n \sum_{i=1}^{n} \|Y_i - T(X_i)\|^2 . \end{array}$$

The operator $\gamma_n : \mathcal{L}_2(\mathbb{H}) \to \mathbb{R}$ is defined in the spirit of other regression contrast introduced (e.g.) by [2, 6] and stands for an empirical version of the risk (6). The estimator $\widehat{S}_{m_1,m_2}$ associated with the model V_{m_1,m_2} is by definition:

$$\widehat{S}_{m_1,m_2} \in \arg\min_{T \in V_{m_1,m_2}} \gamma_n(T). \tag{7}$$

The following proposition gives a condition for the existence of $\widehat{S}_{m_1,m_2}$.

Proposition 1. *Let A and Y_ϕ be the matrices defined as*

$$A = (\langle \Gamma_n \phi_j, \phi_k \rangle)_{(j,k)\in[\![1,m_1]\!]^2} \quad \textit{and} \quad Y_\phi = (\langle \Delta_n \phi_j, \phi_k \rangle)_{(j,k)\in[\![1,m_1]\!]\times[\![1,m_2]\!]}.$$

If A is invertible, then $\widehat{S}_{m_1,m_2}$ in (7) is uniquely defined as

$$\widehat{S}_{m_1,m_2} = \sum_{j=1}^{m_1}\sum_{k=1}^{m_2} \widehat{b}_{j,k}\phi_k \otimes \phi_j,$$

with $\widehat{b} = (\widehat{b}_{j,k})_{(j,k)\in[\![1,m_1]\!]\times[\![1,m_2]\!]}$ is the matrix $\widehat{b} = A^{-1}Y_\phi$.

In the rest of this document, we focus on the basis of principal components. Recall that, by definition, the empirical covariance operator Γ_n is self-adjoint. Moreover, since it is a finite-rank operator, it is also a compact operator. Then, Γ_n is diagonalizable in a Hilbertian basis, denoted by $(\widehat{\varphi}_j)_{j\geq 1}$. We denote by $(\widehat{\lambda}_j)_{j\geq 1}$ its eigenelements, which are sorted in the decreasing order. The $(\widehat{\varphi}_j)_{j\geq 1}$ is called the empirical PCA basis of X. Notice that the operator Γ_n is not invertible, since it has finite rank at most equal to n. This

means that the eigenvalues $(\widehat{\lambda}_j)_{j\geq 1}$ are zero from a given rank. Let us introduce its pseudo-inverse, $\Gamma^{\dagger}_{n,m_1}$, defined for an index $m_1 \in \mathbb{N}\backslash\{0\}$ by

$$\Gamma^{\dagger}_{n,m_1} = \sum_{j=1}^{m_1} \frac{1}{\widehat{\lambda}_j} \widehat{\varphi}_j \otimes \widehat{\varphi}_j,$$

for $m_1 \leq m_{\max}$, with $m_{\max} = \max_{m\geq 1}\{\widehat{\lambda}_m > 0\}$ is the rank from which the eigenvalues are equal to zero, and $\Gamma^{\dagger}_{n,m_1} = \Gamma^{\dagger}_{n,m_{\max}}$ for $m_1 > m_{\max}$.

By choosing the orthonormal basis presented above, the obtained projection estimator can be written as $\widehat{S}_{m_1,m_2} = \widehat{\Pi}_{m_2}\Delta_n\Gamma^{\dagger}_{n,m_1}$, where $\widehat{\Pi}_{m_2}$ is the projection operator onto the finite dimensional subspace $\mathrm{Span}\{\widehat{\varphi}_k, k = 1,\dots,m_2\}$. The estimator $\widehat{S}_{m_1,m_2}$ can be compared to the estimator of [13] which writes $\widehat{S}^{CM}_{m_1} = \Delta_n\Gamma^{\dagger}_{n,m_1}$. Our choice is based on the fact that the initial regression problem comes down to estimate the kernel $\mathcal{S} \in L^2([0,1]^2)$ of the operator S, which brings out two projection dimensions. Our estimate is thus the same as the estimator with "double truncation" of [15] (see their equation (7) p.19), even if they do not introduce it as a minimum of contrast estimator. The definition of $\widehat{S}_{m_1,m_2}$ as an operator that minimizes a contrast allows us to derive non-asymptotic upper-bounds for the prediction error, and to propose a data-driven way to select the best projection dimensions.

3.3. **Upper and lower bounds of the estimation risk**

We provide sharp upper bounds for the estimation risk of the estimator $\widehat{S}_{m_1,m_2}$, for any but fixed $(m_1,m_2) \in (\mathbb{N}\backslash\{0\})^2$. We also establish a lower bound for the prediction risk, to ensure that the collection of estimates is reasonable.

3.3.1. *Assumptions*

To achieve optimal theoretical results, we need to make some assumptions. The first one is a regularity assumption on $S\Gamma^{1/2}$ which is assumed to belong to an ellipsoid space defined by

$$\mathcal{W}^R_{\alpha,\beta} = \left\{ T \in \mathcal{L}_2(\mathbb{H}), \sum_{j=1}^{+\infty}\sum_{r=1}^{+\infty} \eta_\alpha(j)\psi_\beta(r)\langle T(\varphi_j),\varphi_r\rangle^2 \leq R^2 \right\},$$

where $\alpha,\beta > 0$ and for all $\gamma > 0$, the functions η_γ is defined such that $\eta_\gamma(j) \asymp j^\gamma$ or $\eta_\gamma(j) \asymp \exp(j^\gamma)$, and the same for ψ_γ. In the sequel, we speak about the "polynomial case" or the "exponential case". The reason why the regularity concerns the operator $S\Gamma^{1/2}$ is that the considered risk is linked with the Hilbert-Schmidt norm as

$$\mathrm{MSPE}(\widehat{S}_n) = \mathbb{E}\|(\widehat{S}_n - S)\Gamma^{1/2}\|^2_{\mathrm{HS}},$$

where $\Gamma^{1/2}$ is square root of the operator Γ. Other moment and regularity assumptions on X and on the noise ε are needed, but not presented here. For more details on this assumptions one can refer to the paper on the subject [8].

Theorem 1 below gives a sharp upper bound of the maximal prediction risk of the estimator $\widehat{S}_{m_1,m_2}$ with respect to the projection dimensions m_1 and m_2.

Theorem 1. *Assume that we are in the case where the function ψ_β is polynomial with $\beta > 6$ or exponential. Assume also that there exists $\nu > 0$ such that $\lambda_j \leq j^{-1-\nu}$, for any $j \geq 1$. Under the assumptions of Theorem 1, we have the following bound of the non-asymptotic maximal prediction risk of $\widehat{S}_{m_1,m_2}$.*

$$\inf_{\substack{m_1,m_2\in\mathbb{N}\backslash\{0\}\\ m_1\leq n/\ln^2(n)}} \sup_{S\Gamma^{1/2}\in\mathcal{W}^R_{\alpha,\beta}} \mathrm{MSPE}(\widehat{S}_{m_1,m_2}) \leq \inf_{\substack{m_1\in\mathbb{N}\backslash\{0\}\\ m_1\leq n/\ln^2(n)}} \left\{ \sigma^2_\varepsilon \frac{m_1}{n} + \frac{3}{\eta_\alpha(m_1)} \right\} + \frac{c}{n},$$

where c is a positive constant.

In Theorem 1 appears a bias-variance trade-off. Surprisingly, the m_2 parameter does not appear in the upper bound. This phenomenon was also observed by [15]. Indeed, the variance term does not depend on it (the variance is in m_1/n), and we show that the bias tends to 0 when m_2 tends to $+\infty$. The control of the bias requires the use of tools from the perturbation theory [16]. We show that the bound of Theorem 1 is optimal in the minimax sense: it matches the following lower bound.

Theorem 2. *We have the two following convergence decay for the minimax estimation risk, up to a constant* $C > 0$.

(1) If $\eta_\alpha(j) \asymp j^\alpha$ *(polynomial case) then,*

$$\inf_{\widehat{S}_n} \sup_{S\Gamma^{1/2} \in \mathcal{W}^R_{\alpha,\beta}} \mathrm{MSPE}(\widehat{S}_n) \geq C n^{-\frac{\alpha}{\alpha+1}}.$$

(2) If $\eta_\alpha(j) \asymp \exp(j^\alpha)$ *(exponential case) then,*

$$\inf_{\widehat{S}_n} \sup_{S\Gamma^{1/2} \in \mathcal{W}^R_{\alpha,\beta}} \mathrm{MSPE}(\widehat{S}_n) \geq C \frac{(\ln(n))^{1/\alpha}}{n}.$$

3.4. Model selection

We perform adaptive model selection, which does not depend on the unknown smoothness of the model S, but only on the available data. We recall that, for given projection dimensions m_1 and m_2, we estimate the operator S by $\widehat{S}_{m_1,m_2} = \widehat{\Pi}_{m_2} \Delta_n \Gamma^\dagger_{n,m_1}$, where the operators $\widehat{\Pi}_{m_2}$, Δ_n and $\Gamma^\dagger_{n,m_1}$. The idea is to propose a procedure which automatically selects the optimal projection dimensions m_1 and m_2, that is the best estimator in the collection $(\widehat{S}_{m_1,m_2})_{m_1,m_2}$.

According to the result of Theorem 2, we choose $m_2 \to +\infty$ and we select m_1 in the collection $\mathcal{M}_n = \{1, \dots, N_n\}$, where the size of the collection N_n satisfied $N_n \leq \lfloor n/\ln^2(n) \rfloor$ where $\lfloor \cdot \rfloor$ is the floor function, associating to each x in $\mathbb{R}$ the largest integer less or equal to x. Thus, the issue we consider now is the choice of an estimator in the collection $(\widehat{S}_{m_1,\infty})_{m_1 \in \mathcal{M}_n}$, where $\widehat{S}_{m_1,\infty} = \Delta_n \Gamma^\dagger_{n,m_1}$ corresponds in fact to the estimator of [13]. The method we use is derived from the model selection tools developed by [4], as in [6] or [12]. A clear and detailed account is given in [20]. We want to select the "best" estimator in the collection $(\widehat{S}_{m_1,\infty})_{m_1 \in \mathcal{M}_n}$, that is the one which has the smaller risk. Since the risk is unknown in practice, the oracle $m_1^* = \arg\min_{m_1 \in \mathcal{M}_n} \mathrm{MSPE}(\widehat{S}_{m_1,\infty})$ is also unknown, and the risk $\mathrm{MSPE}(\widehat{S}_{m_1,\infty})$ should be replaced by an empirical counterpart. Moreover, the contrast function is an empirical version of the risk, the first idea is to choose $\arg\min_{m_1 \in \mathcal{M}_n} \gamma_n(\widehat{S}_{m_1,\infty})$. However, since the contrast function decreases when m_1 grows, the choice of $\arg\min_{m_1 \in \mathcal{M}_n} \gamma_n(\widehat{S}_{m_1,\infty})$ will lead to the selection of the largest index in the collection $\mathcal{M}_n$. One of the main idea of model selection theory is to introduce a penalty to balance this decrease, usually of the order of the variance. The dimension parameter m_1 is choosen as the one which minimizes a penalized contrast function,

$$\widehat{m}_1 = \arg\min_{m_1 \in \mathcal{M}_n} \left(\gamma_n(\widehat{S}_{m_1,\infty}) + \mathrm{pen}(m_1) \right), \tag{8}$$

where pen is the penalty function defined as $\mathrm{pen} : m_1 \mapsto 8(1+\delta)\sigma_\varepsilon^2 m_1/n$, with $\delta > 0$ a numerical constant that is tuned in practice. Notice that, when m_1 is fixed, $\gamma_n(\widehat{S}_{m_1,m_2})$ decreases with m_2 by definition and $\mathrm{pen}(m_1)$ does not depend on m_2. Thus, $(\widehat{m}_1, +\infty)$ is also a solution of the minimization problem

$$\min_{(m_1,m_2) \in \mathcal{M}_n \times \mathbb{N} \setminus \{0\} \cup \{+\infty\}} \left(\gamma_n(\widehat{S}_{m_1,m_2}) + \mathrm{pen}(m_1) \right).$$

With this writing, the selection procedure has strong similarities with the usual model selection procedures when two dimensions have to be selected (see e.g. [18, 24]). Here, the specificity is that the penalty criterion does not depend on m_2 (since the variance term only depends on m_1). This makes it possible to consider, in an equivalent way, the criterion (8) we have defined, which focuses on m_1 only.

Oracle-type inequality. Theorem 3 proves that the penalty term introduced above has the good order of magnitude to automatically realize the best bias-variance trade-off. In the statement of the result, and in the sequel, $\|\cdot\|_n$ is the empirical norm defined for all operator T as $\|T\|_n^2 = 1/n \sum_{i=1}^n \|T(X_i)\|^2$ and $\widehat{\Pi}^{op}_{m_1,\infty}$ is the orthogonal projection onto the closure of $V_{m_1,\infty} = \mathrm{Span}\{\widehat{\varphi}_k \otimes \widehat{\varphi}_j,\ 1 \leq j \leq m_1,\ m_2 \geq 1\}$.

Theorem 3. *For all $\zeta > 0$,*

$$\mathbb{E}\|S - \widehat{S}_{\widehat{m}_{1},\infty}\|_n^2 \leq (1+\zeta) \inf_{m_1 \in \mathcal{M}_n} \left\{ \mathbb{E}\|S - \widehat{\Pi}^{op}_{m_1,\infty} S\|_n^2 + c(\zeta)\,\mathrm{pen}(m_1) \right\} + \frac{C'}{n},$$

for a constant $C' > 0$ which does not depend neither on n, nor on m_1 and $c(\zeta) = (2+\zeta)/(1+\zeta)$.

Theorem 3 proves that the selected estimator achieves the best bias-variance compromise, up to a multiplicative constant, and the addition of the term C'/n, which is negligible. Then it achieves the minimax rate and, since the dimension selection procedure does not require the knowledge of the unknown regularity α, it is adaptive. A similar result could be obtained for the risk MSPE, but at the price of additional technicalities. Indeed, to obtain such result it is necessary to prove that, with sufficiently large probability, the quantity $\|S\|_n^2 / \mathrm{MSPE}(S)$ is lower bounded by a constant, for all $S \in V_{m_1,m_2}$ which is a random space (depending on the data $X_1, \ldots, X_n$). We could draw inspiration e.g. from the proof of [6, Lemma 6].

3.5. Application to a real data case

In this section, we present a simulation study on a real data case. Other implementations on simulated and real data cases are available on [8].

The data we study are the electricity consumption of appliances curve of a low energy house located in Stambrudge (Belgium). The dataset is freely available on UCI Machine Learning Repository `https://archive.ics.uci.edu/` and has been studied by [7]. It consists on measurements on 24 variables every 10 minutes from 11th january, 2016, 5pm to 27th may, 2016, 6pm. The variable of interest is the consumption of appliances, which is the main source of energy consumption. The data consists of a d-dimensional times series, with $d = 24$. It is first transformed into a sample of functional data by splitting the data day by day. We can deduce from the variable selection study conducted in [26] that the most important variable to predict appliances electricity consumption of day i is the appliances electricity consumption of day $i-1$, and that a log-transformation of the covariates seems to lead to better results. Then, in our study, the variable to predict Y_i is the log of the appliances energy consumption of day i and X_i is the log of appliances energy consumption of day $i-1$. The data are also recentered. We present in Figure 3 the original and transformed data.

Another difficulty for the estimation procedure is that it requires the knowledge of the trace of the noise operator σ_ε^2, which is unknown in practice. To get around this difficulty, we adapt the method proposed in [6], consisting in replacing the unknown quantity σ_ε^2 in criterion (8) by the contrast $\gamma_n(\widehat{S}_{m_1,\infty})$. In model selection in regression contexts, this method shows strong similarities with the one of [3]. In the context of the functional linear model with scalar output, it has been proven in [6] that the estimator selected by this fully data-driven criterion verifies an oracle-type inequality, achieves the same minimax rates as the estimator selected by the criterion depending on the noise variance and that it does not change significantly the practical performances of the estimator.

As suggested by the simulation study, the value of κ is also fixed to $\kappa = 0.6$. To study the selected dimension, the risk of the estimators and their stability, we perform cross-validation of the sample: for each day i, we calculate the selected dimension $\widehat{m}_1^{(-i)}$ and the L^2-prediction error of the estimator $\widehat{S}^{-i}_{\widehat{m}_1^{(-i)},\infty}$ calculated from the sample $\{(X_j, Y_j), j \neq i\}$. The results are presented in Figure 4. The dimension selection procedure is quite stable, selecting more than 80% of time the dimension $\widehat{m}_1 = 11$ and the L^2-prediction error does not explode for some observations.

We also plot in Figure 5, for three well-chosen days i ($i = 104$ is the day for which the distance $\|Y_i - \widehat{Y}_i^{-i}\|$ is minimal, $i = 4$ corresponds to the median prediction error and $i = 83$ to the maximal prediction error), the true value of Y_i and its prediction $\widehat{Y}_i^{-i} = \widehat{S}^{-i}_{\widehat{m}_1^{(-i)},\infty}(X_i)$.

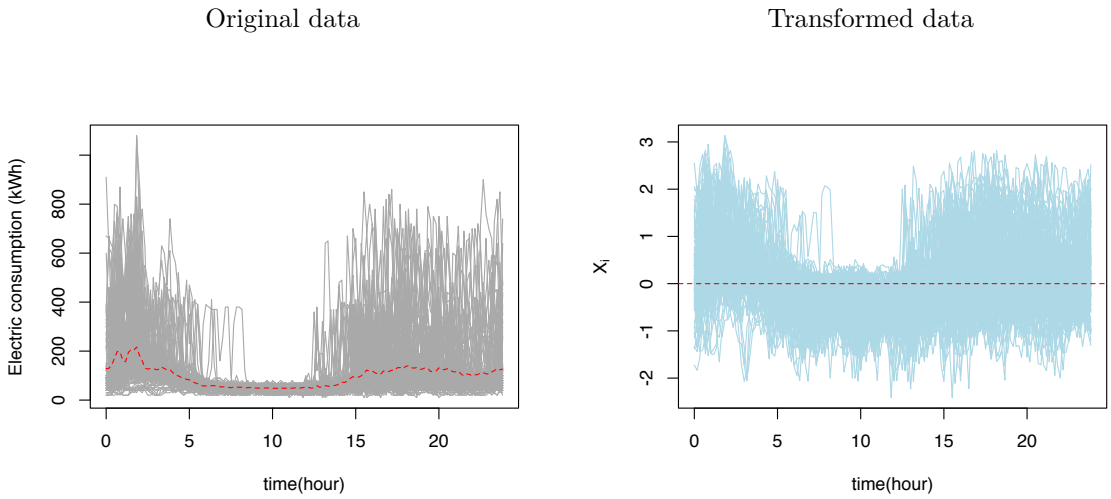

FIGURE 3. Evolution of electric consumption of appliances during $n = 136$ days (original data, thin gray lines) and functions of the sample (transformed data : centered version of the logarithm of the original data, thin blue lines). The dashed red lines are the empirical mean of each sample.

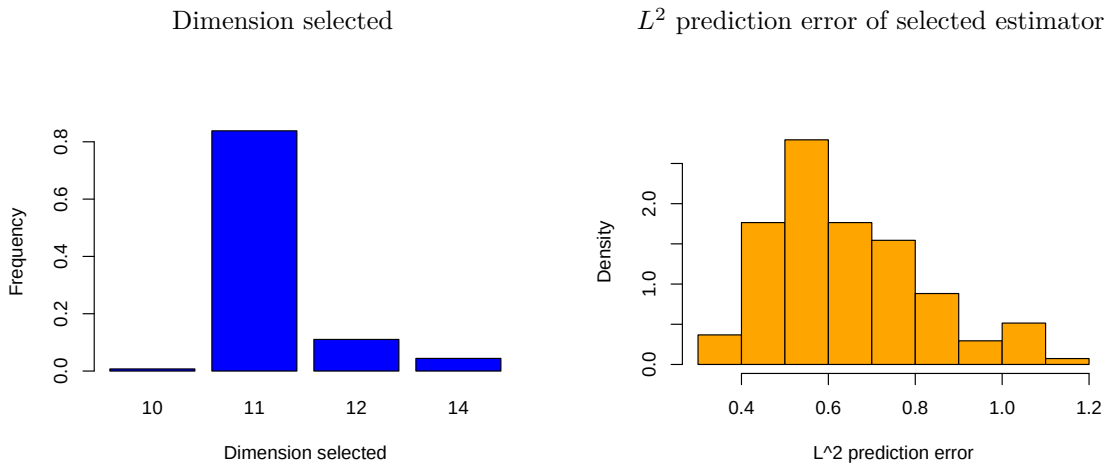

FIGURE 4. Dimension selected and L^2 prediction error $\|Y_i - \widehat{Y}_i^{(-i)}\|$ of the estimator calculated from each cross-validated sample.

Figure 6 represents, for the same days, the prediction of appliances energy consumption (after adding the mean and taking the exponential).

We see in Figure 6 that prediction captures trends well and that the worst prediction seems to be due to a brutal change of behavior of the appliances electricity consumption which is quite hard to predict and may be due to external factors (hence unavoidable with our model).

REFERENCES

[1] Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

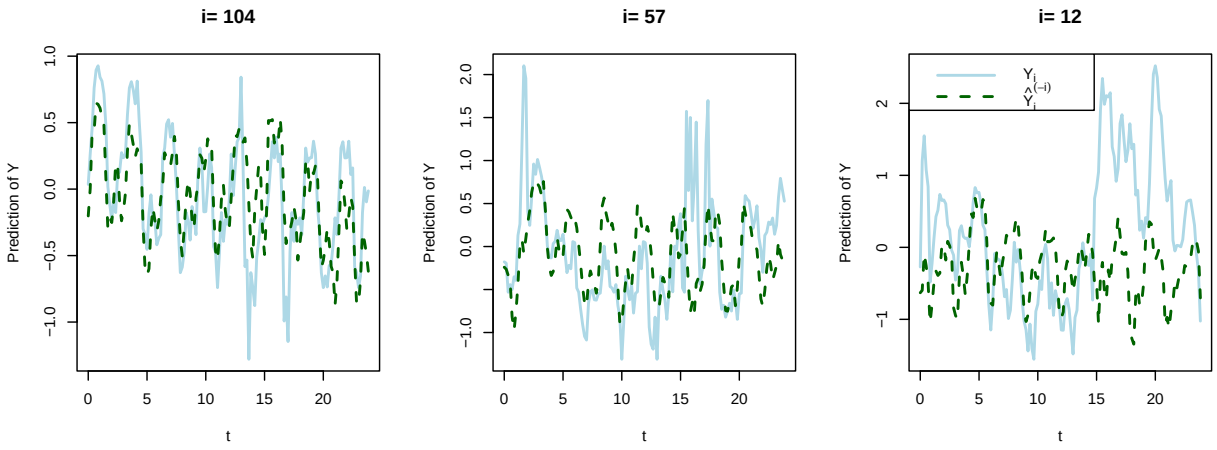

FIGURE 5. Cross-validated prediction $\widehat{Y}_i^{-i}$ made for three days (the days where the prediction is best, median and worst).

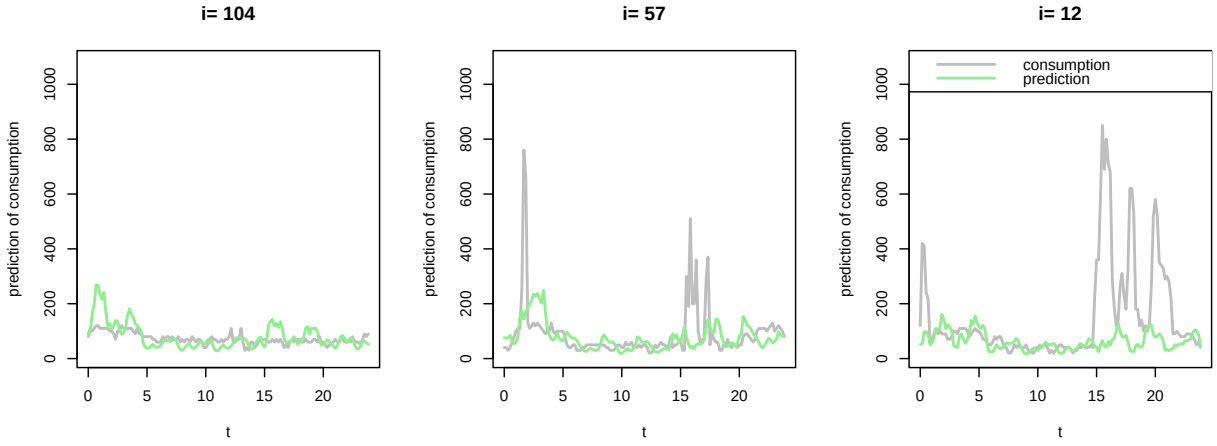

FIGURE 6. Prediction of appliances energy consumption.

[2] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Stat.*, 6:127–146, 2002.

[3] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Estimator selection in the Gaussian setting. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(3):1092 – 1119, 2014.

[4] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields*, 113(3):301–413, 1999.

[5] Lucien Birgé, Pascal Massart, et al. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[6] Élodie Brunel, André Mas, and Angelina Roche. Non-asymptotic adaptive prediction in functional linear models. *J. Multiv. Anal.*, 143:208–232, 2016.

[7] Luis M. Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140, 2017.

[8] Gaëlle Chagny, Anouar Meynaoui, and Angelina Roche. Adaptive nonparametric estimation in the functional linear model with functional output. *arXiv preprint arXiv:2203.00518*, 2022.

[9] Jiawei Chang and Terry Lyons. Insertion algorithm for inverting the signature of a path. *arXiv preprint arXiv:1907.08423*, 2019.
[10] Kuo-Tsai Chen. Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula. *Annals of Mathematics*, pages 163–178, 1957.
[11] Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.
[12] Fabienne Comte and Jan Johannes. Adaptive functional linear regression. *Ann. Statist.*, 40(6):2765–2797, 2012.
[13] Christophe Crambes and André Mas. Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli*, 19(5B):2627–2651, 2013.
[14] Peter K Friz and Nicolas B Victoir. *Multidimensional stochastic processes as rough paths: theory and applications*, volume 120. Cambridge University Press, 2010.
[15] Masaaki Imaizumi and Kengo Kato. PCA-based estimation for functional linear regression with functional responses. *J. Multiv. Anal.*, 163:15–36, 2018.
[16] Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.
[17] Patrick Kidger and Terry Lyons. Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU. *arXiv preprint arXiv:2001.00706*, 2020.
[18] Claire Lacour. Adaptive estimation of the transition density of a Markov chain. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 43(5):571–597, 2007.
[19] Terry J Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998.
[20] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
[21] Xavier Pennec and Vincent Arsigny. Exponential barycenters of the canonical Cartan connection and invariant means on Lie groups. In *Matrix information geometry*, pages 123–166. Springer, 2012.
[22] Imanol Perez Arribas, Guy M Goodwin, John R Geddes, Terry Lyons, and Kate EA Saunders. A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational psychiatry*, 8(1):1–7, 2018.
[23] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, Mar 2011.
[24] Sandra Plancade. Adaptive estimation of the conditional cumulative distribution function from current status data. *J. Statist. Planng Inf.*, 143(9):1466–1485, 2013.
[25] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, 2005.
[26] Angelina Roche. Variable selection and estimation in multivariate functional linear regression via the LASSO. https://hal.archives-ouvertes.fr/hal-01725351, 2021.
[27] Nozomi Sugiura and Shigeki Hosoda. Machine learning technique using the signature method for automated quality control of Argo profiles. *Earth and Space Science*, 7(9):e2019EA001019, 2020.
[28] Weixin Yang, Lianwen Jin, and Manfei Liu. Deepwriterid: An end-to-end online text-independent writer identification system. *IEEE Intelligent Systems*, 31(2):45–53, 2016.