Linear functional regression with truncated signatures

Adeline Fermanian

Version of Record: https://www.sciencedirect.com/science/article/pii/S0047259X22000483
Manuscript_1a8876271e31c792c1a0e9c5dd2ad14f

Functional linear regression with truncated signatures

Adeline Fermanian[a,*]

[a]*Sorbonne Université, CNRS, LPSM, Paris, France*

Abstract

We place ourselves in a functional regression setting and propose a novel methodology for regressing a real output on vector-valued functional covariates. This methodology is based on the notion of signature, which is a representation of a function as an infinite series of its iterated integrals. The signature depends crucially on a truncation parameter for which an estimator is provided, together with theoretical guarantees. An empirical study on both simulated and real-world datasets shows that the resulting methodology is competitive with traditional functional linear models, in particular when the functional covariates take their values in a high dimensional space.

Keywords: Functional data analysis, Linear regression, Signatures.
2020 MSC: Primary 62R10, Secondary 60L10

1. Introduction

In a classical regression setting, a real output Y is described by a finite number of predictors. A typical example would be to model the price of a house as a linear function of several characteristics such as surface area, number of rooms, location, and so on. These predictors are typically encoded as a vector in $\mathbb{R}^p$, $p \in \mathbb{N}^*$. However, some applications do not fall within this setting. For example, in medicine, a classical task consists of predicting the state of a patient (for example, ill or not) from the recording of several physiological variables over some time. The input data is then a function of time and not a vector. Similarly, sound recognition or stock market prediction tasks both consist of learning from time series, possibly multidimensional. Then, a natural idea is to extend the linear model to this more general setting, where one wants to predict from a functional input, of the form $X : [0, 1] \to \mathbb{R}^d$, $d \geq 1$.

This casts our problem into the field of functional data analysis and more specifically within the framework of functional linear regression [31, 39]. This rich domain has undergone considerable developments in recent decades, as illustrated by the monographs of Ramsay and Silverman [40] and Ferraty and Vieu [13], and the review by Morris [36]. One of the core principles of functional data analysis is to represent input functions on a set of basis functions, for example, splines, wavelets, or the Fourier basis. Another approach also consists in extracting relevant handcrafted features, depending on the field of application. For example, [4] and [42] provide overviews of learning methods specific to speech and human action recognition, respectively.

In this article, we build on the work of Levin et al. [25] and explore a novel approach to linear functional regression, called the signature linear model. Its main strength is that it is naturally adapted to vector-valued functions, which is not the case with most of the methods previously mentioned. Its principle is to represent a function by its signature, defined as an infinite series of its iterated integrals. Signatures date back from the 60s when Chen [9] showed that a smooth path can be faithfully represented by its iterated integrals and it has been at the center of rough path theory in the 90s [15, 30]. Rough path theory has seen extraordinary developments in recent times, and, in particular, has gained attention from the machine learning community. Indeed, signatures combined with (deep) learning algorithms have been successfully applied in various fields, such as characters recognition [23, 28, 44, 45], human action recognition [26, 46], speech emotion recognition [43], medicine [2, 32, 34, 35], or finance [3]. We refer the reader to Chevyrev and Kormilitzin [10] for an introduction to signatures in machine learning, and to Fermanian [12] for a more recent overview.

*Corresponding author. Email address: adeline.fermanian@sorbonne-universite.fr

We stress again that the main advantage of the signature approach is that it can handle multidimensional input functions, that is, functions $X : [0,1] \to \mathbb{R}^d$ where $d \geq 2$, whereas traditional methods were designed for real-valued functions. Many modern datasets come in this form with a large dimension d. Moreover, the signature method requires little assumptions on the regularity of X and encodes nonlinear geometric information, that is, gives rise to interpretable regression coefficients. Finally, it is theoretically grounded by good approximation properties: any continuous function can be approximated arbitrarily well by a linear function of the truncated signature [22].

Since any continuous function of X can be approximated by a linear function on its truncated signature, the estimation of a regression function boils down to the estimation of the coefficients in this scalar product. The truncation order of the signature is therefore a crucial parameter as it controls the complexity of the model. Thus, in our quest for a linear model on the signature, one of the main purposes of our article will be to estimate this parameter. With an estimator of the truncation order at hand, the methodology is complete and the signature linear model can be applied to both simulated and real-world data, demonstrating its good performance for practical applications.

To summarize, our document is organized as follows. First, in Section 2, we set the mathematical framework of functional regression and recall the definition of the signature and its main properties. Then, in Section 3, we introduce our model, called 'signature linear model', and define estimators of its parameters. Their rates of convergence are given in Section 4. Finally, Section 5 is devoted to the practical implementation of the signature linear model. We conclude by demonstrating its performance on both simulated and real-world datasets in Section 6.

For the sake of clarity, the proofs of the mathematical results are postponed to Section 8. The code is completely reproducible and available at `https://github.com/afermanian/signature-regression`.

2. Mathematical framework

2.1. Functional linear regression

We place ourselves in a functional linear regression setting with scalar responses: we are given a dataset $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where the pairs (X_i, Y_i) are independent and identically distributed copies of a random couple (X, Y), where X is a (random) function, $X : [0,1] \to \mathbb{R}^d$, $d \geq 1$, and Y a real random variable. Our goal is to approximate the regression function $f(X) = \mathbb{E}[Y|X]$ by a parametrized linear function f_θ and to build an estimator of θ.

In the univariate case, that is when $d = 1$, the classical functional linear model [14, 21] writes

$$Y = \alpha + \int_0^1 X(t)\beta(t)dt + \varepsilon, \tag{1}$$

where $\alpha \in \mathbb{R}, \beta : [0,1] \to \mathbb{R}$ and ε is a random noise. The functional coefficients β and the functional covariates X_i are then expanded on basis functions:

$$\beta(t) = \sum_{k=1}^{K} b_k \phi_k(t), \qquad X_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t), \tag{2}$$

where $\phi_1, \dots, \phi_K$ are a set of real-valued basis functions (for example the monomials $1, t, t^2, \dots, t^K$ or the Fourier basis). Equation (1) can then be rewritten in terms of the c_{ik}s and b_ks, which brings the problem back to the well-known multivariate linear regression setting. Different approaches can then be used in terms of choice of basis functions and regularization [see 40, Chapter 15]. Note that another common approach is functional principal components regression [6, 7]. The idea is to perform a functional principal components analysis (fPCA) on X, which gives a representation of X as a sum of K orthonormal principal components, and to use these as basis functions ϕ_ks.

In both cases, the functional nature of the problem is dealt with by projecting the functions X on a smaller linear space, spanned by basis functions. This basis expansion is not straightforward to extend to the vector-valued case, that is when $d > 1$, the common approach being to expand each coordinate of X independently. This amounts to assuming that there are no interactions between coordinates, which is a strong assumption and not an efficient representation when the coordinates are highly correlated. Moreover, to our knowledge, the only theoretical results in the vector-valued case are found in the domain of longitudinal data analysis [17, 37]. In this case, the different coordinates are assumed to be repeated measurements of a quantity of interest on a patient and each coordinate is given a parametric

model, in the same spirit as ANOVA models. These parametric models do not apply in the general case when the coordinates may correspond to different quantities such as the evolution of different stocks or the x-y-z coordinates of a pen trajectory.

The signature approach removes the need to make such assumptions: the focus moves from finding a functional model for X to finding a basis for functions of X. In other words, instead of using a basis of functions, we use a basis of functions of functions. In a regression setting, this shift of perspective is particularly adequate since the object of interest is the regression function $f(X)$ and not X itself. The whole approach is based on the signature transformation, which takes as input a function X and outputs an infinite vector of coefficients known to characterize X under some smoothness assumptions. In particular, there are no assumptions on the structure of dependence in the different coordinates of X. In other words, the signature is naturally adapted to the vector-valued case.

Before we delve into the signature linear model, we gently introduce the notion of signature and review some of its important properties.

2.2. The signature of a path

We give here a brief presentation of signatures but the reader is referred to Lyons et al. [30] or Friz and Victoir [15] for a more involved mathematical treatment with proofs. To follow the vocabulary from rough path theory, we will often call the functional covariate $X : [0, 1] \to \mathbb{R}^d$ a path. Our basic assumption is that X is of bounded variation, i.e., it has finite length.

Definition 1. Let $X : [0, 1] \to \mathbb{R}^d$, $t \mapsto (X_t^1, \dots, X_t^d)^\top$. The total variation of X is defined by

$$\|X\|_{TV} = \sup_{\mathcal{I}} \sum_{(t_0,\dots,t_k)\in\mathcal{I}} \|X_{t_i} - X_{t_{i-1}}\|,$$

where the supremum is taken over all finite subdivisions of $[0, 1]$, and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. The set of paths of bounded variation is then defined by

$$BV(\mathbb{R}^d) = \{X : [0, 1] \to \mathbb{R}^d \,|\, \|X\|_{TV} < \infty\}.$$

We recall that $BV(\mathbb{R}^d)$ endowed with the norm $\|X\|_{BV(\mathbb{R}^d)} = \|X\|_{TV} + \sup_{t\in[0,1]} \|X_t\|$ is a Banach space. We stress that the basis functions traditionnaly used in functional data analysis are of bounded variation. The assumption that $X \in BV(\mathbb{R}^d)$ is therefore much less restrictive than assuming an expansion such as (2). This assumption allows to define Riemann-Stieljes integrals along paths, which puts us in a position to define the signature.

Definition 2. Let $X \in BV(\mathbb{R}^d)$ and $I = (i_1, \dots, i_k) \subset \{1, \dots d\}^k$, $k \geq 1$, be a multi-index of length k. The signature coefficient of X along the index I on $[0, 1]$ is defined by

$$S^I(X) = \int_{0\leq u_1<\cdots<u_k\leq 1} \cdots \int dX_{u_1}^{i_1} \dots dX_{u_k}^{i_k}. \tag{3}$$

$S^I(X)$ is then said to be a signature coefficient of order k.

The signature of X is the sequence containing all signature coefficients, i.e.,

$$S(X) = (1, S^{(1)}(X), \dots, S^{(d)}(X), S^{(1,1)}(X), S^{(1,2)}(X), \dots, S^{(i_1,\dots,i_k)}(X), \dots).$$

The signature of X truncated at order m, denoted by $S^m(X)$, is the sequence containing all signature coefficients of order lower than or equal to m, that is

$$S^m(X) = (1, S^{(1)}(X), S^{(2)}(X), \dots, S^{\overbrace{(d,\dots,d)}^{\text{length } m}}(X)).$$

Note that the assumption that $X \in BV(\mathbb{R}^d)$ may be relaxed: the signature may still be defined when the Riemann-Stieljes integrals are not well-defined. For example, the signature of the Brownian motion may be defined via

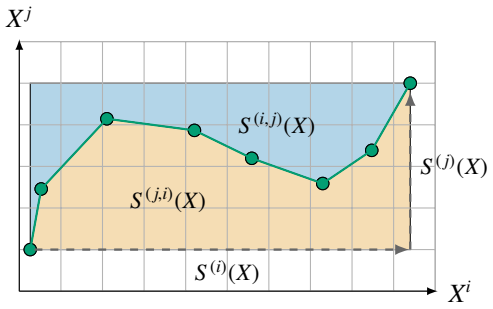


Fig. 1: Geometric interpretation of the signature coefficients. The terms $S^{(i)}(X)$ and $S^{(j)}(X)$ are the increments of the coordinates i and j respectively. The terms $S^{(i,j)}$ and $S^{(j,i)}$ correspond to the areas of the blue and orange regions respectively.

Stratonovitch integrals [24]. Integrating paths that are not of bounded variation is actually one of the motivations behind the definition of the signature in rough path theory.

A crucial feature of the signature is that it encodes the geometric properties of the path, as shown in Fig. 1. Indeed, coefficients of order 1 correspond to the increments of the path in each coordinate and the coefficients of order 2 correspond to areas outlined by the path. For higher orders of truncation, the signature contains information about the joint evolution of tuples of coordinates. Moreover, it is clear from its definition as an integral that the signature is independent of the time parametrization [15, Proposition 7.10] and that it is invariant by translation. Therefore, the signature looks at functions as purely geometric objects, without any information about sampling frequency, speed, or travel time, hence the terminology of 'paths'.

Note that the definition can be extended to paths defined on any interval $[s,t] \subset \mathbb{R}$ by changing the integration bounds in (3). Moreover, it is clear that there are d^k signature coefficients of order k. The signature truncated at order m is therefore a vector of dimension $s_d(m)$, where

$$s_d(m) = \sum_{k=0}^{m} d^k = \frac{d^{m+1}-1}{d-1} \quad \text{if} \quad d \geq 2, \qquad s_d(m) = m+1 \quad \text{if} \quad d = 1.$$

Thus, provided $d \geq 2$, the size of $S^m(X)$ increases exponentially with m and polynomially with d—some typical values are presented in Table 1.

Table 1: Typical values of $s_d(m)$, the size of the signature of a path $X \in BV(\mathbb{R}^d)$ truncated at order m.

	$d=2$	$d=3$	$d=6$
$m=1$	2	3	6
$m=2$	6	12	42
$m=5$	62	363	9330
$m=7$	254	3279	335922

The set of coefficients of order k can be seen as an element of the kth tensor product of $\mathbb{R}^d$ with itself, denoted by $(\mathbb{R}^d)^{\otimes k}$. For example, the d coefficients of order 1 can be written as a vector, and the d^2 coefficients of order 2 as a matrix, i.e.,

$$\begin{pmatrix} S^{(1)}(X) \\ \vdots \\ S^{(d)}(X) \end{pmatrix} \in \mathbb{R}^d, \qquad \begin{pmatrix} S^{(1,1)}(X) & \dots & S^{(1,d)}(X) \\ \vdots & \ddots & \vdots \\ S^{(d,1)}(X) & \dots & S^{(d,d)}(X) \end{pmatrix} \in \mathbb{R}^{d\times d} \approx (\mathbb{R}^d)^{\otimes 2}.$$

Similarly, coefficients of order 3 can be written as a tensor of order 3, and so on. Then, $S(X)$ can be seen as an element of the tensor algebra

$$\mathbb{R} \oplus \mathbb{R}^d \oplus (\mathbb{R}^d)^{\otimes 2} \oplus \cdots \oplus (\mathbb{R}^d)^{\otimes k} \oplus \cdots .$$

Although not fundamental in the present paper, this structure of tensor algebra is the right space to understand properties of the signature [15, 30].

Let us give two examples of paths and their signatures.

Example 1. Let X be a parametrized curve: for any $t \in [0,1]$, $X_t = (t, f(t))$, where $f : \mathbb{R} \to \mathbb{R}$ is a smooth function. Then,

$$S^{(1)}(X) = \int_0^1 dX_t^1 = \int_0^1 dt = 1, \quad S^{(2)}(X) = \int_0^1 dX_t^2 = \int_0^1 f'(t)dt = f(1) - f(0),$$

where f' denotes the derivative of f. Similarly, the signature coefficient along $(1,2)$ is

$$S^{(1,2)}(X) = \int_0^1 \int_0^t dX_u^1 dX_t^2 = \int_0^1 \Big(\int_0^t du \Big) f'(t)dt = \int_0^1 t f'(t)dt = f(1) - \int_0^1 f(t)dt.$$

Example 2. Let X be a d-dimensional linear path:

$$X_t = \begin{pmatrix} X_t^1 \\ \vdots \\ X_t^d \end{pmatrix} = \begin{pmatrix} a_1 + b_1 t \\ \vdots \\ a_d + b_d t \end{pmatrix}.$$

Then, for any index $I = (i_1, \dots, i_k) \subset \{1, \dots, d\}^k$, the signature coefficient along I is

$$S^{(i_1,\dots,i_k)}(X) = \int_{0 \le u_1 < \cdots < u_k \le 1} \cdots \int dX_{u_1}^{i_1} \cdots dX_{u_k}^{i_k} = \int_{0 \le u_1 < \cdots < u_k \le 1} \cdots \int b_{i_1} du_1 \cdots b_{i_k} du_k = \frac{b_{i_1} \dots b_{i_k}}{k!}. \tag{4}$$

It is clear here that the signature is invariant by translation: $S(X)$ depends only on the slope of X and not on the initial position $(a_1, \dots, a_d)^\top \in \mathbb{R}^d$.

We now recall a series of properties of the signature that motivate the definition of the signature linear model. The first important property provides a criterion for the uniqueness of signatures.

Proposition 1. *Assume that $X \in BV(\mathbb{R}^d)$ contains at least one monotone coordinate, then $S(X)$ characterizes X up to translations and reparametrizations.*

This is a sufficient condition, a necessary one has been derived by Hambly and Lyons [19] and is based on the construction of an equivalence relation between paths, called tree-like equivalence. For any path $X \in BV(\mathbb{R}^d)$, the time-augmented path $\tilde{X}_t = (X_t, t)^\top \in BV(\mathbb{R}^{d+1})$ satisfies the assumption of Proposition 1, which ensures signature uniqueness. Enriching the path with new dimensions is actually a classic part of the learning process when signatures are used, and is discussed by Fermanian [12] and Morrill et al. [33]. We will always use this time-augmentation transformation before computing signatures.

The next proposition states that the signature linearizes functions of X and is the core motivation of the signature linear model. We refer the reader to Levin et al. [25], Theorem 3.1, for a proof in a similar setting.

Proposition 2. *Let $D \subset BV(\mathbb{R}^d)$ be a compact set of paths that such that, for any $X \in D$, $X_0 = 0$, and denote by $\tilde{X} = (X_t, t)_{t \in [0,1]}^\top$ the associated time-augmented path. Let $f : D \to \mathbb{R}$ be a continuous function. Then, for every $\varepsilon > 0$, there exist $m^* \in \mathbb{N}$, $\beta^* \in \mathbb{R}^{s_d(m^*)}$, such that, for any $X \in D$,*

$$\left| f(X) - \langle \beta^*, S^{m^*}(\tilde{X}) \rangle \right| \le \varepsilon,$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product on $\mathbb{R}^{s_d(m^*)}$.

This proposition is a consequence of the Stone-Weierstrass theorem. The classical Weierstrass approximation theorem states that every real-valued continuous function on a closed interval can be uniformly approximated by a polynomial function. Linear forms on the signature can, therefore, be thought of as the equivalent of polynomial functions for paths. The assumption that $X_0 = 0$ is due to the fact that signatures are invariant by translation: no information about the initial position of the path is contained in signatures.

Finally, the following bound on the norm of the truncated signature allows to control the rate of decay of signature coefficients of high order—see Lyons [29, Lemma 5.1] for a proof.

Proposition 3. *Let $X : [0,1] \to \mathbb{R}^d$ be a path in $BV(\mathbb{R}^d)$. Then, for any $m \geq 0$,*

$$\|S^m(X)\| \leq \sum_{k=0}^{m} \frac{\|X\|_{TV}^k}{k!} \leq e^{\|X\|_{TV}}.$$

3. The signature linear model

3.1. Presentation of the model

We are now in a position to present the signature linear model. Recall that our goal is to model the relationship between a real random variable $Y \in \mathbb{R}$ and a random input path $X \in BV(\mathbb{R}^d)$. Without loss of generality, we now assume that $d \geq 2$ and that X has been augmented with time—in other words, one coordinate of X is $t \mapsto t$. Proposition 2 then motivates the following model which was first introduced in a slightly different form by [25]: we assume that there exists $m \in \mathbb{N}$, $\beta_m^* \in \mathbb{R}^{s_d(m)}$, such that

$$\mathbb{E}[Y|X] = \langle \beta_m^*, S^m(X) \rangle, \quad \mathrm{Var}(Y|X) \leq \sigma^2 < \infty. \tag{5}$$

We consider throughout the article the smallest $m^* \in \mathbb{N}$ such that there exists $\beta_{m^*}^* \in \mathbb{R}^{s_d(m^*)}$ satisfying

$$\mathbb{E}[Y|X] = \langle \beta_{m^*}^*, S^{m^*}(X) \rangle.$$

In other words, we assume a regression model, where the regression function is a linear form on the signature. A few comments are in order.

From an approximation point of view, this model is very general. Indeed, by Proposition 2, the only requirements for model (5) to be valid are the continuity of the regression function $f(X) = \mathbb{E}[Y|X]$ and the fact that $S(X)$ must characterize the random path X. The latter is ensured by using a time augmentation, that is, considering $\tilde{X}_t = (X_t, t)$, and by fixing the initial value, for example $X_0 = 0$. Then, under the assumption that the data is in a compact set—which will be guaranteed later on by assumption (H_K)—, for any threshold $\varepsilon > 0$, there exist $m^* \in \mathbb{N}$ and $\beta_{m^*}^* \in \mathbb{R}^{s_d(m^*)}$ such that

$$\left| \mathbb{E}[Y|X] - \langle \beta_{m^*}^*, S^{m^*}(X) \rangle \right| \leq \varepsilon.$$

In other words, we know that (the first part of) model (5) is true up to an error of ε. A striking fact is that no assumption that $\mathbb{E}[Y|X]$ is linear in X is needed, contrary to functional models of the form (1).

It is instructive to further compare this model to the functional model (1). Much fewer assumptions on X are needed: it is only assumed to be of finite variation, whereas in (1) it has to have a finite basis expansion. Moreover, our model is directly adapted to the vector-valued case. Finally, it depends directly on a finite vector $\beta_{m^*}^*$, whereas (1) is written in terms of a function β, which must itself be written on basis functions. Note that the choice of basis needs to be adapted to each particular application, whereas the signature linear model only depends on two parameters. In a nutshell, it is a more general model with fewer hyperparameters.

It can be noticed that, since the first term of signatures is always equal to 1, this regression model contains an intercept: when $m^* = 0$, (5) is a constant model. There are two unknown quantities in model (5): m^* and $\beta_{m^*}^*$. The parameter m^* is the truncation order of the signature of X and controls the model size, whereas $\beta_{m^*}^*$ is the vector of regression coefficients, whose size $s_d(m^*)$ depends on m^*.

The signature truncation order m^* is a key quantity in this model and influences the rest of the study. Indeed, it controls the number of coefficients and therefore the computational feasibility of the whole method. However, it is in general little discussed in the literature and small values are picked arbitrarily, regardless of the model used on top of signatures. For example, [28] consider values of m up to 2, [44] up to 3, Arribas et al. [2] and Lai et al. [23] up to 4, [45] up to 5 , and [46] up to 8. Thus, one of our main objectives is to establish a rigorous procedure to estimate m^*, and, to this end, we define a consistent estimator of m^*. As we will see later, a simple estimator of $\beta_{m^*}^*$, and therefore of the regression function, is then also obtained.

3.2. Estimating the truncation order

Let $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be a set of i.i.d. observations drawn according to the law of (X, Y). We use the approach of penalized empirical risk minimization. For the moment, let us fix a certain truncation order $m \in \mathbb{N}$, and let $\alpha > 0$ denote a fixed positive number. Then, the ball in $\mathbb{R}^{s_d(m)}$ of radius α centered at 0 is denoted by

$$B_{m,\alpha} = \{\beta \in \mathbb{R}^{s_d(m)} \mid \|\beta\| \leq \alpha\},$$

where $\|\cdot\|$ stands for the Euclidean norm, whatever the dimension. By a slight abuse of notation, the sequence $(B_{m,\alpha})_{m\in\mathbb{N}}$ can be seen as a nested sequence of balls, i.e., $B_{0,\alpha} \subset B_{1,\alpha} \subset \cdots \subset B_{m,\alpha} \subset B_{m+1,\alpha} \subset \cdots$. From now on, we will only consider coefficients within these balls. Therefore, we assume that the true coefficient $\beta^*_{m^*}$ lies within such a ball, i.e., we make the assumption

(H_α) There exists $\alpha > 0$ such that $\beta^*_{m^*} \in B_{m^*,\alpha}$.

On the one hand, for a fixed truncation order m, the theoretical risk is defined by $\mathcal{R}_m(\beta) = \mathbb{E}(Y - \langle \beta, S^m(X)\rangle)^2$. Then, the minimal theoretical risk for a certain truncation order m, is defined by

$$L(m) = \inf_{\beta\in B_{m,\alpha}} \mathcal{R}_m(\beta) = \mathcal{R}_m(\beta^*_m),$$

where $\beta^*_m \in \operatorname{argmin}_{\beta\in B_{m,\alpha}} \mathcal{R}_m(\beta)$ (note that the existence of β^*_m is ensured by convexity of the problem). Since the sets $(B_{m,\alpha})_{m\in\mathbb{N}}$ are nested, L is a decreasing function of m. Its minimum is attained at $m = m^*$, and, provided $m \geq m^*$, $L(m)$ is then constant and equal to

$$\mathcal{R}(\beta^*_{m^*}) = \mathbb{E}(Y - \langle \beta^*_{m^*}, S^{m^*}(X)\rangle)^2 = \mathbb{E}(\mathrm{Var}(Y|X)) \leq \sigma^2.$$

On the other hand, the empirical risk with signature truncated at order m is defined by $\widehat{\mathcal{R}}_{m,n}(\beta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \beta, S^m(X_i)\rangle)^2$, where $\beta \in B_{m,\alpha}$. The minimum of $\widehat{\mathcal{R}}_{m,n}$ over $B_{m,\alpha}$ is denoted by $\widehat{L}_n(m)$ and defined as

$$\widehat{L}_n(m) = \min_{\beta\in B_{m,\alpha}} \widehat{\mathcal{R}}_{m,n}(\beta) = \widehat{\mathcal{R}}_{m,n}(\widehat{\beta}_m),$$

where $\widehat{\beta}_m$ denotes a point in $B_{m,\alpha}$ where the minimum is attained. Note that $\beta \mapsto \widehat{\mathcal{R}}_{m,n}(\beta)$ is a convex function so $\widehat{\beta}_m$ exists. We point out that minimizing $\widehat{\mathcal{R}}_{m,n}$ over $B_{m,\alpha}$ is equivalent to performing a Ridge regression with a certain regularization parameter which depends on α.

To summarize, for a fixed truncation order m, a Ridge regression gives the best parameter $\widehat{\beta}_m$ to model Y as a linear form on the signature of X truncated at order m. Recall that our goal is to find a truncation order $\widehat{m}$ close to the true one m^*. Since the $(B_{m,\alpha})_{m\in\mathbb{N}}$ are nested, the sequence $(\widehat{L}_n(m))_{m\in\mathbb{N}}$ decreases with m. Indeed, increasing m makes the set of parameters larger and therefore decreases the empirical risk. An estimator of m^* can then be defined by a trade-off between this decreasing empirical risk and an increasing function that penalizes the number of coefficients:

$$\widehat{m} = \min\Big(\operatorname*{argmin}_{m\in\mathbb{N}}(\widehat{L}_n(m) + \mathrm{pen}_n(m))\Big),$$

where $m \mapsto \mathrm{pen}_n(m)$ is an increasing function of m that will be defined in Theorem 1. If the minimum is reached by several values, we set $\widehat{m}$ to the smallest one. The procedure is illustrated in Fig. 2 for a toy dataset which will be described in Section 5.2.

Now that we have an estimate of m^*, which is a key ingredient in establishing the whole process of the expected signature method, and before presenting the whole procedure, we justify the estimator by some theoretical results.

4. Performance bounds

In this section, we show that it is possible to calibrate a penalization that ensures exponential convergence of $\widehat{m}$ to m^*. In addition to (H_α), we need the following assumption:

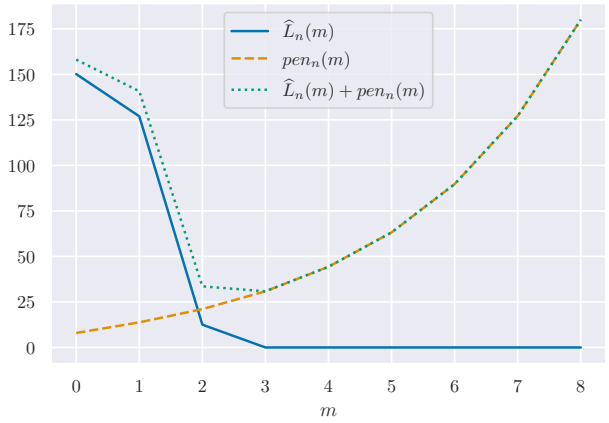


Fig. 2: The functions $m \mapsto \widehat{L}_n(m)$ (blue solid curve), $m \mapsto \text{pen}_n(m)$ (orange dashed curve) and $m \mapsto \widehat{L}_n(m) + \text{pen}_n(m)$ (green dotted curve) for a toy dataset. In this case, the value of $\widehat{m}$ is $\widehat{m} = 3$.

(H_K) there exists $K_Y > 0$ and $K_X > 0$ such that almost surely $|Y| \leq K_Y$ and $\|X\|_{TV} \leq K_X$.

The assumption (H_K) says that the trajectories have a length uniformly bounded by K_X and that the responses Y live in a compact set. These assumptions are quite different from the ones in functional linear models of the form (1). Indeed, concerning the regularity of X, they typically assume that X is in L^2 and that its coefficients c_{ik} in the basis expansion (2) decrease sufficiently fast. We therefore trade an assumption that the functions have a nice basis decomposition for a compactness property, which seems a reasonable choice for practical applications. For example, any discrete-time time-series model observed over a finite horizon, such as ARIMA, satisfies (H_K). Any continuously differentiable function with bounded derivative also satisfies (H_K). Note also that (H_K) does not depend strongly on the dimension d, whereas the assumptions of functional linear models become very stringent in this case; they typically assume an additive relationship between Y and the different coordinates of X. We shall also use the constant K, defined by

$$K = 2(K_Y + \alpha e^{K_X})e^{K_X}. \tag{6}$$

The main result of the section is the following.

Theorem 1. *Let* $K_{\text{pen}} > 0$, $0 < \rho < \frac{1}{2}$, *and*

$$\text{pen}_n(m) = K_{\text{pen}} n^{-\rho} \sqrt{s_d(m)}. \tag{7}$$

Let n_0 *be the smallest integer satisfying*

$$(n_0)^{\tilde{\rho}} \geq (432K\alpha\sqrt{\pi} + K_{\text{pen}})\Big(\frac{2\sqrt{s_d(m^*+1)}}{L(m^*-1)-\sigma^2} + \frac{\sqrt{2s_d(m^*+1)}}{K_{\text{pen}}\sqrt{d^{m^*+1}}}\Big), \tag{8}$$

where $\tilde{\rho} = \min(\rho, \frac{1}{2} - \rho)$. *Then, under the assumptions* (H_α) *and* (H_K), *for any* $n \geq n_0$,

$$\mathbb{P}(\widehat{m} \neq m^*) \leq C_1 \exp\left(-C_2 n^{1-2\rho}\right),$$

where the constants C_1 *and* C_2 *are defined by*

$$C_1 = 74 \sum_{m>0} e^{-C_3 s_d(m)} + 148m^*, \quad C_3 = \frac{K_{\text{pen}}^2 d^{m^*+1}}{128 s_d(m^*+1)(72K^2\alpha^2 + K_Y^2)}, \tag{9}$$

and

$$C_2 = \frac{1}{16(1152K^2\alpha^2 + K_Y^2)} \min\Big(\frac{K_{\text{pen}}^2 d^{m^*+1}}{8 s_d(m^*+1)}, L(m^*-1) - \sigma^2\Big). \tag{10}$$

This theorem provides a non-asymptotic bound on the convergence of $\widehat{m}$. It implies the almost sure convergence of $\widehat{m}$ to m^*. We can note that the penalty decreases slowly with n (more slowly than a square-root) and increases with m exponentially, i.e., as $d^{m/2}$. The penalty includes an arbitrary constant K_{pen}. Its value that minimizes n_0 is

$$K^*_{\text{pen}} = \sqrt{\frac{(L(m^*-1)-\sigma^2)432\sqrt{\pi}\alpha K}{d^{m^*+1}}},$$

and, in practice, it is calibrated with the slope heuristics method of [5], described in Section 5. The proof of Theorem 1 is based on chaining tail inequalities that bound uniformly the tails of the risk. We refer the reader to Section 8 for a detailed proof.

To give some insights into this estimator it is interesting to look at the behavior of the constants when different quantities vary.

- If the dimension of the path d gets large, then $d^{m^*+1} \sim s_d(m^*+1)$ and the constants C_1 and C_2 stay of the same order (provided that the risk $L(m^*-1)$ stays constant). Therefore, the quality of the bound does not change in high dimensions. However, the constant n_0 increases at the rate of $O(d^{m^*/2\tilde{\rho}})$: we neeed exponentially more data when d grows.

- If the true truncation parameter m^* is large, that is, the regression function $\mathbb{E}[Y|X]$ depends on higher-order terms of the signature, the same phenomenon is observed except that C_1 increases linearly: C_2 and C_3 stay of the same order, $C_1 \sim 148m^*$, and n_0 increases at the rate of $O(d^{m^*/2\tilde{\rho}})$. It is not surprising: when m^* increases, the size of the coefficient $\beta^*_{m^*}$ increases and therefore more data are needed to estimate it.

- If α increases, n_0 and C_1 increase while C_2 decreases. In other words, more data is needed and the quality of the estimator deteriorates. Indeed, when α gets larger, the parameter spaces $B_{m,\alpha}$ gets larger for any m so estimation is harder.

- The last quantity of interest is $L(m^*-1)-\sigma^2 \leq L(m^*-1)-L(m^*)$, which measures the difference of risk between a smaller model and the model truncated at m^*. By definition, it is a strictly positive quantity. When it gets close to zero, it means that a model truncated at m^*-1 is almost as good as a model truncated at m^*. We can see that when this difference decreases, n_0 increases and C_2 decreases: it is harder to find that a truncation order of m^* is better than m^*-1, therefore the estimator $\widehat{m}$ deteriorates.

With an estimator of $\widehat{m}$ at hand, one can simply choose to estimate $\beta^*_{m^*}$ by $\widehat{\beta}_{\widehat{m}}$, which gives an estimator of the regression function in model (5). As a by-product of Theorem 1, we then get the following bound.

Corollary 1. *Under the assumptions (H_α) and (H_K), for any $n \geq n_0$,*

$$\mathbb{E}\Big(\langle\hat{\beta}_{\widehat{m}}, S^{\widehat{m}}(X)\rangle - \langle\beta^*_{m^*}, S^{m^*}(X)\rangle\Big)^2 \leq \frac{C_5}{\sqrt{n}} + C_6 e^{-C_2 n^{1-2\rho}},$$

where the constants C_5 and C_6 are defined by

$$C_5 = 36K\alpha\sqrt{\pi}(m^*+1)\sqrt{s_d(m^*)}, \quad C_6 = 2664K\alpha\sqrt{\pi}\sum_{m>m^*}\sqrt{s_d(m)}e^{-C_3 s_d(m)} + 2\alpha^2 e^{K_X}C_1.$$

The proof is given in Section 8. This rate of convergence in $O(n^{-1/2})$ is similar to the ones usually obtained for functional linear models when $d=1$, except that much less assumptions are needed on the path X. Indeed, the rates obtained on the regression function usually depend on regularity assumptions on X and β in (1). For example, it can depend on the Fourier coefficients of X [18], on the number of Lipschitz-continuous derivatives of β [8], or on the periodicity of X [27]. We can note that when the true coefficient m^* gets larger, prediction is more difficult and the bound increases. This is also the case when K increases, which amounts to allowing larger values for Y and X.

We stress that in both Theorem 1 and Corollary 1, the constant α is assumed to be fixed. In practice, it is unknown and is typically selected via cross-validation. Taking this into account in the theoretical analysis would be an interesting extension for future work. We have now all the ingredients necessary to implement this signature linear model. Before looking at its performance on real-world datasets, we present in the next section the complete methodology from a computational point of view.

5. Computational aspects

5.1. The signature linear model algorithm

The first step towards practical application is to be able to compute signatures efficiently. Typically, the input data consists of arrays of sampled values of X. We choose to interpolate the sampled points linearly, and therefore our problem reduces to computing signatures of piecewise linear paths. To this end, equation (4) gives the signature of a linear path and Chen's theorem [9], stated below, provides a formula to compute recursively the signature of a concatenation of paths.

Let $X : [s,t] \to \mathbb{R}^d$ and $Y : [t,u] \to \mathbb{R}^d$ be two paths, $0 \leq s < t < u \leq 1$. The concatenation of X and Y, denoted by $X * Y$, is defined as the path from $[s,u]$ to $\mathbb{R}^d$ such that, for any $v \in [s,u]$,

$$(X * Y)_v = \begin{cases} X_v, & \text{if } v \in [s,t], \\ X_t + Y_v - Y_t, & \text{if } v \in [t,u]. \end{cases}$$

Proposition 4 (Chen). *Let $X : [s,t] \to \mathbb{R}^d$ and $Y : [t,u] \to \mathbb{R}^d$ be two paths with bounded variation. Then, for any multi-index $(i_1, \dots, i_k) \subset \{1, \dots, d\}^k$,*

$$S^{(i_1,\dots,i_k)}(X * Y) = \sum_{\ell=0}^{k} S^{(i_1,\dots,i_\ell)}(X) \cdot S^{(i_{\ell+1},\dots,i_k)}(Y). \tag{11}$$

This proposition is an immediate consequence of the linearity property of integrals [30, Theorem 2.9]. Therefore, to compute the signature of a piecewise linear path, it is sufficient to iterate the following two steps:

1. Compute with (4) the signature of a linear section of the path;
2. Concatenate it to the other pieces with Chen's formula (11).

This procedure is implemented in the Python library `iisignature` [41]. Thus, for a sample consisting of p points in $\mathbb{R}^d$, if we consider the path formed by their linear interpolation, the computation of the path signature truncated at level m takes $O(pd^m)$ operations. The complexity is therefore linear in the number of sampled points but exponential in the truncation order m, which emphasizes once more the importance of the choice of $\widehat{m}$.

Algorithm 1: Pseudo-code for the signature linear model.

Data: $\{(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)\}$
Result: Estimators $\hat{m}$ and $\hat{\beta}_{\hat{m}}$

Interpolate linearly the columns of $\mathbf{x}_i$ so as to have a set of continuous piecewise linear paths $X_i : [0,1] \to \mathbb{R}^d$, $1 \leq i \leq n$. Add a time dimension, i.e., consider the path $\widetilde{X}_i : [0,1] \to \mathbb{R}^{d+1}$, where $\widetilde{X}_i^j = X_i^j$ for $1 \leq j \leq d$, and $X_{i,t}^{d+1} = t$, $t \in [0,1]$.

Select the Ridge regularization parameter λ by cross validation on the regression model with $\{S^1(\widetilde{X}_1), \dots, S^1(\widetilde{X}_n)\}$ as predictors.

for $m = 1, \dots, M$ **do**

  Compute signatures truncated at level m: $\{S^m(\widetilde{X}_1), \dots, S^m(\widetilde{X}_n)\}$.

  Fit a Ridge regression on the pairs $\{(S^m(\widetilde{X}_1), Y_1), \dots, (S^m(\widetilde{X}_n, Y_n)\}$. Compute its squared loss $\widehat{L}_n(m)$.

  Compute the penalization $\mathrm{pen}_n(m) = K_{\mathrm{pen}} \frac{\sqrt{s_d(m)}}{n^\rho}$.

Choose $\widehat{m} = \underset{0 \leq m \leq M}{\mathrm{argmin}} \left(\widehat{L}_n(m) + \mathrm{pen}_n(m)\right)$.

Compute $\widehat{\beta}_{\widehat{m}}$ by fitting a Ridge regression on $\{(S^{\widehat{m}}(\widetilde{X}_1), Y_1), \dots, (S^{\widehat{m}}(\widetilde{X}_n, Y_n)\}$: $\widehat{\beta}_{\widehat{m}} = (\mathbf{S}^\top \mathbf{S} + \lambda \mathbf{I})^{-1} \mathbf{S}^\top \mathbf{Y}$, where $\mathbf{S} \in \mathbb{R}^{n \times s_d(\widehat{m})}$ is the matrix which rows are the signatures of the inputs $S^{\widehat{m}}(\widetilde{X}_i)^\top$, $\mathbf{I} \in \mathbb{R}^{s_d(\widehat{m}) \times s_d(\widehat{m})}$ is the identity matrix, and $\mathbf{Y} = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^d$ is the vector of responses.

In practice, we are given a dataset $\{(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)\}$, where, for any $1 \leq i \leq n$, $Y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^{d \times p_i}$. The columns of the matrix $\mathbf{x}_i$ correspond to values of a process X_i in $\mathbb{R}^d$ sampled at p_i different times. We fix $M \in \mathbb{N}$ such

that, for any $m \geq M$, the function $m \mapsto \widehat{L}_n(m) + \text{pen}_n(m)$ is strictly increasing and apply the procedure described in Algorithm 1.

Note that in the first step of Algorithm 1 there exist other choices for the embedding of the matrix $\mathbf{x_i}$ into a continuous path $\widetilde{X}_i$ [12]. The parameter ρ is set to 0.4. The constant K_{pen} is calibrated with the so-called slope heuristics method, first proposed by [5].

5.2. *A toy example*

This section is devoted to illustrating the different steps of Algorithm 1 and the convergence of the estimator $\widehat{m}$ with simulated data. We first simulate a dataset $\{(\mathbf{x_1}, Y_1), \ldots, (\mathbf{x_n}, Y_n)\}$ following the signature model (5).

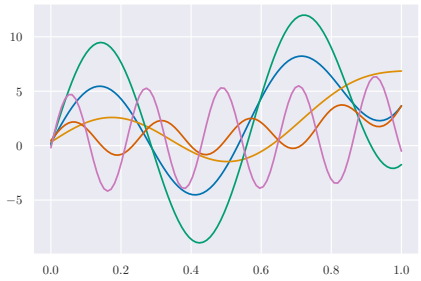


Fig. 3: One sample X_i from model (12) with $d = 5$.

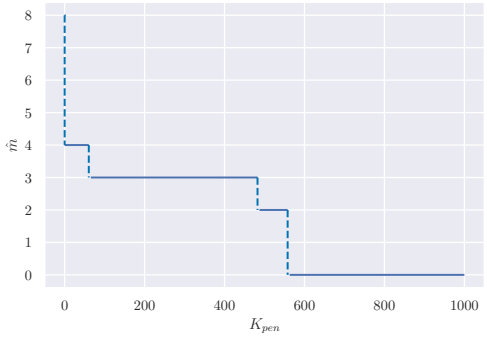


Fig. 4: Selection of K_{pen} with the slope heuristics method.

For any $1 \leq i \leq n$, let $X_i : [0,1] \to \mathbb{R}^d$, $X_{i,t} = (X^1_{i,t}, \ldots, X^d_{i,t})$ be defined by

$$X^k_{i,t} = \alpha^k_{i,1} + 10\alpha^k_{i,2} \sin\Big(\frac{2\pi t}{\alpha^k_{i,3}}\Big) + 10(t - \alpha^k_{i,4})^3, \quad 1 \leq k \leq d, \tag{12}$$

where the parameters $\alpha^k_{i,\ell}$, $1 \leq \ell \leq 4$ are sampled uniformly on $[0,1]$. Let $(t_0, \ldots, t_{p-1})$ be a regular partition of $[0,1]$ of length p, the matrix of the path values

$$\mathbf{x_i} = (x^k_{i,j})_{\substack{1 \leq k \leq d \\ 1 \leq j \leq p}} \in \mathbb{R}^{d \times p}$$

is then a discretization of X_i on $[0,1]$: $x^k_{i,j} = X^k_{i,t_j}$. It will cause no confusion to use the same notation $\mathbf{x_i}$ to denote the matrix of values of X_i on the partition $(t_0, \ldots, t_{p-1})$ and their piecewise linear interpolation. Fig. 3 shows one sample $\mathbf{x_i}$ with $p = 100$ and $d = 5$.

For any $m^* \in \mathbb{N}$, the output Y_i is now defined as $Y_i = \langle \beta, S^{m^*}(\mathbf{x_i}) \rangle + \varepsilon_i$, where ε_i is a uniform random variable on $[-100, 100]$ and β is given by

$$\beta_j = \frac{1}{1000} u_j, \quad 1 \leq j \leq s_d(m^*),$$

where u_j is sampled uniformly on $[0,1]$. Then, m^* is estimated with the procedure described in Algorithm 1 for different sample sizes n. To select the constant K_{pen}, we use the dimension jump method, that is we plot $\widehat{m}$ as a function of K_{pen}, find the value of K_{pen} that corresponds to the first big jump of $\widehat{m}$ and fix K_{pen} to be equal to twice this value. For a recent account of the theory of slope heuristics, we refer the reader to the review by [1]. For example, for $m^* = 5$,$d = 2$, and $n = 50$, plotting $\widehat{m}$ against K_{pen} yields Fig. 4. In this case, K_{pen} is selected at 100.

We fix $d = 2$, $m^* = 5$, and $K_{\text{pen}} = 20$. For different sample sizes n, we iterate Algorithm 1 twenty times. In Fig. 5, a histogram of the values taken by $\widehat{m}$ is plotted against n. We can see that when n increases, the estimator converges to the true value $m^* = 5$. For $n = 500$ we always pick $\widehat{m} = 5$ over the twenty iterations.

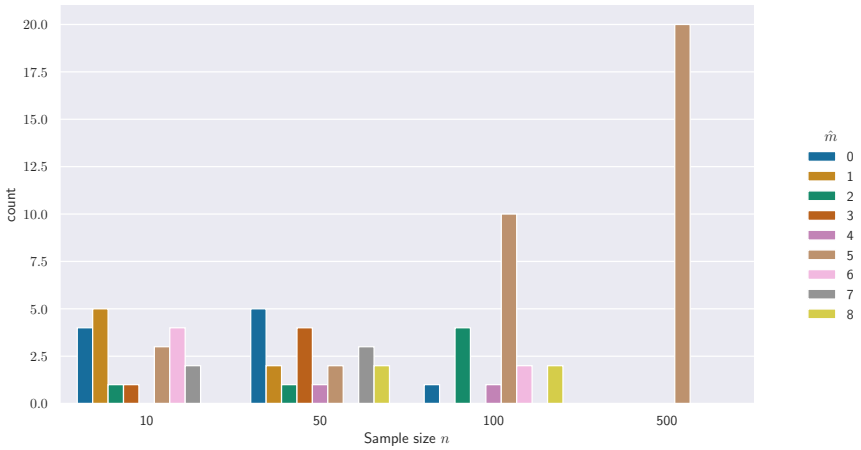


Fig. 5: Histogram of $\widehat{m}$ as a function of n over 20 iterations. The functional predictors X are simulated following (12) and the response Y follows the linear model on signatures with $m^* = 5$. The hyperparameters are $\rho = 0.4$ and $K_{\text{pen}} = 20$.

6. Experimental results

Now that we have a complete procedure at hand, we demonstrate in this section its performance compared to canonical approaches in functional data analysis. We show in particular that it performs better in high dimensions, that is when d is large.

We compare our model to the functional linear model with basis functions presented in Section 2.1, to functional principal component regression (fPCR), and to functional k-nearest neighbors regression. The first models are parametric linear models, while the k-nearest neighbors is nonlinear and nonparametric. Concerning the functional linear model, we consider two choices for the basis $\phi_1, \dots, \phi_K$, namely the B-Spline and Fourier basis [see 40]. Then, the approach consists in projecting the function $X : [0, 1] \to \mathbb{R}^d$ onto the ϕ_is, coordinate by coordinate. The number K of basis functions is selected via cross-validation (with a minimum of 4 and maximum of 14 for Fourier and B-Splines, and a minimum of 1 and a maximum of 6 for the fPCR). For the fPCR, we first smooth the functional covariates with 7 B-Splines. The number of neighbors is selected by cross-validation with a minimum of 1 and a maximum of 9. This procedure is implemented with the Python package `scikit-fda` [38]. In Subsections 6.1 and 6.2, since the focus is on the performance of the signature linear model and to simplify the computations, we select $\widehat{m}$ via cross-validation. For the real-world dataset of Subsection 6.3, it is estimated as described in the previous section.

6.1. Smooth paths

Our goal is to see the influence of the dimension d on the quality of the different models: the signature linear model and the 3 linear functional models. To this end, we simulate some paths following model (12) and predict the average value of the path at the next time step. More precisely, let $(t_0, t_1, \dots, t_p)$ be a partition of $[0, 1]$ of length $p + 1$, then we sample X_i following (12) and let

$$\mathbf{x_i} = (X_{i,t_0} | \cdots | X_{i,t_{p-1}}) \in \mathbb{R}^{d \times p}, \quad Y_i = \frac{1}{d} \sum_{k=1}^{d} X^k_{i,t_p} + \varepsilon_i,$$

where ε_i are i.i.d uniform random variables on $[-1, 1]$. We let d vary on a grid from 1 to 11, simulate some train and test data, and assess the performance of the model with the mean squared error (MSE) on the test set. We iterate the procedure 20 times, which gives, for each model (signature, Fourier, B-Spline, and fPCR), a boxplot of errors, shown in Fig. 6.

It is clear that when d increases, the signature gets better relatively to the 4 other models. We can also note that the B-Spline basis performs best in low dimensions, which is not surprising since the data has a 3rd order polynomial term—see (12). However, even though the B-Spline basis is particularly well-adapted to the data, it is outperformed by the signature linear model when the dimension becomes too large (starting from $d = 7$).

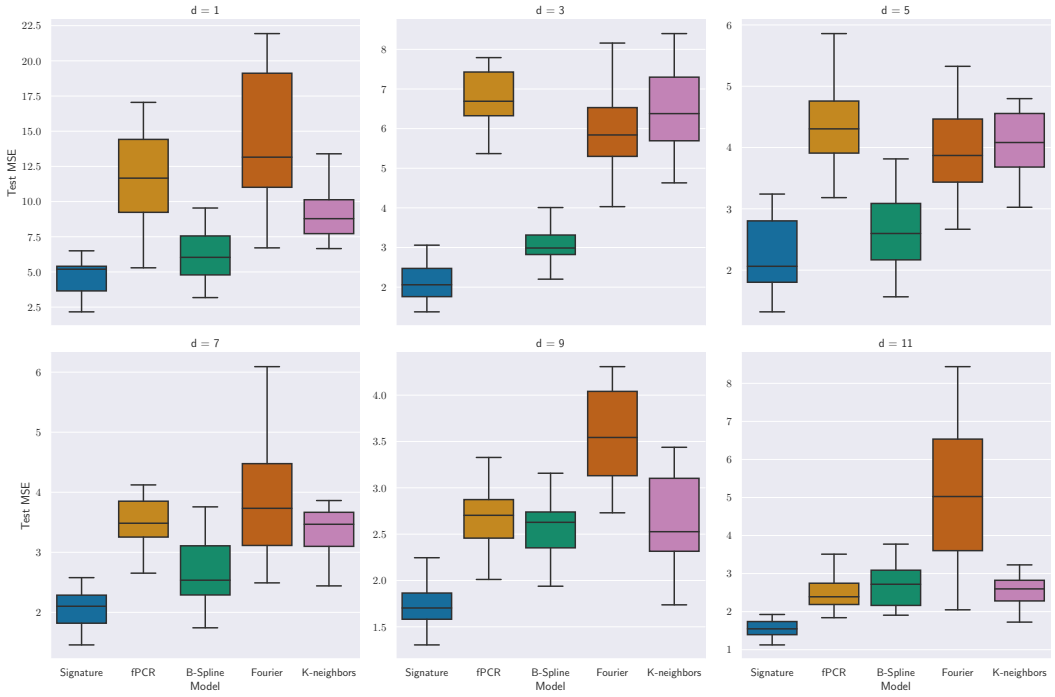


Fig. 6: Test MSE for different regression models when the inputs follow (12) and Y is the mean response at the next time step.

6.2. *Gaussian processes*

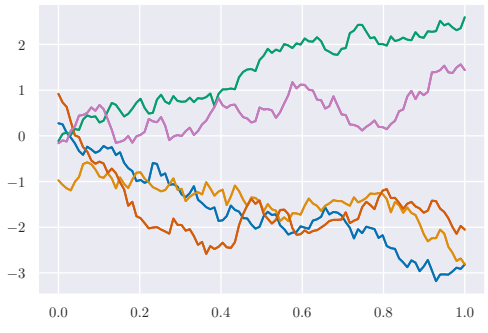


Fig. 7: One sample X from the Gaussian process model (13) with $d = 5$

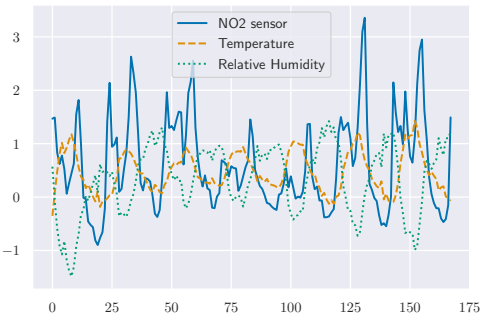


Fig. 8: One sample from the Air Quality dataset.

We continue this simulation study with more complex paths: Gaussian processes. Let $d \geq 1$, $1 \leq i \leq n$, we define the path $X_i = (X_{i,t}^1, \ldots, X_{i,t}^d)_{t\in[0,1]}$ by

$$X_{i,t}^k = \alpha_i^k t + \xi_{i,t}^k, \quad 1 \leq k \leq d, \quad t \in [0,1], \tag{13}$$

where α_i^k is sampled uniformly in $[-3,3]$ and ξ_i^k is a Gaussian process with exponential covariance matrix (with length-scale 1). The response is the norm of the trend slope: $Y_i = \|\alpha_i\| + \varepsilon_i$, where ε_i is uniformly sampled on $[-1,1]$. Fig. 7 shows a realization of X_i with $d = 5$.

We vary the dimension d on the same grid as before and iterate the whole procedure 20 times, which gives the results in Fig. 9. We can see that for these more complicated paths, the signature is better than the 3 linear models even for $d = 1$, but similar to the k-neighbors regression. The difference in performance with B-Spline and Fourier basis increases a lot with d, wheras the k-neighbors model is quite stable.

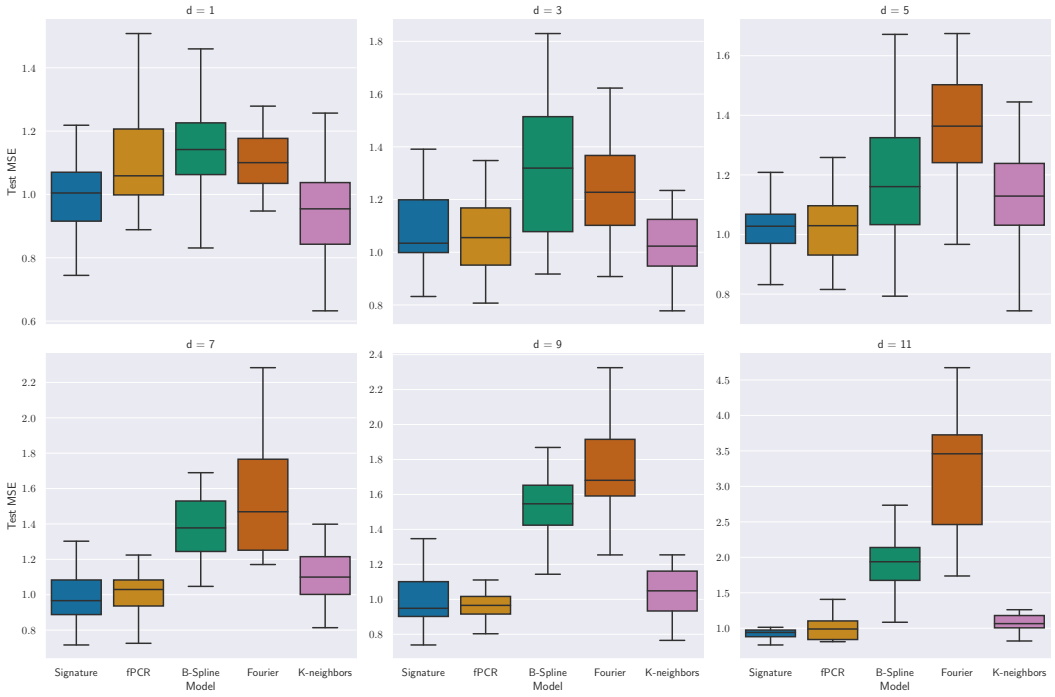


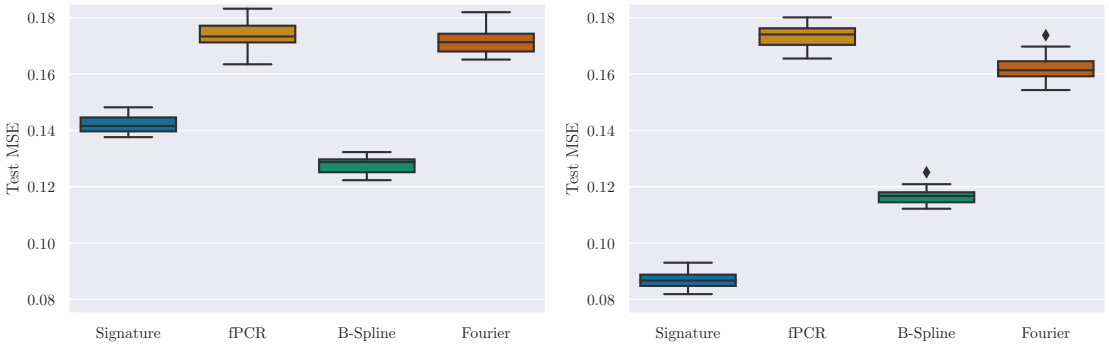
Fig. 9: Test MSE for different regression models when the inputs are gaussian processes with a random linear trend, as defined by (13), and the response is the norm of the trend slope.

6.3. Air quality dataset

We conclude this section with a study of the UCI "Air Quality Data Set" [11]. The data contains the hourly averaged response from 5 metal oxide chemical sensors recorded in a polluted area in Italy during a year (from March 2004 to February 2005). Ground truth concentrations are also included, together with temperature and humidity values. We restrict our analysis to the study of the concentration of nitrogen dioxide (N02), and more precisely to the prediction of the ground truth value of NO2 at the next hour. We consider two situations for the predictor function X: a univariate and a multivariate case. In the univariate case, we are given the values of the sensor recording the concentration of NO2 during the previous 7 days. In this case, the data is in dimension $d = 1$ and sampled at $p = 168$ values. In the multivariate case, we add the information of temperature and relative humidity to X, making it a path in dimension $d = 3$. We show in Fig. 8 one sample in the multivariate case (in the univariate case, X consists only of the blue solid curve).

We perform 20 random train/test splits and show in Fig. 10 a boxplot of the test MSE for each model. We do not consider the k-neighbors regression due to its prohibitive running time for the sample size of this dataset (6156 training samples and 3033 test samples). Indeed, the other models take a few minutes to run while the k-neighbors regression takes two hours. We can see that in the univariate case, the B-Splines perform best. However, when more information is taken into account, that is, in the multivariate case, the signature model has the smallest error. The error of the three other models almost does not change when information about temperature and humidity is added, whereas the error of the signature linear model is divided by 2. We conclude that signatures can extract relevant information from multivariate time series. It should be noted that this type of data is increasingly common in modern applications, as the capabilities for recording and storing data are only getting better.

To conclude, we represent in Fig. 11 the values of the regression vector $\widehat{\beta}_{\widetilde{m}}$ to illustrate its interpretation. We observe that the two largest coefficients are the ones corresponding to $S^{(1)}(X)$ and $S^{(1,1)}(X)$: they both correspond to the variation in NO2 concentration during the period (last value minus initial value). It is therefore not surprising that this is a key quantity to predict the concentration of NO2 at the next hour. We can also comment on the large absolute value of some coefficients of order 2, for example, the one corresponding to $S^{(2,1)}(X)$. The value $S^{(2,1)}(X)$ is the



(a) Univariate case.

(b) Multivariate case.

Fig. 10: Test MSE for different regression models for the Air Quality dataset.

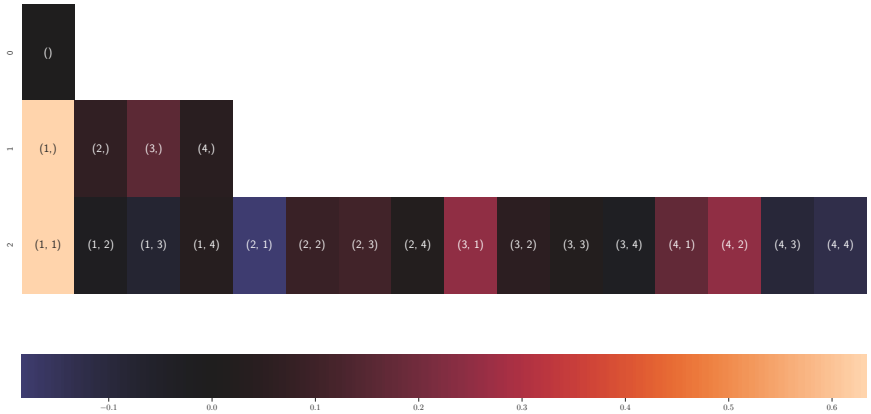


Fig. 11: Heatmap of the first coefficients of $\widehat{\beta}_{\widehat{m}}$ for the Air quality dataset with a truncation order of 4. The vertical axis represents the order of the coefficients: on top the coefficient of order 0, then the 4 coefficients of order 1, then the 16 coefficients of order 2. The color corresponds to the value of the coefficient.

area under the curve (Temperature, NO2), as explained in Fig. 1. The corresponding coefficient, therefore, contains information about the importance of the joint evolution of Temperature and NO2 to predict future concentration. For example, if it is positive, it means that a common increase in Temperature and NO2 will give rise to a larger concentration of NO2. In other words, there is an interaction between Temperature and NO2 concentration. A similar analysis can be done for the curve (Humidity, NO2), which corresponds to the coefficient (3,1). Finally, the large value of the coefficient corresponding to $S^{(4,1)}(X)$, which is equal to the area under the curve (Time, NO2) is also not surprising: it counts the total quantity of concentration of NO2 during the period.

To conclude, the coefficients obtained with the signature linear model have a geometric interpretation, which is often valuable for practical applications. Contrary to the coefficients in traditional functional linear models, they are global measures of interaction between coordinates: there is no time-specific interpretation as there would be for $\beta(t)$ in (1). We refer to Giusti and Lee [16] for more details on the interpretation of signatures, in particular as a measure of causality between different coordinates.

7. Conclusion and perspectives

In this paper, we have provided a complete and ready-to-use methodology to implement the signature linear model. This led us to define a consistent estimator of the signature truncation order. We show on both simulated and real-world datasets that this model performs better than traditional functional linear models when the functional data is vector-valued, especially in high dimensions.

The signature is a flexible tool for summarizing multidimensional time series and can be used in various contexts. This study is just a first step towards understanding how it should be used in a statistical setting and there are a lot of potential extensions. First, we restricted our study to the setting of linear regression, however, signatures are just as relevant in classification or unsupervised learning settings. Moreover, the problem of the high dimension of the regression coefficient, due to its exponential dependence on m, is the major limitation of the signature linear model. In this article, we dealt with it by carefully choosing the truncation order. However, this is not the only option. For example, regularization approaches that induce a sparsity pattern on this coefficient, or the use of a related lower-dimensional object called the logsignature, are two interesting directions.

8. Proofs

Proof of Theorem 1

This section is devoted to the proof of Theorem 1. We will use extensively results from [20]. The next two lemmas first show that it is sufficient to obtain a uniform tail bound on the risk to control the convergence of $\widehat{m}$.

Lemma 1. *For any $m \in \mathbb{N}$,*

$$\left|\widehat{L}_n(m) - L(m)\right| \leq \sup_{\beta \in B_{m,\alpha}} \left|\widehat{\mathcal{R}}_{m,n}(\beta) - \mathcal{R}_m(\beta)\right|.$$

Proof: Introducing $\widehat{\mathcal{R}}_{m,n}(\beta_m^*)$ yields

$$\widehat{L}_n(m) - L(m) = \widehat{\mathcal{R}}_{m,n}(\widehat{\beta}_m) - \mathcal{R}_m(\beta_m^*) = \widehat{\mathcal{R}}_{m,n}(\widehat{\beta}_m) - \widehat{\mathcal{R}}_{m,n}(\beta_m^*) + \widehat{\mathcal{R}}_{m,n}(\beta_m^*) - \mathcal{R}_m(\beta_m^*).$$

Since $\widehat{\beta}_m$ minimises $\widehat{\mathcal{R}}_{m,n}$ over $B_{m,\alpha}$, $\widehat{\mathcal{R}}_{m,n}(\widehat{\beta}_m) - \widehat{\mathcal{R}}_{m,n}(\beta_m^*) \leq 0$, which gives

$$\widehat{L}_n(m) - L(m) \leq \widehat{\mathcal{R}}_{m,n}(\beta_m^*) - \mathcal{R}_m(\beta_m^*) \leq \sup_{\beta \in B_{m,\alpha}} |\widehat{\mathcal{R}}_{m,n}(\beta) - \mathcal{R}_m(\beta)|.$$

In the same manner, $L(m) - \widehat{L}_n(m) \leq \sup_{\beta \in B_{m,\alpha}} |\widehat{\mathcal{R}}_{m,n}(\beta) - \mathcal{R}_m(\beta)|$, which proves the lemma. □

Lemma 2. *For any $m > m^*$, $\mathbb{P}(\widehat{m} = m) \leq \mathbb{P}(2 \sup_{\beta \in B_{m,\alpha}} |\widehat{\mathcal{R}}_{m,n}(\beta) - \mathcal{R}(\beta)| \geq \mathrm{pen}_n(m) - \mathrm{pen}_n(m^*))$.*

Proof: For any $m \in \mathbb{N}$,

$$\mathbb{P}(\widehat{m} = m) \leq \mathbb{P}\left(\widehat{L}_n(m) + \mathrm{pen}_n(m) \leq \widehat{L}_n(m^*) + \mathrm{pen}_n(m^*)\right) = \mathbb{P}\left(\widehat{L}_n(m^*) - \widehat{L}_n(m) \geq \mathrm{pen}_n(m) - \mathrm{pen}_n(m^*)\right).$$

Recall that, by definition of model (5), $m \mapsto L(m)$ is a decreasing function and that its minimum is attained at $m = m^*$. Therefore, for any $m \in \mathbb{N}$, $L(m^*) \leq L(m)$, and Lemma 1 yields

$$\begin{aligned}\widehat{L}_n(m^*) - \widehat{L}_n(m) &= \widehat{L}_n(m^*) - L(m^*) + L(m^*) - L(m) + L(m) - \widehat{L}_n(m) \leq \widehat{L}_n(m^*) - L(m^*) + L(m) - \widehat{L}_n(m) \\ &\leq \sup_{\beta \in B_{m^*,\alpha}} |\widehat{\mathcal{R}}_{m,n}(\beta) - \mathcal{R}_m(\beta)| + \sup_{\beta \in B_{m,\alpha}} |\widehat{\mathcal{R}}_{m,n}(\beta) - \mathcal{R}_m(\beta)|.\end{aligned}$$

For $m > m^*$, $B_{m^*,\alpha} \subset B_{m,\alpha}$, which gives $\widehat{L}_n(m^*) - \widehat{L}_n(m) \leq 2\sup_{\beta \in B_{m,\alpha}} |\widehat{\mathcal{R}}_{m,n}(\beta) - \mathcal{R}_m(\beta)|$, and the proof is complete. □

From now on, we denote by $Z_{m,n}$ the centered empirical risk for signatures truncated at m: for any $\beta \in B_{m,\alpha}$,

$$Z_{m,n}(\beta) = \widehat{\mathcal{R}}_{m,n}(\beta) - \mathcal{R}_m(\beta) = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle \beta, S^m(X_i)\rangle\right)^2 - \mathbb{E}(Y - \langle \beta, S^m(X)\rangle)^2.$$

We will now derive a uniform tail bound on $Z_{m,n}(\beta)$, which is the main result needed to prove Theorem 1. In a nutshell, we show that $(Z_{m,n}(\beta))_{\beta \in B_{m,\alpha}}$ is a subgaussian process for some appropriate distance, and then use a chaining tail inequality [20, Theorem 5.29] on $Z_{m,n}$.

Lemma 3. *Under the assumptions (H_α) and (H_K), for any $m \in \mathbb{N}$, the process $(Z_{m,n}(\beta))_{\beta \in B_{m,\alpha}}$ is subgaussian for the distance*

$$D(\beta, \gamma) = \frac{K}{\sqrt{n}} \|\beta - \gamma\|, \tag{14}$$

where the constant K is defined by (6).

Proof: By definition, it is clear that $\mathbb{E}Z_{m,n}(\beta) = 0$ for any $\beta \in B_{m,\alpha}$. Let $\ell_{(X,Y)} \colon B_{m,\alpha} \to \mathbb{R}$ be given by $\ell_{(X,Y)}(\beta) = (Y - \langle \beta, S^m(X) \rangle)^2$. We first prove that $\ell_{(X,Y)}$ is K-Lipschitz. For any $\beta, \gamma \in B_{m,\alpha}$,

$$\begin{aligned}
|\ell_{(X,Y)}(\beta) - \ell_{(X,Y)}(\gamma)| &= \left| (Y - \langle \beta, S^m(X) \rangle)^2 - (Y - \langle \gamma, S^m(X) \rangle)^2 \right| \\
&\leq 2 \max \left(\left| Y - \langle \beta, S^m(X) \rangle \right|, \left| Y - \langle \gamma, S^m(X) \rangle \right| \right) \times \left| \langle \beta - \gamma, S^m(X) \rangle \right| \\
&\quad \text{(because } |a^2 - b^2| \leq 2 \max(|a|, |b|)|a - b| \text{)} \\
&\leq 2 \max \left(\left| Y - \langle \beta, S^m(X) \rangle \right|, \left| Y - \langle \gamma, S^m(X) \rangle \right| \right) \times \left\| S^m(X) \right\| \|\beta - \gamma\| \\
&\quad \text{(by the Cauchy-Schwartz inequality).}
\end{aligned}$$

Moreover, by the triangle inequality and Cauchy-Schwartz inequality,

$$\left| Y - \langle \beta, S^m(X) \rangle \right| \leq |Y| + \left\| S^m(X) \right\| \|\beta\| \leq K_Y + \alpha \left\| S^m(X) \right\|,$$

and, by Proposition 3, $\left\| S^m(X) \right\| \leq e^{\|X\|_{TV}} \leq e^{K_X}$. Consequently, $\left| Y - \langle \beta, S^m(X) \rangle \right| \leq K_Y + \alpha e^{K_X}$, and

$$\left| \ell_{(X,Y)}(\beta) - \ell_{(X,Y)}(\gamma) \right| \leq 2(K_Y + \alpha e^{K_X}) e^{K_X} \|\beta - \gamma\| = K \|\beta - \gamma\|.$$

Therefore, by Hoeffding's lemma [20, Lemma 3.6], $\ell_{(X,Y)}(\beta) - \ell_{(X,Y)}(\gamma)$ is a subgaussian random variable with variance proxy $K^2 \|\beta - \gamma\|^2$, which gives, for $\lambda \geq 0$,

$$\mathbb{E} \exp \left(\lambda \Big(\ell_{(X,Y)}(\beta) - \ell_{(X,Y)}(\gamma) - \mathbb{E}(\ell_{(X,Y)}(\beta) - \ell_{(X,Y)}(\gamma)) \Big) \right) \leq \exp \left(\frac{\lambda^2 K^2 \|\beta - \gamma\|^2}{2} \right).$$

From this, it follows that

$$\begin{aligned}
\mathbb{E} e^{\lambda \left(Z_{m,n}(\beta) - Z_{m,n}(\gamma) \right)} &= \mathbb{E} \exp \left(\frac{\lambda}{n} \sum_{i=1}^{n} \ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma) - \mathbb{E} \Big(\ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma) \Big) \right) \\
&= \prod_{i=1}^{n} \mathbb{E} \exp \left(\frac{\lambda}{n} \Big(\ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma) - \mathbb{E}(\ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma)) \Big) \right) \\
&\leq \exp \left(\frac{\lambda^2 K^2 \|\beta - \gamma\|^2}{2n} \right) = \exp \left(\frac{\lambda^2 D(\beta, \gamma)^2}{2} \right),
\end{aligned}$$

where $D(\beta, \gamma) = \frac{K \|\beta - \gamma\|}{\sqrt{n}}$, which completes the proof. □

We can now derive a maximal tail inequality for $Z_{m,n}(\beta)$.

Proposition 5. *Under the assumptions (H_α) and (H_K), for any $m \in \mathbb{N}$, $x > 0$, $\beta_0 \in B_{m,\alpha}$,*

$$\mathbb{P} \Big(\sup_{\beta \in B_{m,\alpha}} Z_{m,n}(\beta) \geq 108 \sqrt{\pi} K \alpha \sqrt{\frac{s_d(m)}{n}} + Z_{m,n}(\beta_0) + x \Big) \leq 36 \exp \Big(- \frac{x^2 n}{144 K^2 \alpha^2} \Big),$$

where the constant K is defined by (6).

Proof: By Lemma 3, $Z_{m,n}$ is a subgaussian process for D, defined by (14). So, we may apply Theorem 5.29 of [20] to $Z_{m,n}$ on the metric space $(B_{m,\alpha}, D)$:

$$\mathbb{P}\Big(\sup_{\beta \in B_{m,\alpha}} Z_{m,n}(\beta) - Z_{m,n}(\beta_0) \geq 36 \int_0^\infty \sqrt{\log(N(\varepsilon, B_{m,\alpha}, D))} d\varepsilon + x \Big) \leq 36 \exp\Big(- \frac{x^2 n}{36 \times 4K^2\alpha^2} \Big),$$

where $N(\varepsilon, B_{m,\alpha}, D)$ is the ε-covering number of $B_{m,\alpha}$ with respect to D, and where we use that

$$\operatorname{diam}(B_{m,\alpha}) = \frac{2K\alpha}{\sqrt{n}}.$$

Moreover, $N(\varepsilon, B_{m,\alpha}, D) = N(\frac{\sqrt{n}}{K}\varepsilon, B_{m,\alpha}, \|\cdot\|)$, and so, by Lemma 5.13 of van Handel [20],

$$N(\varepsilon, B_{m,\alpha}, D) \leq \left(\frac{3K\alpha}{\sqrt{n}\varepsilon} \right)^{s_d(m)} \text{ if } \varepsilon < \frac{K\alpha}{\sqrt{n}},$$

and $N(\varepsilon, B_{m,\alpha}, D) = 1$ otherwise. Therefore,

$$\begin{aligned} \int_0^\infty \sqrt{\log(N(\varepsilon, B_{m,\alpha}, D))} d\varepsilon &= \int_0^{\frac{K\alpha}{\sqrt{n}}} \sqrt{\log(N(\varepsilon, B_{m,\alpha}, D))} d\varepsilon \\ &\leq \int_0^{\frac{K\alpha}{\sqrt{n}}} \sqrt{s_d(m) \log\left(\frac{3K\alpha}{\sqrt{n}\varepsilon} \right)} d\varepsilon \\ &\leq 3K\alpha \sqrt{\frac{s_d(m)}{n}} \int_0^\infty 2x^2 \exp\left(-x^2\right) dx = 3K\alpha \sqrt{\frac{s_d(m)}{n}} \sqrt{\pi}, \end{aligned} \tag{15}$$

where in the second inequality we use the change of variable $x = \sqrt{\log\left(\frac{2K\alpha}{\sqrt{n}\varepsilon}\right)}$. □

Since $\mathbb{P}(\widehat{m} \neq m^*) = \mathbb{P}(\widehat{m} > m^*) + \mathbb{P}(\widehat{m} < m^*)$, we divide the proof into two cases. Let us first consider $m > m^*$ in the next proposition.

Proposition 6. *Let $0 < \rho < \frac{1}{2}$, and $\operatorname{pen}_n(m)$ be defined by (7): $\operatorname{pen}_n(m) = K_{\text{pen}} n^{-\rho} \sqrt{s_d(m)}$. Let n_1 be the smallest integer satisfying*

$$n_1 \geq \left(\frac{432\sqrt{\pi} K\alpha \sqrt{s_d(m^*+1)}}{K_{\text{pen}}(\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)})} \right)^{1/(\frac{1}{2}-\rho)}. \tag{16}$$

Then, under the assumptions (H_α) and (H_K), for any $m > m^*$, $n \geq n_1$,

$$\mathbb{P}\left(\widehat{m} = m\right) \leq 74 \exp\left(- C_3(n^{1-2\rho} + s_d(m))\right),$$

where the constant C_3 is defined by

$$C_3 = \frac{K_{\text{pen}}^2 d^{m^*+1}}{128 s_d(m^*+1)(72K^2\alpha^2 + K_Y^2)}.$$

Proof: Let

$$u_{m,n} = \frac{1}{2}(\operatorname{pen}_n(m) - \operatorname{pen}_n(m^*)) = \frac{K_{\text{pen}}}{2} n^{-\rho}\Big(\sqrt{s_d(m)} - \sqrt{s_d(m^*)}\Big).$$

As $m \mapsto \operatorname{pen}_n(m)$ is increasing in m, it is clear that $u_{m,n} > 0$ for any $m > m^*$. From Lemma 2, we see that

$$\mathbb{P}\left(\widehat{m} = m\right) \leq \mathbb{P}\Big(\sup_{\beta \in B_{m,\alpha}} \left|Z_{m,n}(\beta)\right| > u_{m,n}\Big) = \mathbb{P}\Big(\sup_{\beta \in B_{m,\alpha}} Z_{m,n}(\beta) > u_{m,n}\Big) + \mathbb{P}\Big(\sup_{\beta \in B_{m,\alpha}} (-Z_{m,n}(\beta)) > u_{m,n}\Big).$$

We focus on the first term of the inequality, the second can be handled in the same way since Proposition 5 also holds when $Z_{m,n}(\beta)$ is replaced by $-Z_{m,n}(\beta)$. Let β_0 be a fixed point in $B_{m,\alpha}$ that will be chosen later, we have

$$\begin{aligned}\mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) > u_{m,n}\Big) &= \mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) > u_{m,n},\, Z_{m,n}(\beta_0) \le \frac{u_{m,n}}{2}\Big) + \mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) > u_{m,n},\, Z_{m,n}(\beta_0) > \frac{u_{m,n}}{2}\Big)\\ &\le \mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) > \frac{u_{m,n}}{2} + Z_{m,n}(\beta_0)\Big) + \mathbb{P}\Big(Z_{m,n}(\beta_0) > \frac{u_{m,n}}{2}\Big). \qquad (17)\end{aligned}$$

We treat each term separately. The first one is handled by Proposition 5. To this end, we need to ensure that $\frac{u_{m,n}}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}}$ is positive. By definition,

$$\begin{aligned}\frac{u_{m,n}}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}} &= \frac{K_{\text{pen}}}{2}n^{-\rho}\Big(\sqrt{s_d(m)} - \sqrt{s_d(m^*)}\Big) - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}}\\ &= \sqrt{s_d(m)}n^{-\rho}\frac{K_{\text{pen}}}{2}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m)}} - \frac{2\times 108\sqrt{\pi}K\alpha}{K_{\text{pen}}}n^{\rho-\frac{1}{2}}\Big).\\ &\ge \sqrt{s_d(m)}n^{-\rho}\frac{K_{\text{pen}}}{2}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}} - \frac{216\sqrt{\pi}K\alpha}{K_{\text{pen}}}n^{\rho-\frac{1}{2}}\Big).\end{aligned}$$

Let $n_1 \in \mathbb{N}$ be such that

$$1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}} - \frac{216\sqrt{\pi}K\alpha}{K_{\text{pen}}}n_1^{\rho-\frac{1}{2}} > \frac{1}{2}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\Big) \Leftrightarrow n_1 > \Big(\frac{432\sqrt{\pi}K\alpha\sqrt{s_d(m^*+1)}}{K_{\text{pen}}(\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)})}\Big)^{1/(\frac{1}{2}-\rho)},$$

then, for any $n \ge n_1$,

$$\frac{u_{m,n}}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}} \ge \sqrt{s_d(m)}n^{-\rho}\frac{K_{\text{pen}}}{4}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\Big) > 0.$$

Hence, Proposition 5 applied to $x = \frac{u_{m,n}}{2} - 108\sqrt{\pi}K\alpha\sqrt{\frac{s_d(m)}{n}}$ now shows that, for $n \ge n_1$,

$$\begin{aligned}\mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) > \frac{u_{m,n}}{2} + Z_{m,n}(\beta_0)\Big) &\le 36\exp\Big(-\frac{n}{144K^2\alpha^2}\Big(\frac{u_{m,n}}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}}\Big)^2\Big)\\ &\le 36\exp\Big(-\frac{s_d(m)n^{1-2\rho}K_{\text{pen}}^2}{144K^2\alpha^2\times 16}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\Big)^2\Big)\\ &= 36\exp\Big(-\kappa_1 s_d(m)n^{1-2\rho}\Big), \qquad (18)\end{aligned}$$

where

$$\kappa_1 = \frac{K_{\text{pen}}^2}{2304K^2\alpha^2}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\Big)^2.$$

We now turn to the second term of (17). Since $\big|Y - \langle\beta_0, S^m(X)\rangle\big|^2 \le (K_Y + \|\beta_0\|e^{K_X})^2$ a.s., Hoeffding's inequality yields,

for $n \geq n_1$,

$$\begin{aligned}
\mathbb{P}\Big(Z_{m,n}(\beta_0) > \frac{u_{m,n}}{2}\Big) &\leq \exp\Big(- \frac{n u_{m,n}^2}{8(K_Y + \|\beta_0\| e^{K_X})^2}\Big) \\
&= \exp\Big(- \frac{n^{1-2\rho} K_{\text{pen}}^2 \Big(\sqrt{s_d(m)} - \sqrt{s_d(m^*)}\Big)^2}{32(K_Y + \|\beta_0\| e^{K_X})^2}\Big) \\
&\leq \exp\Big(- \frac{n^{1-2\rho} K_{\text{pen}}^2 s_d(m)}{32(K_Y + \|\beta_0\| e^{K_X})^2}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\Big)^2\Big) \\
&= \exp\Big(-\kappa_2 n^{1-2\rho} s_d(m)\Big), \qquad (19)
\end{aligned}$$

where

$$\kappa_2 = \frac{K_{\text{pen}}^2}{32(K_Y + \|\beta_0\| e^{K_X})^2}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\Big)^2.$$

Combining (18) with (19), we obtain

$$\begin{aligned}
\mathbb{P}\Big(\sup_{\beta \in B_{m,\alpha}} Z_{m,n}(\beta) > u_{m,n}\Big) &\leq 36 \exp\big(- \kappa_1 n^{1-2\rho} s_d(m)\big) + \exp\big(- \kappa_2 n^{1-2\rho} s_d(m)\big) \leq 37 \exp\big(- \kappa_3 n^{1-2\rho} s_d(m)\big) \\
&\leq 37 \exp\Big(- \frac{\kappa_3}{2}(n^{1-2\rho} + s_d(m))\Big),
\end{aligned}$$

where $\kappa_3 = \min(\kappa_1, \kappa_2)$. The same proof works for the process $(-Z_{m,n}(\beta))$, and consequently

$$\mathbb{P}\left(\widehat{m} = m\right) \leq 2 \times 37 \exp\Big(- \frac{\kappa_3}{2}(n^{1-2\rho} + s_d(m))\Big).$$

We are left with the task of choosing an optimal β_0. Since

$$\kappa_3 = \min(\kappa_1, \kappa_2) = \frac{K_{\text{pen}}^2}{32}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\Big)^2 \min\Big(\frac{1}{72K^2\alpha^2}, \frac{1}{(K_Y + \|\beta_0\| e^{K_X})^2}\Big),$$

it is clear that κ_3 is maximal at $\beta_0 = 0$, which yields

$$\kappa_3 = \frac{K_{\text{pen}}^2}{32}\Big(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\Big)^2 \min\Big(\frac{1}{72K^2\alpha^2}, \frac{1}{K_Y^2}\Big).$$

Noting that

$$\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)} = \sqrt{d^{m^*+1} + s_d(m^*)} - \sqrt{s_d(m^*)} \geq \sqrt{\frac{d^{m^*+1}}{2}},$$

where we have used the fact that for $a, b \geq 0$, $\sqrt{a} + \sqrt{b} \geq \sqrt{2}\sqrt{a+b}$, letting

$$C_3 = \frac{1}{2} \times \frac{K_{\text{pen}}^2 d^{m^*+1}}{64 s_d(m^*+1)(72K^2\alpha^2 + K_Y^2)}$$

completes the proof. □

To treat the case $m < m^*$, we need a rate of convergence of $\widehat{L}_n$. This can be obtained with arguments similar to the previous proof.

Proposition 7. *For any $\varepsilon > 0$, $m \in \mathbb{N}$, let $n_2 \in \mathbb{N}$ be the smallest integer such that*

$$n_2 \geq \frac{432^2 K^2 \pi \alpha^2 s_d(m)}{\varepsilon^2}. \qquad (20)$$

Then, for any $n \geq n_2$,

$$\mathbb{P}(|\widehat{L}_n(m) - L(m)| > \varepsilon) \leq 74 \exp\big(-C_4 n\varepsilon^2\big),$$

where the constant C_4 is defined by

$$C_4 = \Big(2(1152K^2\alpha^2 + K_Y^2)\Big)^{-1}. \tag{21}$$

Proof: By Lemma 1,

$$\mathbb{P}(|\widehat{L}_n(m) - L(m)| > \varepsilon) \leq \mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} |Z_{m,n}(\beta)| > \varepsilon\Big) = \mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) > \varepsilon\Big) + \mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} (-Z_{m,n}(\beta)) > \varepsilon\Big).$$

Let us fix $\beta_0 \in B_{m,\alpha}$, we can now proceed as in Proposition 6. Since, for $n \geq n_2$,

$$\frac{\varepsilon}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}} > \frac{\varepsilon}{4} > 0,$$

Hoeffing's inequality and Proposition 5 show that

$$\begin{aligned}
\mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) > \varepsilon\Big) &\leq \mathbb{P}\Big(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) > \frac{\varepsilon}{2} + Z_{m,n}(\beta_0)\Big) + \mathbb{P}\Big(Z_{m,n}(\beta_0) > \frac{\varepsilon}{2}\Big) \\
&\leq 36\exp\Big(-\frac{n}{144K^2\alpha^2}\Big(\frac{\varepsilon}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}}\Big)^2\Big) + \exp\Big(-\frac{n\varepsilon^2}{2(K_Y + \|\beta_0\|e^{K_X})^2}\Big) \\
&\leq 36\exp\Big(-\frac{n\varepsilon^2}{2304K^2\alpha^2}\Big) + \exp\Big(-\frac{n\varepsilon^2}{2(K_Y + \|\beta_0\|e^{K_X})^2}\Big) \leq 37\exp\big(-\kappa_4 n\varepsilon^2\big),
\end{aligned}$$

where

$$\kappa_4 = \min\Big(\frac{1}{2304K^2\alpha^2}, \frac{1}{2(K_Y + \|\beta_0\|e^{K_X})^2}\Big).$$

The same analysis can be done to $(-Z_{m,n}(\beta))$, and so $\mathbb{P}\big(|\widehat{L}_n(m) - L(m)| > \varepsilon\big) \leq 74\exp\big(-\kappa_4 n\varepsilon^2\big)$. Moreover, taking $\beta_0 = 0$ gives

$$\kappa_4 = \min\left(\frac{1}{2304K^2\alpha^2}, \frac{1}{2(K_Y + \|\beta_0\|e^{K_X})^2}\right) \geq \frac{1}{2(1152K^2\alpha^2 + K_Y^2)} = C_4,$$

which completes the proof. □

This allows us to treat the case $m < m^*$.

Proposition 8. *Let $0 < \rho < \frac{1}{2}$ and $\mathrm{pen}_n(m)$ be defined by (7). Let n_3 be the smallest integer satisfying*

$$n_3 \geq \Big(\frac{2\sqrt{s_d(m^*)}}{L(m^*-1) - \sigma^2}(432K\alpha\sqrt{\pi} + K_{\mathrm{pen}})\Big)^{1/\rho}. \tag{22}$$

Then, under the assumptions (H_α) and (H_K), for any $m < m^*$, $n \geq n_3$,

$$\mathbb{P}\left(\widehat{m} = m\right) \leq 148\exp\Big(-n\frac{C_4}{4}\big(L(m) - L(m^*) - \mathrm{pen}_n(m^*) + \mathrm{pen}_n(m)\big)^2\Big),$$

where the constant C_4 is defined by (21).

Proof: This is a consequence of Proposition 7. For any $m < m^*$,

$$\begin{aligned}
\mathbb{P}(\widehat{m} = m) &\leq \mathbb{P}\Big(\widehat{L}_n(m) - \widehat{L}_n(m^*) \leq \mathrm{pen}_n(m^*) - \mathrm{pen}_n(m)\Big) \\
&= \mathbb{P}\Big(\widehat{L}_n(m^*) - L(m^*) + L(m) - \widehat{L}_n(m) \geq L(m) - L(m^*) - \big(\mathrm{pen}_n(m^*) - \mathrm{pen}_n(m)\big)\Big) \\
&\leq \mathbb{P}\Big(\big|\widehat{L}_n(m) - L(m)\big| \geq \frac{1}{2}\big(L(m) - L(m^*) - \mathrm{pen}_n(m^*) + \mathrm{pen}_n(m)\big)\Big) \\
&\quad + \mathbb{P}\Big(\big|\widehat{L}_n(m^*) - L(m^*)\big| \geq \frac{1}{2}\big(L(m) - L(m^*) - \mathrm{pen}_n(m^*) + \mathrm{pen}_n(m)\big)\Big).
\end{aligned}$$

In order to apply Proposition 7, we first need to ensure that $L(m)-L(m^*)-\text{pen}_n(m^*)+\text{pen}_n(m)$ is strictly positive. Recall that $m \mapsto L(m)$ is a decreasing function, minimal at $m = m^*$ and then bounded by σ^2. Recall also that $m \mapsto \text{pen}_n(m)$ is strictly increasing. This gives, for $m < m^*$:

$$L(m) - L(m^*) - \text{pen}_n(m^*) + \text{pen}_n(m) > L(m^* - 1) - \sigma^2 - K_{\text{pen}} n^{-\rho} \sqrt{s_d(m^*)}.$$

This implies that it is enough that

$$L(m^* - 1) - \sigma^2 - K_{\text{pen}} n^{-\rho} \sqrt{s_d(m^*)} > \frac{1}{2}(L(m^* - 1) - \sigma^2) \tag{23}$$

to ensure that $L(m) - L(m^*) - \text{pen}_n(m^*) + \text{pen}_n(m) > 0$. This yields a first condition on n_3:

$$n_3 \geq \left(\frac{2K_{\text{pen}} \sqrt{s_d(m^*)}}{L(m^* - 1) - \sigma^2}\right)^{\frac{1}{\rho}}. \tag{24}$$

However, to apply Proposition 7, we also need n_3 to satisfy (20) , which writes

$$n_3 \geq \frac{432^2 K^2 \pi \alpha^2 s_d(m)}{(L(m) - L(m^*) - \text{pen}_n(m^*) + \text{pen}_n(m))^2}.$$

If n_3 satisfies (24), we can bound the right-hand side uniformly in m:

$$\frac{432^2 K^2 \pi \alpha^2 s_d(m)}{(L(m) - L(m^*) - \text{pen}_n(m^*) + \text{pen}_n(m))^2} \leq \frac{4 \times 432^2 K^2 \pi \alpha^2 s_d(m^*)}{(L(m^* - 1) - \sigma^2)^2} = \left(\frac{2 \times 432 K \alpha \sqrt{\pi s_d(m^*)}}{L(m^* - 1) - \sigma^2}\right)^2.$$

We can assume that this quantity is larger than 1, as otherwise the condition on n_3 will be trivially satisfied. Then, as $\rho < \frac{1}{2}$, it is enough for n_3 to satisfy

$$n_3 \geq \max\left(\frac{2K_{\text{pen}} \sqrt{s_d(m^*)}}{L(m^* - 1) - \sigma^2}, \frac{2 \times 432 K \alpha \sqrt{\pi s_d(m^*)}}{L(m^* - 1) - \sigma^2}\right)^{1/\rho},$$

or in a more compact form that

$$n_3 \geq \left(\frac{2(K_{\text{pen}} + 432 K \alpha \sqrt{\pi}) \sqrt{s_d(m^*)}}{L(m^* - 1) - \sigma^2}\right)^{1/\rho}.$$

We conclude by applying Proposition 7 to both terms with

$$\varepsilon = \frac{1}{2}(L(m) - L(m^*) - \text{pen}_n(m^*) - \text{pen}_n(m)).$$

□

We are now in a position to prove Theorem 1.

Proof: [Proof of Theorem 1] The result is a consequence of Propositions 6 and 8. For this, we first need to ensure that the conditions on n (16) and (22) are satisfied. Thus, we need to bound

$$M = \max\left(\left(\frac{2\sqrt{s_d(m^*)}}{L(m^* - 1) - \sigma^2}(432 K \alpha \sqrt{\pi} + K_{\text{pen}})\right)^{1/\rho}, \left(\frac{432 \sqrt{\pi} K \alpha \sqrt{s_d(m^* + 1)}}{K_{\text{pen}}(\sqrt{s_d(m^* + 1)} - \sqrt{s_d(m^*)})}\right)^{1/(\frac{1}{2} - \rho)}\right).$$

If $\tilde{\rho} = \min(\rho, \frac{1}{2} - \rho)$, then

$$\begin{aligned} M &\leq \left((432 K \alpha \sqrt{\pi} + K_{\text{pen}}) \sqrt{s_d(m^* + 1)} \max\left(\frac{2}{L(m^* - 1) - \sigma^2}, \frac{1}{K_{\text{pen}}(\sqrt{s_d(m^* + 1)} - \sqrt{s_d(m^*)})}\right)\right)^{1/\tilde{\rho}} \\ &\leq \left((432 K \alpha \sqrt{\pi} + K_{\text{pen}}) \sqrt{s_d(m^* + 1)}\left(\frac{2}{L(m^* - 1) - \sigma^2} + \frac{\sqrt{2}}{K_{\text{pen}} \sqrt{d^{m^*+1}}}\right)\right)^{1/\tilde{\rho}}. \end{aligned}$$

Therefore, condition (8) implies that (16) and (22) are satisfied. Splitting the probability $\mathbb{P}(\widehat{m} \neq m^*)$ into two terms now gives

$$\mathbb{P}(\widehat{m} \neq m^*) = \mathbb{P}(\widehat{m} > m^*) + \mathbb{P}(\widehat{m} < m^*) \leq \sum_{m>m^*} \mathbb{P}(\widehat{m} = m) + \sum_{m<m^*} \mathbb{P}(\widehat{m} = m).$$

On the one hand, Theorem 6 shows that, for $n \geq n_0$,

$$\sum_{m>m^*} \mathbb{P}(\widehat{m} = m) \leq 74 e^{-C_3 n^{1-2\rho}} \sum_{m>m^*} e^{-C_3 s_d(m)},$$

and, on the other hand, Proposition 8 gives

$$\sum_{m<m^*} \mathbb{P}(\widehat{m} = m) \leq 148 \sum_{m=0}^{m^*-1} \exp\Big(-\frac{C_4}{4} n(L(m) - L(m^*) - \mathrm{pen}_n(m^*) + \mathrm{pen}_n(m))\Big) \leq 148 m^* \exp\big(-\frac{C_4}{8} n(L(m^*-1) - \sigma^2)\big),$$

where we have used that for $n \geq n_0$, (23) is true. Letting

$$\kappa_5 = \min\Big(C_3, \frac{C_4(L(m^*-1) - \sigma^2)}{8}\Big)$$

yields

$$\mathbb{P}(\widehat{m} \neq m^*) \leq 74 e^{-\kappa_5 n^{1-2\rho}} \sum_{m>0} e^{-C_3 s_d(m)} + 148 m^* e^{-\kappa_5 n} \leq C_1 e^{-\kappa_5 n^{1-2\rho}},$$

where

$$C_1 = 74 \sum_{m>0} e^{-C_3 s_d(m)} + 148 m^*.$$

To complete the proof, it remains to find a lower bound on κ_5:

$$\begin{aligned}\kappa_5 = \min\Big(C_3, \frac{C_4(L(m^*-1) - \sigma^2)}{8}\Big) &= \min\Big(\frac{K_{\mathrm{pen}}^2 d^{m^*+1}}{128 s_d(m^*+1)(72K^2\alpha^2 + K_Y^2)}, \frac{L(m^*-1) - \sigma^2}{16(1152K^2\alpha^2 + K_Y^2)}\Big) \\ &\geq \frac{1}{16(1152K^2\alpha^2 + K_Y^2)} \min\Big(\frac{K_{\mathrm{pen}}^2 d^{m^*+1}}{8 s_d(m^*+1)}, L(m^*-1) - \sigma^2\Big) = C_2. \qquad (25)\end{aligned}$$

□

Proof of Corollary 1

First, let us note that $\mathbb{E}(\langle \widehat{\beta}_{\widehat{m}}, S^{\widehat{m}}(X)\rangle - \langle \beta^*_{m^*}, S^{m^*}(X)\rangle)^2 = \mathbb{E}(\mathcal{R}_{\widehat{m}}(\widehat{\beta}_{\widehat{m}})) - \mathcal{R}_{m^*}(\beta^*_{m^*})$. Moreover, we have a.s.

$$\begin{aligned}\mathcal{R}_{\widehat{m}}(\widehat{\beta}_{\widehat{m}}) - \mathcal{R}_{m^*}(\beta^*_{m^*}) &= \mathcal{R}_{\widehat{m}}(\widehat{\beta}_{\widehat{m}}) - \mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) + \mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) - \mathcal{R}_{m^*}(\beta^*_{m^*}) \\ &= \mathcal{R}_{\widehat{m}}(\widehat{\beta}_{\widehat{m}}) - \widehat{\mathcal{R}}_{\widehat{m},n}(\widehat{\beta}_{\widehat{m}}) + \widehat{\mathcal{R}}_{\widehat{m},n}(\widehat{\beta}_{\widehat{m}}) - \widehat{\mathcal{R}}_{\widehat{m},n}(\beta^*_{\widehat{m}}) + \widehat{\mathcal{R}}_{\widehat{m},n}(\beta^*_{\widehat{m}}) - \mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) + \mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) - \mathcal{R}_{m^*}(\beta^*_{m^*}) \\ &\leq \mathcal{R}_{\widehat{m}}(\widehat{\beta}_{\widehat{m}}) - \widehat{\mathcal{R}}_{\widehat{m},n}(\widehat{\beta}_{\widehat{m}}) + \widehat{\mathcal{R}}_{\widehat{m},n}(\beta^*_{\widehat{m}}) - \mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) + \mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) - \mathcal{R}_{m^*}(\beta^*_{m^*}) \\ &\leq 2 \sup_{\beta \in B_{\widehat{m},\alpha}} |\widehat{\mathcal{R}}_{\widehat{m},n}(\beta) - \mathcal{R}_{\widehat{m}}(\beta)| + \mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) - \mathcal{R}_{m^*}(\beta^*_{m^*}).\end{aligned}$$

We decompose the proof into two lemmas.

Lemma 4.

$$\mathbb{E}[\sup_{\beta \in B_{\widehat{m},\alpha}} |\widehat{\mathcal{R}}_{\widehat{m},n}(\beta) - \mathcal{R}_{\widehat{m}}(\beta)|] \leq 36 K\alpha \sqrt{\frac{\pi}{n}}\Big((m^*+1)\sqrt{s_d(m^*)} + 74 e^{-C_3 n^{1-2\rho}} \sum_{m>m^*} \sqrt{s_d(m)} e^{-C_3 s_d(m)}\Big),$$

where the constant C_3 is defined by (6).

Proof: From Corollary 5.25 of [20] and (15), for any $m \in \mathbb{N}$,

$$\mathbb{E}\Big(\sup_{\beta \in B_{m,\alpha}} |\widehat{\mathcal{R}}_{m,n}(\beta) - \mathcal{R}_m(\beta)| \Big) \leq 12 \int_0^\infty \sqrt{\log(N(B_{m,\alpha}, D, \varepsilon))} = 36K\alpha \sqrt{s_d(m)} \sqrt{\frac{\pi}{n}},$$

where $N(B_{m,\alpha}, D, \varepsilon)$ is the ε-covering number of $B_{m,\alpha}$ with respect to the distance D, defined by (14). This gives, for $m = \widehat{m}$,

$$\mathbb{E}\Big(\sup_{\beta \in B_{\widehat{m},\alpha}} |\widehat{\mathcal{R}}_{\widehat{m},n}(\beta) - \mathcal{R}_{\widehat{m}}(\beta)| \Big) \leq 36K\alpha \sqrt{\frac{\pi}{n}} \mathbb{E}\Big(\sqrt{s_d(\widehat{m})} \Big).$$

To compute this expectation, Proposition 6 yields

$$\begin{aligned}
\mathbb{E}\Big(\sqrt{s_d(\widehat{m})} \Big) &= \sum_{m \leq m^*} \sqrt{s_d(m)} \mathbb{P}(\widehat{m} = m) + \sum_{m > m^*} \sqrt{s_d(m)} \mathbb{P}(\widehat{m} = m) \\
&\leq (m^* + 1) \sqrt{s_d(m^*)} + \sum_{m > m^*} \sqrt{s_d(m)} 74 \exp\big(- C_3 (n^{1-2\rho} + s_d(m)) \big) \\
&\leq (m^* + 1) \sqrt{s_d(m^*)} + e^{-C_3 n^{1-2\rho}} \sum_{m > m^*} \sqrt{s_d(m)} 74 \exp\big(- C_3 s_d(m) \big),
\end{aligned}$$

which completes the proof. □

Lemma 5. $\mathbb{E}(\mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) - \mathcal{R}_{m^*}(\beta^*_{m^*})) = 2\alpha^2 e^{K_X} C_1 e^{-C_2 n^{1-2\rho}}$, *where the constants* C_1 *and* C_2 *are defined by* (9) *and* (10).

Proof: Since, for any $m \in \mathbb{N}$, $\langle \beta^*_m, S^m(X) \rangle^2 \leq \|\beta^*_k\|_2^2 \|S^m(X)\|_2^2 \leq \alpha^2 e^{K_X}$, it follows that

$$\begin{aligned}
\mathbb{E}(\mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) - \mathcal{R}_{m^*}(\beta^*_{m^*})) &= \mathbb{E}\Big(\big(Y - \langle \beta^*_{\widehat{m}}, S^{\widehat{m}}(X) \rangle\big)^2 - \big(Y - \langle \beta^*_{m^*}, S^{m^*}(X) \rangle\big)^2 \Big) \\
&= \mathbb{E}\Big(\big(\langle \beta^*_{m^*}, S^{m^*}(X) \rangle + \varepsilon - \langle \beta^*_{\widehat{m}}, S^{\widehat{m}}(X) \rangle\big)^2 - \varepsilon^2 \Big) \\
&= \mathbb{E}\Big(\big(\langle \beta^*_{m^*}, S^{m^*}(X) \rangle - \langle \beta^*_{\widehat{m}}, S^{\widehat{m}}(X) \rangle\big)^2 \Big) \leq 2\alpha^2 e^{K_X} \mathbb{P}(\widehat{m} \neq m^*).
\end{aligned}$$

By Theorem 1, this yields $\mathbb{E}(\mathcal{R}_{\widehat{m}}(\beta^*_{\widehat{m}}) - \mathcal{R}_{m^*}(\beta^*_{m^*})) \leq 2\alpha^2 e^{K_X} C_1 e^{-C_2 n^{1-2\rho}}$, where C_1 and C_2 are defined by (9) and (10). □

Letting $C_5 = 36K\alpha \sqrt{\pi}(m^* + 1) \sqrt{s_d(m^*)}$, and $C_6 = 2664K\alpha \sqrt{\pi} \sum_{m > m^*} \sqrt{s_d(m)} e^{-C_3 s_d(m)} + 2\alpha^2 e^{K_X} C_1$, since, by (25), $C_2 \leq C_3$, we conclude that

$$\mathbb{E}(\langle \widehat{\beta}_{\widehat{m}}, S^{\widehat{m}}(X) \rangle - \langle \beta^*_{m^*}, S^{m^*}(X) \rangle)^2 \leq \frac{C_5}{\sqrt{n}} + C_6 e^{-C_2 n^{1-2\rho}}.$$

Acknowledgments

This work was supported by a grant from Région Ile-de-France. I would like to thank Gérard Biau (Sorbonne Université) and Benoît Cadre (Université Rennes 2) for stimulating discussions and insightful suggestions. I also thank the Editor and two anonymous referees for their careful reading of the paper and constructive comments, which led to a substantial improvement of the article.

References

[1] S. Arlot, Minimal penalties and the slope heuristics: a survey, Journal de la Société Française de Statistique 160 (2019) 1–106.

[2] I. P. Arribas, G. M. Goodwin, J. R. Geddes, T. Lyons, K. E. Saunders, A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder, Translational psychiatry 8 (2018) 1–7.

[3] I. P. Arribas, C. Salvi, L. Szpruch, Sig-SDEs model for quantitative finance, arXiv:2006.00218 (2020).

[4] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, et al., Automatic speech recognition and speech variability: A review, Speech communication 49 (2007) 763–786.

[5] L. Birgé, P. Massart, Minimal penalties for gaussian model selection, Probability Theory and Related Fields 138 (2007) 33–73.

[6] É. Brunel, A. Mas, A. Roche, Non-asymptotic adaptive prediction in functional linear models, Journal of Multivariate Analysis 143 (2016) 208–232.
[7] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Statistics & Probability Letters 45 (1999) 11–22.
[8] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statistica Sinica (2003) 571–591.
[9] K.-T. Chen, Integration of paths—a faithful representation of paths by non-commutative formal power series, Transactions of the American Mathematical Society 89 (1958) 395–407.
[10] I. Chevyrev, A. Kormilitzin, A primer on the signature method in machine learning, arXiv:1603.03788 (2016).
[11] S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical 129 (2008) 750–757.
[12] A. Fermanian, Embedding and learning with signatures, Computational Statistics & Data Analysis 157 (2021) 107148.
[13] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer, New York, 2006.
[14] L. E. Frank, J. H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109–135.
[15] P. K. Friz, N. B. Victoir, Multidimensional Stochastic Processes as Rough Paths: Theory and Applications, volume 120 of Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 2010.
[16] C. Giusti, D. Lee, Iterated integrals and population time series analysis, in: Topological Data Analysis, Springer, 2020, pp. 219–246.
[17] S. Greven, C. Crainiceanu, B. Caffo, D. Reich, Longitudinal functional principal component analysis, in: Recent Advances in Functional Data Analysis and Related Topics, Springer, 2011, pp. 149–154.
[18] P. Hall, J. L. Horowitz, et al., Methodology and convergence rates for functional linear regression, The Annals of Statistics 35 (2007) 70–91.
[19] B. Hambly, T. Lyons, Uniqueness for the signature of a path of bounded variation and the reduced path group, The Annals of Mathematics 171 (2010) 109–167.
[20] R. van Handel, Probability in high dimension, Technical Report, Princeton University, 2014.
[21] T. Hastie, C. Mallows, [a statistical view of some chemometrics regression tools]: Discussion, Technometrics 35 (1993) 140–143.
[22] F. J. Király, H. Oberhauser, Kernels for sequentially ordered data, Journal of Machine Learning Research 20 (2019) 1–45.
[23] S. Lai, L. Jin, W. Yang, Online signature verification using recurrent neural network and length-normalized path signature descriptor, in: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 1, IEEE, pp. 400–405.
[24] Y. Le Jan, Z. Qian, Stratonovich's signatures of brownian motion determine brownian sample paths, Probability Theory and Related Fields 157 (2013) 209–223.
[25] D. Levin, T. Lyons, H. Ni, Learning from the past, predicting the statistics for the future, learning an evolving system, arXiv:1309.0260 (2013).
[26] C. Li, X. Zhang, L. Jin, LPSNet: a novel log path signature feature based hand gesture recognition framework, in: 2017 IEEE International Conference on Computer Vision Workshop, pp. 631–639.
[27] Y. Li, T. Hsing, On rates of convergence in functional linear regression, Journal of Multivariate Analysis 98 (2007) 1782–1804.
[28] M. Liu, L. Jin, Z. Xie, Ps-lstm: Capturing essential sequential online information with path signature and lstm for writer identification, in: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 1, IEEE, pp. 664–669.
[29] T. Lyons, Rough paths, signatures and the modelling of functions on streams, arXiv:1405.4537 (2014).
[30] T. Lyons, M. Caruana, T. Lévy, Differential Equations driven by Rough Paths, volume 1908 of Lecture Notes in Mathematics, Springer, Berlin, 2007.
[31] B. D. Marx, P. H. Eilers, Generalized linear regression on sampled signals and curves: a P-spline approach, Technometrics 41 (1999) 1–13.
[32] P. Moore, T. Lyons, J. Gallacher, Using path signatures to predict a diagnosis of Alzheimer's disease, PloS ONE 14 (2019).
[33] J. Morrill, A. Fermanian, P. Kidger, T. Lyons, A generalised signature method for multivariate time series feature extraction, arXiv:2006.00873 (2020).
[34] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, T. Lyons, The signature-based model for early detection of sepsis from electronic health records in the intensive care unit, International Conference in Computing in Cardiology (2019).
[35] J. H. Morrill, A. Kormilitzin, A. J. Nevado-Holgado, S. Swaminathan, S. D. Howison, T. J. Lyons, Utilization of the signature method to identify the early onset of sepsis from multivariate physiological time series in critical care monitoring, Critical Care Medicine 48 (2020) e976–e981.
[36] J. S. Morris, Functional regression, Annual Review of Statistics and Its Application 2 (2015) 321–359.
[37] S. Y. Park, A.-M. Staicu, Longitudinal functional data analysis, Stat 4 (2015) 212–226.
[38] C. Ramos-Carreño, J. L. Torrecilla, A. Suárez, Scikit-fda: A python package for functional data analysis, in: 3rd International Workshop on Advances in Functional Data Analysis, volume 5.
[39] J. O. Ramsay, C. Dalzell, Some tools for functional data analysis, Journal of the Royal Statistical Society. Series B (Methodological) 53 (1991) 539–561.
[40] J. O. Ramsay, B. W. Silverman, Functional Data Analysis. 2nd Edition., Springer, New York, 2005.
[41] J. Reizenstein, B. Graham, Algorithm 1004: The iisignature library: Efficient calculation of iterated-integral signatures and log signatures, ACM Transactions on Mathematical Software (2020).
[42] P. Turaga, R. Chellappa, V. S. Subrahmanian, O. Udrea, Machine recognition of human activities: A survey, IEEE Transactions on Circuits and Systems for Video technology 18 (2008) 1473–1488.
[43] B. Wang, M. Liakata, H. Ni, T. Lyons, A. J. Nevado-Holgado, K. Saunders, A path signature approach for speech emotion recognition, in: Interspeech 2019, pp. 1661–1665.
[44] W. Yang, L. Jin, M. Liu, Chinese character-level writer identification using path signature feature, dropstroke and deep cnn, in: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp. 546–550.
[45] W. Yang, L. Jin, M. Liu, DeepWriterID: An end-to-end online text-independent writer identification system, IEEE Intelligent Systems 31 (2016) 45–53.
[46] W. Yang, T. Lyons, H. Ni, C. Schmid, L. Jin, J. Chang, Developing the path signature methodology and its application to landmark-based human action recognition, arXiv:1707.03993 (2017).