ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Signature and statistical learning

Master Thesis

Adeline Fermanian

Friday 26th October, 2018

Advisors: Prof. Peter Bühlmann, Prof. Gérard Biau, Prof. Benoît Cadre
Department of Mathematics, Seminar for Statistics, ETH Zürich

Abstract

In this thesis, we introduce the notion of signature of a path as feature set in a statistical learning framework. The signature dates back from the 60s when Chen noticed in [4] that a path can be represented by its iterated integrals and it has been at the centre of Lyons' rough paths theory in the 90s. The signature encodes geometric properties of a multidimensional path: it can be viewed as a non-parametric dimension reduction technique. Moreover, a lot of real-world data can be represented as a path evolving with time, think for example of handwriting recognition from character trajectories, market prediction from financial time series, analysis of medical sensors... The signature transformation combined with a learning algorithm has achieved state of the art results for several of these applications, see, e.g., [34]. This justifies the need of statistical investigation of the signature properties. In the following, we review the theory of signature and we investigate its applications in statistical learning. We look more closely at the signature transformation in a regression framework, as presented in [21], and derive a convergence rate. In light of promising results in the literature, we undertake some empirical tests on the signature transformation and its performance compared to other classical algorithms.

Acknowledgements

I would first like to thank my supervisors, Gérard and Benoît, who have offered me the opportunity to work on this thesis. They have always been available to help me when I needed to and, by many thoughtful gestures, have made these last months a really nice time. I am also grateful for the opportunity to continue working on this topic as a Phd student and I am looking forward to the three years to come! In this regard, I would like to thank the DIM Math Innov for awarding me a Phd grant. Then, I would like to thank Lorenzo Zambotti for his help in this work. I am also very grateful to all the Phd students and professors at LPSM for welcoming me in the lab and having many nice chats over a coffee.

Finally, I would like to thank Peter Bühlmann for accepting to be my supervisor at ETH, as well as all the team from the Seminar for Statistics. I am grateful for the dedication of the lecturers, who have aroused my interest in Statistics during my master. I am also thankful to Markus Kalish and Fadoua Balabdaoui who have taken some time to discuss my educational choices.

Contents

List of Figures

Chapter 1

Introduction

1.1 The problem

Given the ubiquity of time series data and the increase in storage capacity, there has been a lot of interest in statistical methods for functional data. To name only a few, market evolution in finance, electrocardiograms or gait evolution in medicine, meteorological records, geo-tracking of people or cars, are all functional data. One may want to classify or cluster them into different categories (for example discriminate between a normal or diseased medical record) or to predict future values from the ones observed until now (weather or financial stocks prediction). For this, traditional methods, which deal with a finite number of features, need to be extended. The main difficulty lays in the high dimensional nature of these data. A lot of focus has thus lied in finding good representations or good similarity measures between two data streams. Note that the problem of dealing with functional object appeared in different scientific communities who thus have a different vocabulary. We will talk indistinguishably of functional data, data stream or time series. What we mean by these terms is that we want to learn from a function $X : [0, 1] \to \mathbb{R}^d$ an output Y which can be either numerical or discrete.

To reduce the dimension of data streams, traditional methods often require strong assumptions on their functional nature or on the underlying probabilistic models. We want to investigate here what is called the signature method, which requires no strong assumption on the functional data and can be used in a fully non-parametric statistical model. This method treats data streams as paths and represents them on a small number of coefficients who encapsulate their geometric properties.

More precisely, the signature of a path is an infinite series consisting of its iterated integrals. Iterated integrals have been introduced in the seminal work [4] of Chen in the middle of the 60s. The notion has been rediscovered in the 90s with Lyons' theory of rough paths. It has recently received the attention of the machine learning community because of a series of successful applications. To cite some of them, [34] have achieved state of the art results for handwriting recognition with a recurrent neural network combined with signature features up to order 3, [12] has used the same approach for characters recognition, [22] have used Lasso regression with signature features up to order 4 for classification of financial data streams, [17] for detection of bipolar disorders and [35] for human action recognition. For a more complete review of recent applications, we refer the reader to [6]. To our current knowledge, no systematic comparison of predictive power of the signature transformation has been undertaken. It has also led to some statistical articles, [24] in statistical inference for stochastic differential equation

and [16] who define a "sequential kernel" based on the signature.

Let us now introduce the signature as a natural object when solving controlled differential equations (see [11] for a systematic treatment of this topic).

1.2 The origin of signature

The signature of a path naturally appears when looking at Picard's iterations for controlled ordinary differential equations of the form

$$dY_t = g(Y_t)dX_t, \quad Y_0 = y_0, \tag{1.1}$$

with $X : [0,1] \to E := \mathbb{R}^d$, $Y : [0,1] \to F := \mathbb{R}^e$, and $g : F \to \mathcal{L}(E,F)$, $\mathcal{L}(E,F)$ being the set of linear functionals from E to F (that one can identify with the set of $e \times d$ real matrices). That Y solves equation (1.1) means that for all $t \in [0,1]$,

$$Y_t = y_0 + \int_0^t g(Y_s)dX_s.$$

One can obtain solutions by Picard iterations. Let, for any path $Y : [0,1] \to \mathbb{R}^d$,

$$F(Y)_t = y_0 + \int_0^t g(Y_s)dX_s.$$

Then, a solution of (1.1) is a fixed point of F and one can approach it by defining a sequence of paths

$$Y^{k+1} = F(Y^k),$$

with an arbitrary initial path Y^0. Under suitable assumptions one can prove that Y^k converges to Y.

Let us look at a particular case when g is linear. In this case, we can rewrite $g(Y_s)dX_s$ as $h(dX_s)Y_s$ with $h : \mathbb{R}^d \to \mathbb{R}^{e\times e}$ linear. We begin with the constant path $Y_t^0 = y_0$ and we define iteratively

$$Y_t^1 = y_0 + \int_0^t h(dX_s)Y_s^0 = \left(1 + \int_0^t h(dX_s)\right) y_0,$$

$$Y_t^2 = y_0 + \int_0^t h(dX_s)Y_s^1 = \left(1 + \int_0^t h(dX_s) + \int_0^t \int_0^s h(dX_u)h(dX_s)\right) y_0.$$

Letting $h^{\otimes k}(e_1 \otimes \cdots \otimes e_k) = h(e_1)\cdots h(e_k)$, we have, by linearity,

$$Y_t^2 = \left(1 + h\left(\int_0^t dX_s\right) + h^{\otimes 2}\left(\int_0^t \int_0^s dX_u \otimes dX_s\right)\right) y_0,$$

and iterating we are left to

$$Y_t^k = \Bigg(1 + h(\int_0^t dX_s) + h^{\otimes 2}(\int_0^t \int_0^s dX_u \otimes dX_s) + \cdots + h^{\otimes k}\left(\int_{0<u_1<u_2<\cdots<u_k<1} \cdots \int dX_{u_1} \otimes \cdots \otimes dX_{u_k}\right)\Bigg) y_0.$$

The expression $\int_{0<u_1<u_2<\cdots<u_k<1} \cdots \int dX_{u_1} \otimes \cdots \otimes dX_{u_k}$ that appears is exactly the signature up to order k of X. We see that it completely determines Y^k. Taking the limit, the infinite signature of X determines the solution Y. It has actually be shown that this also holds for non-linear maps (see [4] and [14] for g Lipschitz).

1.3 Overview

The last section has shown that the signature characterizes solutions of differential equations driven by a certain path and is thus an interesting function of the path. We will investigate this more deeply in this thesis. Chapter 2 provides a detailed exposition of the signature and its properties. Chapter 3 reviews some of the standard statistical methods dealing with functional data and presents our model of interest. Finally, we present some experimental results in chapter 4, which provide evidence for the potential of the signature method in practical applications.

Chapter 2

Signature : theoretical foundations

2.1 Preliminaries

Before defining formally the signature, we need to introduce some notions about bounded variation paths and tensor spaces. In all the following, E is a Banach space of dimension d equipped with a norm $\|.\|$, usually identified with $\mathbb{R}^d$.

2.1.1 Paths of bounded variation

Definition 2.1.1 (Path of bounded variation)**.** Let $X : [0,1] \to E$ be a continuous path. The p-variation of X for $p \geq 1$ is defined by

$$\|X\|_{p-var} = \left(\sup_D \sum_{t_i \in D} \|X_{t_i} - X_{t_{i-1}}\|^p \right)^{1/p},$$

where the supremum is taken over all finite partitions

$$D = \{(t_0, \dots, t_k) | k \geq 0, 0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1\}$$

of $[0,1]$. X is said to be of finite p-variation if its p-variation is finite.

We denote $BV^p(E)$ the set of continuous paths of finite p-variation with values in E.

If $p = 1$, we have

$$\|X\|_{1-var} = \sup_D \sum_{t_i \in D} \|X_{t_i} - X_{t_{i-1}}\| = \lim_{|D| \to \infty} \sum_{t_i \in D} \|X_{t_i} - X_{t_{i-1}}\|,$$

with $|D| = k$ the length of the partition $D = \{0 = t_0 < \cdots < t_k = 1\}$. Moreover, if X is continuously differentiable, we have

$$\|X\|_{1-var} = \int_0^1 \|X'(t)\| dt.$$

If $\|X\|_{1-var} < \infty$ we will also say that X is of bounded variation. Note that $\|.\|_{p-var}$ is a semi-norm (the p-variation of a constant path is null) and that we can define a norm on $BV^p(E)$ by letting

$$\|X\|_{BV^p(E)} = \|X\|_{p-var} + \sup_{t \in [0,1]} \|X_t\|.$$

The norm $\|X\|_{BV^p(E)}$ is called the p-variation norm of X and $\|X\|_{p-var}$ the p-variation of X. It equips $BV^p(E)$ with a Banach structure, see [23].

Proposition 2.1.1. *$(BV^p(E), \|.\|_{BV^p(E)})$ is a Banach space.*

Definition 2.1.2 (Riemann-Stieljes integral)**.** Let X and Y be two functions from $[a,b] \to \mathbb{R}^d$. Let $D^n = \{a = t_i^n < \cdots < t_k^n = b\}$ a partition of $[a,b]$ such that its mesh $\|D^n\| = \sup_{1\leq i\leq k} |t_i^n - t_{i-1}^n|$ tends to 0 when $n \to \infty$ and (s_i^n) a sequence such that for all i and n, $s_i^n \in [t_{i-1}^n, t_i^n]$. Then, if the sum

$$\sum_{i=1}^{n} Y_{s_i}(X_{t_i} - X_{t_{i-1}})$$

converges to a limit I independent of the choice of (s_i^n) and D^n, we say that the Riemann-Stieljes integral of Y against X exists, is equal to I and denote it $\int_a^b Y_t dX_t = \int_a^b Y d$.

It can be shown that if Y is continuous and X of bounded variation, then the Riemann-Stieljes integral $\int Y dX$ exists (see [33] for a proof). Young also proved in [36] that the Riemann-Stieljes integral is well-defined if X has finite p-variation and Y has finite q-variation with $\frac{1}{p} + \frac{1}{q} > 1$.

2.1.2 Tensor space

We now introduce some definitions and notations about the tensor algebra, which is the space of the signature.

Definition 2.1.3 (Tensor product of vector spaces)**.** Let E and F be two vector spaces over a field K. A tensor product of E and F, denoted by $E \otimes F$, is a vector space over the same field K with a bilinear map $\varphi : E \times F \to E \otimes F$ such that for any basis $e = (e_i)_{i\in I}$ of E and $f = (f_j)_{j\in J}$ of F, then

$$\varphi(e \times f) = \{\varphi(e_i, f_j) | e_i \in e, f_j \in f\}$$

is a basis of $E \otimes F$. For any $x \in E$ and $y \in F$, $\varphi(x,y)$ is denoted $x \otimes y$ and called the tensor product of x and y.

One can prove that such a product exists and is unique up to isomorphisms (see appendix A and [26] for more details).

The nth tensor power of a vector space E is defined as the order n tensor product of E with itself:

$$E^{\otimes n} = \underbrace{E \otimes \cdots \otimes E}_{n}.$$

It is useful to identify $E^{\otimes n}$ with the space of homogeneous non-commuting polynomials of degree n. Indeed, let $(e_1, \ldots, e_d)$ be a basis of E, any element of $E^{\otimes n}$ can be written as a sum $\sum_{I=(i_1,\cdots,i_n)\subset\{1,\cdots,d\}^n} \alpha_I e_{i_1} \otimes \cdots \otimes e_{i_n}$, which can be thought of as $\sum \alpha_I X_{i_1} \ldots X_{i_n}$ where $X_1, \ldots, X_n$ are non commuting indeterminates.

Note that by construction $\dim(E \otimes F) = \dim(E) \times \dim(F)$ so that if $E = \mathbb{R}^d$, $E^{\otimes n}$ is of dimension d^n. Then, one can also identify $E^{\otimes n}$ with $\mathbb{R}^{d^n}$, which means that, for example, $E^{\otimes 2}$ can be identified with the space of $d \times d$ matrices.

Definition 2.1.4. We denote by $T((E))$ the space of formal series of tensors of E, i.e.,

$$T((E)) = \{(a_0, \ldots, a_n, \ldots) \mid \forall n \geq 0 \quad a_n \in E^{\otimes n}\},$$

and

$$T^N(E) := \oplus_{k=0}^{N} E^{\otimes k}$$

the truncated tensor space up to order N.

Definition 2.1.5. We endow $T((E))$ with the following operations : for $a, b \in T((E))$, $\lambda \in \mathbb{R}$,

$$a + b = (a_0 + b_0, a_1 + b_1, \ldots, a_n + b_n, \ldots)$$
$$\lambda.a = (\lambda a_0, \lambda a_1, \ldots, \lambda a_n, \ldots)$$
$$a \otimes b = (c_0, c_1, \ldots, c_n, \ldots)$$

with

$$c_n = \sum_{k=0}^{n} a_k \otimes b_{n-k}$$

We also define the projection operator of an infinite tensor object onto its first levels.

Definition 2.1.6. The canonical projection π_N of an element of $T((E))$ on the truncated tensor space $T^N(E)$ is defined by

$$\begin{array}{rccc} \pi_N : & T((E)) & \longrightarrow & T^N(E) \\ & (a_0, \cdots, a_N, a_{N+1}, \cdots) & \longmapsto & (a_0, \cdots, a_N). \end{array}$$

The following proposition is clear from the definitions.

Proposition 2.1.2. *$(T((E)), +, ., \otimes)$ is a real non-commutative algebra with neutral element $\mathbf{1} := (1, 0, \ldots, 0, \ldots)$.*

An element of $T((E))$ will be invertible if and only if $a_0 \neq 0$. Thus, the space $\tilde{T}((E)) = \{a \in T((E)) | \pi^0(a) = 1\}$ is a group.

Proposition 2.1.3. *$(\tilde{T}((E)), \otimes)$ is a Lie group, and for all $a \in \tilde{T}((E))$,*

$$a^{-1} = \sum_{k \geq 0} (\mathbf{1} - a)^{\otimes k}.$$

See appendix B for the definition of a Lie group.

Proof. Let $N > 0$ and let us look at the projection at order N of $a \otimes \sum_{k \geq 0} (\mathbf{1} - a)^{\otimes k}$. We have

$$\begin{aligned} \pi_N \left(a \otimes \sum_{k \geq 0} (\mathbf{1} - a)^{\otimes k} \right) &= \pi_N \left(a \otimes \sum_{k=0}^{N} (\mathbf{1} - a)^{\otimes k} + a \otimes \sum_{k \geq N+1} (\mathbf{1} - a)^{\otimes k} \right) \\ &= \pi_N \left(a \otimes \sum_{k=0}^{N} (\mathbf{1} - a)^{\otimes k} \right) \end{aligned}$$

because the second term contains only elements of order at least $N + 1$ (as $\pi_0(\mathbf{1} - a) = 0$). Moreover,

$$\begin{aligned} a \otimes \sum_{k=0}^{N} (\mathbf{1} - a)^{\otimes k} &= (\mathbf{1} - (\mathbf{1} - a)) \otimes \sum_{k=0}^{N} (\mathbf{1} - a)^{\otimes k} \\ &= \sum_{k=0}^{N} (\mathbf{1} - a)^{\otimes k} - \sum_{k=0}^{N} (\mathbf{1} - a)^{\otimes k+1} \\ &= \mathbf{1} - (\mathbf{1} - a)^{\otimes N+1}. \end{aligned}$$

Therefore,

$$\pi_N(a \otimes \sum_{k=0}^{N} (\mathbf{1} - a)^{\otimes k}) = \mathbf{1}.$$

This is true for any N and the same calculations can be done for $\sum_{k\geq 0}(\mathbf{1}-a)^{\otimes k} \otimes a$. So

$$\sum_{k\geq 0}(\mathbf{1}-a)^{\otimes k} \otimes a = a \otimes \sum_{k\geq 0}(\mathbf{1}-a)^{\otimes k} = \mathbf{1}$$

$\tilde{T}((E))$ is an affine subspace of $T((E))$ so it is a smooth manifold. The operations $\otimes$ and $^{-1}$ are smooth maps (they are polynomials in the coordinates). We can conclude that $\tilde{T}((E))$ is a Lie group. □

As in [16] and [11], we endow the tensor space with the Euclidean scalar product and associated norm.

Definition 2.1.7. Let $a, b \in E^{\otimes k}$, $(e_i)_{i=1}^d$ an orthonormal basis of E. Then, if $a = \sum_{I\subset\{1,\cdots,d\}^k} a_I e_{i_1} \otimes \cdots \otimes e_{i_k}$, $b = \sum_{I\subset\{1,\cdots,d\}^k} b_I e_{i_1} \otimes \cdots \otimes e_{i_k}$, we have

$$\langle a, b\rangle_{E^{\otimes k}} = \sum_{i_1,\cdots,i_k\in\{1,\cdots,d\}} a_{i_1\cdots i_k} b_{i_1\cdots i_k}$$

and

$$\|a\|_{E^{\otimes k}} = \sqrt{\sum_{i_1,\cdots,i_k\in\{1,\cdots,d\}} a_{i_1\cdots i_k}^2}.$$

Note that the norm satisfies the following property: for $n \geq 0$, $0 \leq k \leq n$, $a \in E^{\otimes k}$ and $b \in E^{\otimes n-k}$

$$\|a \otimes b\|_{E^{\otimes n}} = \|a\|_{E^{\otimes k}} \|b\|_{E^{\otimes n-k}}$$

It is thus and admissible norm (see Appendix A.2 for definition).

Definition 2.1.8. For $v, w \in T((E))$ we define

$$\langle v, w\rangle_{T((E))} = \sum_{k\geq 0} \langle v_k, w_k\rangle_{E^{\otimes k}}$$

Note that this "scalar product" in $T((E))$ may be infinite.

2.2 Definition and first examples

Now that we have the necessary definitions on the tensor space, we can define the signature as an element of $T((E))$ and provide some elementary examples.

2.2.1 Definition and notation

Definition 2.2.1. Let $X : [0,1] \to E$ be a path of bounded variation. The signature of X is defined as

$$S(X) = (1, \mathbf{X^1}, \mathbf{X^2}, \cdots, \mathbf{X^n}, \cdots) \in T((E)),$$

where, for each integer n,

$$\mathbf{X^n} = \int\cdots\int_{0<u_1<u_2<\cdots<u_n<1} dX_{u_1} \otimes \cdots \otimes dX_{u_n} \in E^{\otimes n}.$$

The integrals can be understood as Riemann-Stieljes integrals.

Example 2.2.1. If $d = 2$, $X_t = (X_t^1, X_t^2)$, $E^{\otimes n}$ can be identified with $\mathbb{R}^{2^n}$ and one can compute the first two orders of the signature:

$$\mathbf{X^1} = \int_0^1 dX_t = \begin{pmatrix} \int_0^1 dX_t^1 \\ \int_0^1 dX_t^2 \end{pmatrix}$$

$$\mathbf{X^2} = \int_0^1 \int_0^t dX_s \otimes dX_t = \begin{pmatrix} \int_0^1 \int_0^t dX_s^1 dX_t^1 & \int_0^1 \int_0^t dX_s^1 dX_t^2 \\ \int_0^1 \int_0^t dX_s^2 dX_t^1 & \int_0^1 \int_0^t dX_s^2 dX_t^2 \end{pmatrix}.$$

Notation. For every integer $N \geq 1$, the truncated signature of order N is

$$S^N(X) = (1, \mathbf{X^1}, \mathbf{X^2}, \cdots, \mathbf{X^N}) = \pi_N(S(X))$$

Remark. If X is of bounded p-variation with $1 < p < 2$, then the signature is still well-defined if the integrals are understood as Young integrals (see for instance [11]). If $p > 2$, the iterated integrals are no longer uniquely defined. The theory of rough paths generalize the Young integral in these cases. Thus, there exists a notion of signature for highly oscillatory signals, which makes the signature transformation a very general tool.

Notation. We introduce the following notation:

$$\Delta_k = \{(u_1, \cdots, u_k) \in [0,1]^k | 0 < u_1 < \cdots < u_k < 1\}.$$

For a multi-index $I = (i_1, \cdots, i_k) \in \{1, \cdots, d\}^k$, its length is denoted by $|I| = k$ and the coefficient of $\mathbf{X^k}$ corresponding to this multi-index is

$$S^I(X) = \int \cdots \int_{(u_1,\cdots,u_k)\in\Delta_k} dX_{u_1}^{i_1} \ldots dX_{u_N}^{i_N} = \int_{u\in\Delta_k} dX_u^I.$$

We will sometimes consider the signature of a path restricted to a specific interval $X : [s,t] \to E$ and then denote its signature by $S(X)_{[s,t]}$.

The signature is an element of $T((E))$ and the truncated signature $S^N(X)$ is in $T^N(E)$. If $(e_i)_{i=1}^d$ is a basis of E, $(e_{i_1} \otimes \cdots \otimes e_{i_N})_{(i_1,\cdots,i_N)\in\{1,\cdots,d\}^N}$ is a basis of $E^{\otimes N}$. Then for $I = (i_1, \cdots, i_N)$ we can also write the signature as

$$S(X) = 1 + \sum_{N=1}^{\infty} \sum_{|I|=N} S^I(X) e_{i_1} \otimes \cdots \otimes e_{i_N}.$$

2.2.2 Geometric interpretation

We try to give here a geometric interpretation of the first orders of the signature. The first order terms are just the increments of the path: for any $s, t \in [0,1]$,

$$\mathbf{X}_{[s,t]}^1 = X_t - X_s$$

The second order terms are already a little more interesting. Indeed, by definition, the Levy area of a curve (x_t, y_t) in the time interval $[0,1]$ is

$$\mathcal{A} = \frac{1}{2}\left(\int_0^1 x_s dy_s - \int_0^1 y_s dx_s\right).$$

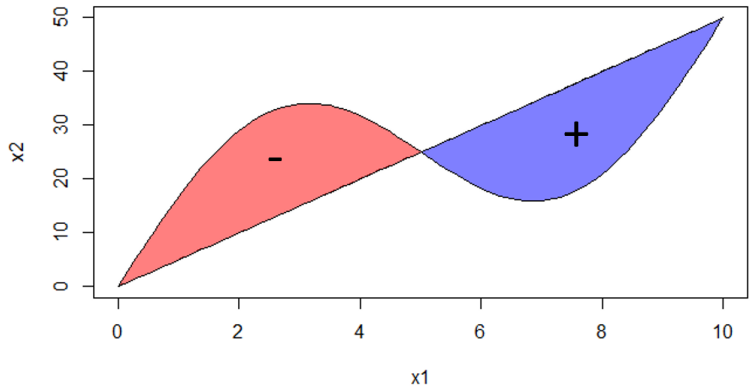


Figure 2.1: Levy area

It is a signed are between the curve and the chord connecting the two endpoints (see figure 2.1). We recognize the coefficients of the signature: if $\mathcal{A}_{i,j}$ is the Levy area of the curve (X_t^i, X_t^j), then

$$\mathcal{A}_{i,j} = \frac{1}{2}(S^{i,j}(X) - S^{j,i}(X).$$

We see that the signature encodes geometric properties of the paths formed by the different pairs of coordinates of X.

2.2.3 Examples

We can now give some examples of paths for which we can directly compute the signature.

Example 2.2.2 (Linear path)**.** If $X : [0,1] \to \mathbb{R}^d$ is a linear path, i.e., $X_t = X_0 + (X_1 - X_0)t$ for $t \in [0,1]$, then for any $I = (i_1, \cdots, i_k) \in \{1, \cdots, d\}^k$,

$$\begin{aligned} S^I(X) &= \int \cdots \int_{0<u_1<u_2<\cdots<u_k<1} dX_{u_1}^{i_1} \dots dX_{u_k}^{i_k} \\ &= \int \cdots \int_{0<u_1<u_2<\cdots<u_k<1} (X_1 - X_0)^{i_1} \cdots (X_1 - X_0)^{i_k} du_1 \cdots du_k \\ &= \prod_{j=1}^{k} (X_1 - X_0)^{i_j} \int \cdots \int_{0<u_1<u_2<\cdots<u_k<1} du_1 \cdots du_k \\ &= \frac{1}{k!} \prod_{j=1}^{k} (X_1 - X_0)^{i_j}. \end{aligned}$$

Example 2.2.3 (Path in one dimension)**.** In one dimension, the signature is directly related to the moments of X. Indeed, let $X : [0,1] \to \mathbb{R}$ be a one-dimensional path. Then, for any $k \geq 0$, $\mathbf{X^k} \in \mathbb{R}$, and

$$\mathbf{X^k} = \int \cdots \int_{0<u_1<u_2<\cdots<u_k<1} dX_{u_1} \dots dX_{u_N} = \frac{1}{k!}(X_1 - X_0)^k.$$

So, if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and X is a time-continuous stochastic process such that X is of bounded variation almost surely, the expected signature is

$$\mathbb{E}[\mathbf{X}^{\mathbf{k}}_{[0,t]}] = \frac{1}{k!}\mathbb{E}[(X_t - X_0)^k].$$

Therefore, in one dimension, the signature only depends on its increment $X_1 - X_0$. We will see in the next section that more interesting geometric interpretations appear for multidimensional paths. Following [16], the signature coefficients can be thought of as ordered moments of the paths.

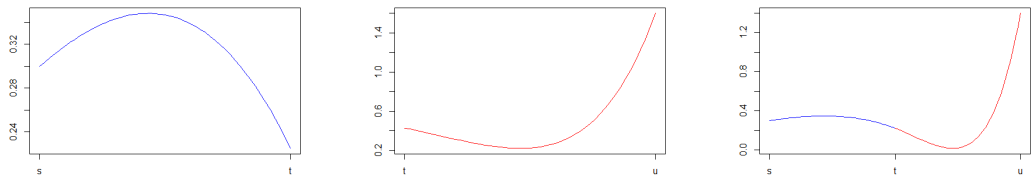
2.3 Properties of the signature

2.3.1 First properties

The signature has a number of interesting properties, which will be useful for computations and statistical inference. We present here some of them with proofs, adapted from [23]. The first property gives a method to compute the signature of a concatenation of paths.

Definition 2.3.1. Let $X : [s,t] \to E$ and $Y : [t,u] \to E$. We let the concatenation of paths (see Figure 2.2) be defined as

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



X Y $X * Y$

Figure 2.2: Concatenation of paths

Theorem 2.3.1 (Chen's identity). *Let $X : [s,t] \to E$ and $Y : [t,u] \to E$ be two continuous paths with bounded variation. Then*

$$S(X * Y)_{[s,u]} = S(X)_{[s,t]} \otimes S(Y)_{[t,u]}.$$

Proof. Let $Z = X * Y : [s,u] \to E$ and $S(Z)$ its signature. We look at the order n term of

its signature:

$$\begin{aligned}\mathbf{Z^n} &= \int\cdots\int_{s<u_1<u_2<\cdots<u_n<u} dZ_{u_1}\otimes\cdots\otimes dZ_{u_n}\\ &= \sum_{k=0}^{n}\int\cdots\int_{s<u_1<\cdots<u_k<t<u_{k+1}<\cdots<u_n<u} dZ_{u_1}\otimes\cdots\otimes dZ_{u_n}\\ &= \sum_{k=0}^{n}\int\cdots\int_{s<u_1<\cdots<u_k<t} dX_{u_1}\otimes\cdots\otimes dX_{u_k}\otimes\int\cdots\int_{t<u_{k+1}<\cdots<u_n<u} dY_{u_{k+1}}\otimes\cdots\otimes dY_{u_n}\\ &= \sum_{k=0}^{n} X^k\otimes Y^{n-k},\end{aligned}$$

where we used Fubini's theorem. Therefore,

$$S(Z) = S(X)\otimes S(Y)$$

□

For paths of bounded p-variation with $p\in(1,2)$, one cannot use Fubini's theorem but one can prove that Chen's identity is still valid (see [23]). An equivalent formulation of Chen's theorem is that the signature is a homomorphism from the monoid of paths $(BV(E),*)$ to the group $(\tilde{T}((E)),\otimes)$. We can now apply this formula to the signature of a piecewise linear path.

Example 2.3.1. Let $X:[0,1]\to\mathbb{R}^d$ be a piecewise linear path. Let us assume that the knots are $0=t_0<t_1<\cdots<t_k=1$. Then, by Chen's identity,

$$S(X) = S(X)_{[t_0,t_1]}\otimes\cdots\otimes S(X)_{[t_{k-1},t_k]}.$$

On each $[t_{j-1},t_j]$, X is a linear path so $S(X)_{[t_{j-1},t_j]}$ can be obtained from example 2.2.2.

The next lemma show that the signature is independent of time parametrisation. It is an immediate consequence from the definition.

Lemma 2.3.1 (Invariance under time reparametrisation)**.** *Let $X:[a,b]\to E$ be a path, $\psi:[a,b]\to[a,b]$ be a reparametrisation (non decreasing surjection), and let $\tilde{X}_t = X_{\psi(t)}$ for $t\in[a,b]$. Then, for every $s,t\in[a,b]$,*

$$S(\tilde{X})_{[s,t]} = S(X)_{[\psi(s),\psi(t)]}.$$

Proof. This is the change of variable formula for Riemann-Stieljes integrals. □

This also means that when we will use the signature as a feature for a learning task, one will lose all information about the speed of the path. As we have seen in section 1.2, the signature characterises solutions of a certain class of differential equations. Conversely, the next lemma shows that the signature itself is a solution of a differential equation. This result will prove extremely useful to derive other properties of the signature.

Lemma 2.3.2. *Let $X:[0,1]\to E$ of bounded variation, $Y:[0,1]\to T^N(E)$ Then the unique solution of the controlled differential equation*

$$dY_t = \pi_N(Y_t\otimes dX_t),\quad Y_0 = (1,0,\cdots,0)$$

is the signature of X truncated at order N i.e. the path from $[0,1]\to T^N(E)$ defined by $t\to S^N(X)_{[0,t]}$.

Proof. Existence and uniqueness are a consequence of Theorem C.1.1, which is a version of Picard's theorem. We can check that the truncated signature is indeed a solution. We have for the order k of the signature on $[0,t]$

$$\begin{aligned}\mathbf{X^k}_{[0,t]} &= \int\cdots\int_{0<u_1<\cdots<u_k<t} dX_{u_1}\otimes\cdots\otimes dX_{u_k}\\ &= \int_0^t \left(\int\cdots\int_{0<u_1<\cdots<u_{k-1}<u_k} dX_{u_1}\otimes\cdots\otimes dX_{u_{k-1}}\right)\otimes dX_{u_k}\\ &= \int_0^t \mathbf{X^{k-1}}_{[0,u]}\otimes dX_u.\end{aligned}$$

Thus,

$$S^k(X)_{[0,t]} = (1,\mathbf{X^1},\cdots,\mathbf{X^k}) = \mathbf{1} + \int_0^t \pi_k(S^{k-1}(X)_{[0,u]}\otimes dX_u).$$

□

Corollary 2.3.1. *The truncated signature map $\pi_N\circ S : BV(E)\to T^N(E)$ is continuous for any $N\geq 0$*

Proof. This is an immediate consequence of the second part of Picard's Theorem C.1.1. □

For a path X of bounded p-variation, we know that its signature exists and that its first term is 1. Then, it is invertible in $\tilde{T}((E))$. One can prove that its inverse is the signature of the path obtained by running X backward in time.

Proposition 2.3.1 (Time-reversal)**.** *Let $X:[0,1]\to E$ a path with bounded p-variation with $p<2$ and let $\overleftarrow{X}_t$ its time-reversal : $\overleftarrow{X}_t = X_{1-t}$ for $t\in[0,1]$. Then,*

$$S(X)\otimes S(\overleftarrow{X}) = \mathbf{1}.$$

Proof. One can use lemma 2.3.2. Let $Z = X * \overleftarrow{X}$. Then, for any Banach space V and any map $f: V\to\mathcal{L}(E,V)$, it is equivalent for a path $Y:[0,1]\to V$ to be solution of

$$dY_t = f(Y_t)dX_t,\quad Y_0=\xi,\quad Y_1=\eta$$

or of

$$dY_t = f(Y_t)d\overleftarrow{X}_t,\quad Y_0=\eta,\quad Y_1=\xi.$$

So any solution of

$$dY_t = f(Y_t)dZ_t,\quad Y_0=\xi$$

will satisfy $Y_2=\xi$ for any function f. If we take as f the function of Lemma 2.3.2, $\mathbf{1}$ is a solution but $S^N(Z)_{[0,t]}$ is also a solution. By uniqueness, for any $N\geq 0$

$$S^N(Z)_{[0,t]} = \mathbf{1}$$

and thus

$$S(Z) = S(X)\otimes S(\overleftarrow{X}) = \mathbf{1}$$

□

Theorem 2.3.2 provides a criterion for uniqueness of the signature. Before stating it, we need to investigate what it means for two paths to have the same signature. It is clear that for two paths X and Y, $S(X) = S(Y)$ does not imply that $X = Y$. Indeed, Lemma 2.3.1 already tells us that if Y is a reparametrisation of X it will have the same signature. Moreover, according to Proposition 2.3.1, $S(X \otimes \overleftarrow{X}) = \mathbf{1}$ so the path $X \otimes \overleftarrow{X}$ has the same signature as any constant path. The situation is actually even more complicated. Indeed if X, Y and Z are non-constant paths, then $S(X \otimes Y \otimes \overleftarrow{Y} \otimes Z \otimes \overleftarrow{Z} \otimes \overleftarrow{X}) = \mathbf{1}$ even if this path cannot be rewritten as $W \otimes \overleftarrow{W}$. The notion of tree-like paths introduced in [14] formalizes this idea of a path being "reducible" to a constant path by certain operations

Definition 2.3.2 (Tree like)**.** A path $X : [0,1] \to E$ is tree-like if there exists a continuous function $h : [0,1] \to [0, +\infty)$ such that $h(0) = h(1) = 0$ and such that for all $s, t \in [0,1]$, $s \leq t$,

$$\|X_s - X_t\| \leq h(s) + h(t) - 2 \inf_{u \in [s,t]} h(u).$$

Two paths X and Y are called tree-like equivalent if $X * \overleftarrow{Y}$ is tree-like (where $\overleftarrow{Y}_t = Y_{1-t}$).

Example 2.3.2 (Tree-like path)**.** If a path is the concatenation of a smaller path with its time reversal, then it is tree-like, see figure 2.3.

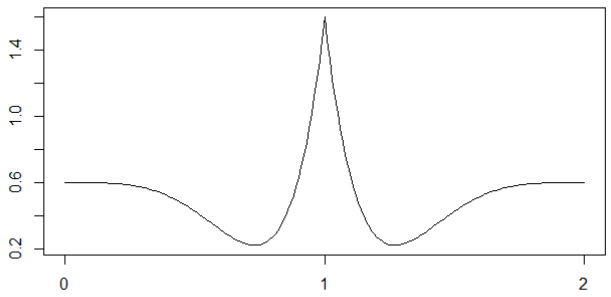


Figure 2.3: Tree-like path

A useful sufficient condition not to be tree-like is to have one monotonous coordinate. The paths we will encounter in real-world applications will usually be tree-like. In order to ensure uniqueness, for any path $X(t)$, we will consider the path $(t, X(t))$ which will then not be tree-like. We can now state the uniqueness theorem, proved in [14].

Theorem 2.3.2 (Uniqueness)**.** *Let X be a continuous path with bounded variation. Then,*

- *$S(X) = \mathbf{1}$ if and only if X is tree-like.*
- *The signature $S(X)$ is unique up to tree-like equivalence.*

In other words, this theorem states that the equivalence relation defined in Definition 2.3.2 is the same as the relation $X \sim Y \iff S(X) = S(Y)$. The result has been extended in [3] to geometric p-rough paths for some $p > 1$. In light of what was said before, the following lemma is a useful sufficient condition to ensure uniqueness of the signature.

Lemma 2.3.3. *Let $X : [0,1] \to E$ be of bounded 1-variation with at least one monotonous coordinate. Then $S(X)$ determines X uniquely.*

Conversely, some work has been done in order to find sufficient conditions under which the signature uniquely defines the path (up to reparametrization), see [2] and [19]. Finally, one can show some results on the norm of the signature. These will give us useful tools for computing rates of convergence of algorithms with a signature involved.

Proposition 2.3.2. *Let $X : [0,1] \to E$ be a path of bounded variation. Recall that $\|X\|_{1-var}$ is its 1-variation. Then, for any $k \geq 0$,*

$$\|\mathbf{X^k}\|_{E^{\otimes k}} \leq \frac{1}{k!}\|X\|_{1-var}^k < \infty$$

and

$$\|S(X)\|_{T((E))} \leq \exp(\|X\|_{1-var}) < \infty.$$

Proof. We have:

$$\begin{aligned}\|\mathbf{X^k}\|_{E^{\otimes k}} &= \left\| \int_{u_1,\cdots,u_k \in \Delta_k} X_{u_1} \otimes \cdots \otimes X_{u_k} \right\|_{E^{\otimes k}} \\ &\leq \int_{u_1,\cdots,u_k \in \Delta_k} \|dX_{u_1} \otimes \cdots \otimes dX_{u_k}\|_{E^{\otimes k}}\end{aligned}$$

by the triangle inequality (proof with Riemman sums or directly with Cauchy Schwartz).

By property of the tensor norm,

$$\|dX_{u_1} \otimes \cdots \otimes dX_{u_k}\|_{E^{\otimes k}} = \|dX_{u_1}\|_E .. \|dX_{u_k}\|_E .$$

Therefore,

$$\int_{u_1,\cdots,u_k \in \Delta_k} \|dX_{u_1}\|_E .. \|dX_{u_k}\|_E = \frac{1}{k!} \int_{u_1,\cdots,u_k \in [0,1]} \|dX_{u_1}\|_E .. \|dX_{u_k}\|_E ,$$

because $[0,1]^k$ can be partitioned into all possible orderings of the variables $u_1, \cdots, u_k$ and there are $k!$ such orderings. We thus get

$$\begin{aligned}\|\mathbf{X^k}\|_{E^{\otimes k}} &\leq \frac{1}{k!} \int_{u_1,\cdots,u_k \in [0,1]} \|dX_{u_1}\|_E \cdots \|dX_{u_k}\|_E \\ &\leq \frac{1}{k!} \int_{u_1 \in [0,1]} \|dX_{u_1}\|_E \cdots \int_{u_k \in [0,1]} \|dX_{u_k}\|_E \\ &\leq \frac{1}{k!} \left(\int_{u \in [0,1]} \|dX_u\|_E \right)^k \\ &\leq \frac{1}{k!} ||X||_{1-var}^k.\end{aligned}$$

The second inequality comes directly from the first one, using triangle inequality and Taylor expansion of the exponential. More precisely, we have

$$\|S(X)\|_{T((E))} = \sqrt{1+\sum_{k\geq 0}\|X^k\|^2_{E^{\otimes k}}} \leq 1+\sum_{k\geq 0}\|X^k\|_{E^{\otimes k}}$$
$$\leq 1+\sum_{k\geq 0}\frac{1}{k!}\|X\|^k_{1-var} = \exp(\|X\|_{1-var}).$$

□

2.3.2 Linear forms on the signature

One remarkable theorem is that the space of linear forms on the signature form an algebra of functions and thus approximate continuous functions arbitrary well (see Theorems 2.3.4 and 2.3.5). Statistically, this means that any non-linear relationship can be modelled by a linear model on the signature (see Section 3.2). For the moment, we present the space of linear forms on the signature and its structure of algebra. The multiplication is the shuffle product of elements of the tensor algebra.

Definition 2.3.3 (Shuffle product)**.** A permutation σ of $\{1,\dots,k+m\}$ is called a (k,m)-shuffle if $\sigma^{-1}(1) < \cdots < \sigma^{-1}(k)$ and $\sigma^{-1}(k+1) < \cdots < \sigma^{-1}(k+m)$

Let $I = (i_1,\dots,i_k)$ and $J = (j_1,\dots,j_m)$ be two multi-indexes with $i_1,\dots,i_k,j_1,\dots,j_m \in \{1,\dots,d\}$. Then, the shuffle product of I and J, denoted $I \sqcup J$, is a finite set of multi-indexes of length $k+m$:

$$I \sqcup J = \{(r_{\sigma(1)},\dots,r_{\sigma(k+m)}) | \sigma \in \text{Shuffles}(k,m)\}$$

with $(r_1,\dots,r_k,r_{k+1,\dots,r_{k+m}}) = (i_1,\dots,i_k,j_1,\dots,j_m)$.

This amounts to shuffling the words I and J without changing the order of their letters. There are $\frac{(m+k)!}{m!k!}$ elements in $I \sqcup J$. Note that the shuffle product is commutative and associative.

Theorem 2.3.3 (Shuffle product identity)**.** *For a path $X : J \to E$ and two multi-indexes $I = (i_1,\dots,i_k) \subset \{1,\dots,d\}^k$ and $J = (j_1,\dots,j_m) \subset \{1,\dots,d\}^k$, we have*

$$S^I(X)S^J(X) = \sum_{K\in I\sqcup J} S^K(X).$$

Proof. The result comes by partitioning the integration domain. Indeed, we have

$$S^I(X)S^J(X) = \int\cdots\int_{0<u_1<\cdots<u_k<1} dX^{i_1}_{u_1}\dots dX^{i_k}_{u_k} \int\cdots\int_{0<t_1<\cdots<t_m<1} dX^{j_1}_{t_1}\dots dX^{j_m}_{u_m}$$
$$= \sum_{\sigma\in\text{Shuffles}(k,m)} \int\cdots\int_{0<v_1<\cdots<v_{k+m}<1} dX^{r_{\sigma(1)}}_{v_1}\dots dX^{r_{\sigma(k+m)}}_{v_{k+m}}$$
$$= \sum_{K\in I\sqcup J} S^K(X),$$

with $(r_1,\dots,r_{k+m}) = (i_1,\dots,i_k,j_1,\dots,j_m)$. □

Example 2.3.3. For $I = \{1\}$ and $J = \{2\}$ we get

$$S^1(X)S^2(X) = S^{12}(X) + S^{21}(X).$$

For $I = \{1\}$ and $J = \{2,3\}$ we get

$$S^1(X)S^{23}(X) = S^{123}(X) + S^{231}(X) + S^{213}(X).$$

Not only is the shuffle product a nice property to compute the signature but it actually characterizes it. Before going further, we need some notation about linear forms on the tensor space.

First, let E^* be the dual space of E, that is the space of linear functions from E to $\mathbb{R}$. We canonically identify the space $T(E^*)$ (tensor space of the dual of E) with $T((E))^*$ the dual of the tensor space. Indeed, if $(e_1^*, \cdots, e_d^*)$ is a basis of E^*, then the set $(e_I^* = e_{i_1}^* \otimes \cdots \otimes e_{i_k}^*)_{\substack{I=(i_1,\cdots,i_k)\in\{1,\ldots,d\}^k \\ k\in\mathbb{N}}}$ is a basis of $T(E^*)$ and we identify it with a basis of $T((E))^*$ by setting

$$e_I^*(e_{j_1} \otimes \cdots \otimes e_{j_k}) = \delta_{i_1 j_1} \ldots \delta_{i_k j_k}.$$

We will now denote $T(E^*)$ the space of linear forms on $T((E))$. We define the shuffle product of two linear forms: let $f^*, g^* \in T(E^*)$, then we can write $f^* = \sum f_I e_I^*$, $g^* = \sum g_I e_I^*$ and we define

$$f^* \sqcup g^* = \sum_{K \in I \sqcup J} f_I g_J e_K^*.$$

A first consequence of the shuffle product property is the following theorem:

Theorem 2.3.4. *The space of linear forms on the signature endowed with the shuffle product is an algebra.*

Proof. We just need to check that the product of two linear forms on the signature is itself a linear form on the signature. It is true by the shuffle product property. □

This implies the following theorem.

Theorem 2.3.5 (Linear approximations)**.** *Let D be a compact subset of $BV^1(J, E)$ of paths that are not tree-like equivalent. Let $f : D \to \mathbb{R}$ continuous (in 1-variation norm). Then, for every $\varepsilon > 0$, there exists $w \in T((E))$ such that, for all $X \in D$,*

$$|f(X) - \langle w, S(X)\rangle| \leq \varepsilon$$

Proof. This is a direct consequence from Stone-Weierstrass theorem (see Appendix C.2). Indeed, let us consider

$$A = \operatorname{span}\left\{ f : X \mapsto \langle e_I, S(X)\rangle_{T((E))} | k \geq 0, I \subset \{1, \cdots, d\}^k \right\}.$$

A is a linear subspace of the set of continuous functions from D to $\mathbb{R}$, denoted $\mathcal{C}(D, \mathbb{R})$. Because of the shuffle property, it is also a sub-algebra. To apply Stone-Weierstrass theorem we must check that it contains a non-zero constant function and that it separates points. The first condition is met because the first term of the signature is one : $X \mapsto \langle e_{i_0}, S(X)\rangle_{T((E))} = 1$. The second condition is met because of the uniqueness of the signature: if $X, Y \in D, X \neq Y$, then $S(X) \neq S(Y)$, which means that at least one of their coordinates differ. By taking the corresponding basis one can find a function f in A such that $f(X) \neq f(Y)$. So A is dense in $\mathcal{C}(D, \mathbb{R})$.

□

We have proved a little more than the result stated : w will have only a finite number of non null coordinates because any function in A can be written as a finite linear combination of $\langle e_{i_1} \otimes \cdots \otimes e_{i_k}, S(X)\rangle_{T((E))}$.

2.3.3 Exponential, logarithm and Lie series

We will now see that the shuffle product property characterizes the signature, in the sense that any element of $T^N(E)$ that satisfies this property is the truncated signature of a path with bounded variation. Moreover, it can be shown that this property is satisfied if and only if the log of the element is a Lie formal series. Let us state this more precisely.

Definition 2.3.4. An element $a \in \tilde{T}((E))$ is group-like if the evaluation mapping

$$\begin{aligned} T(E^*) &\longmapsto \mathbb{R} \\ e^* &\longrightarrow e^*(a) \end{aligned}$$

is a morphism of algebras when $T(E^*)$ is endowed with the shuffle product.

We denote $G^{(*)}$ the space of group-like elements of $\tilde{T}((E))$. In other words, a is group-like if for any $e^*, f^* \in T(E^*)$, we have

$$e^*(a)f^*(a) = (e^* \sqcup f^*)(a).$$

We will see that the range of the signature is a subgroup of $G^{(*)}$ (and a small one), but also that, if $G^{(N)} = \pi_N(G^{(*)})$, then $G^{(N)}$ coincides exactly with the range of the signature truncated at order N. For this, we need to introduce the notion of log signature and Lie formal series.

Definition 2.3.5. Let $a \in T((E))$, then the exponential of a is the function $\exp : T((E)) \to \tilde{T}((E))$

$$\exp(a) = \sum_{k \geq 0} \frac{a^{\otimes k}}{k!}.$$

The logarithm is defined as $\log : \tilde{T}((E)) \to T((E))$

$$\log(a) = \sum_{k \geq 1} \frac{(-1)^k}{k} (\mathbf{1} - a)^{\otimes k}$$

Lemma 2.3.4. *Let $T_0(E) = \{a \in T((E)) : \pi_0(a) = 0\}$ the subset of $T((E))$ of elements whose first term is 0. Then $\exp : T_0(E) \to \tilde{T}((E))$ and $\log : \tilde{T}((E)) \to T_0(E)$ are each-other inverses.*

Example 2.3.4. We can write example 2.2.2 in a more compact form : when $X_t = X_0 + (X_T - X_0)t$ for $t \in [0, 1]$,

$$S(X) = \exp(X_T - X_0).$$

Thus the logarithm of a linear path is just

$$\log(S(X)) = X_T - X_0$$

($X_T - X_0$ is seen as an element of the tensor space with all orders null except the first one equal to $X_T - X_0$).

Definition 2.3.6 (Lie formal series)**.** We endow the tensor algebra with a Lie bracket : for $a, b \in T((E))$, we let

$$[a, b] = a \otimes b - b \otimes a.$$

For F_1 and F_2 linear subspaces of $T((E))$ we denote $[F_1, F_2]$ the linear span of all elements $[a, b]$ such that $a \in F_1$ and $b \in F_2$. Then, we define recursively $L_0 = 0$, $L_1 = E$, $L_2 = [E, L_1] = [E, E]$, $L_3 = [E, L_2] = [E, [E, E]]$,... and for any $n > 0$ $L_{n+1} = [E, L_n]$. L_n is

a linear subspace of $E^{\otimes n}$ and is called the space of homogeneous Lie polynomials of degree n.
Finally, we define the space of Lie formal series over E by

$$\mathcal{L}((E)) = \{l = (l_0, l_1, \cdots, l_n, \cdots) | \forall n \geq 0 \quad l_n \in L_n\}.$$

We have the following fundamental theorem.

Theorem 2.3.6. *For any $a \in \tilde{T}((E))$,*

$$a \in G^{(*)} \Leftrightarrow \log(a) \in \mathcal{L}((E)).$$

We can state the corresponding property for truncated spaces.

Lemma 2.3.5. *Let $\mathcal{L}^{(N)} = \pi_n(\mathcal{L}((E)))$, $\tilde{T}^{(N)}(E) = \pi_n(\tilde{T}((E)))$ and $a \in \tilde{T}((E))$. Then $a \in G^{(N)}$ if and only if $\log(a) \in \mathcal{L}^{(N)}$.*

Finally, the next proposition shows that $G^{(N)}$ is exactly the range of the signature.

Proposition 2.3.3. *Every element of $G^{(N)}$ is the truncated signature of a path of bounded variation. More precisely, for any $p \in [1, 2)$, $G^{(N)}$ is the range of the function*

$$\pi_N \circ S : BV^p(E) \to T^N(E).$$

Lemma 2.3.5 and proposition 2.3.3 tell us that $a \in T^N(E)$ is the truncated signature of a bounded variation path if and only if $\log(a) \in \mathcal{L}^{(N)}$. $\mathcal{L}^{(N)}$ having a much lower dimension than $\tilde{T}^{(N)}(E)$ (see section 4.2), we will use the log signature of a path instead of its signature as a feature.

Chapter 3

Learning with functional data

Back to the statistical problem, we can now describe more precisely our setting. We want to model a response Y (a real number or an integer) from a function $X \in BV(\mathbb{R}^d)$. We are provided a discretized version of X, so in the end we still have a vector but highly dimensional and with highly correlated covariates. We call X a function because its "nature" is functional and it has some regularity properties. In this section, we will review some strategies adopted in such a statistical setting. In its more general form, we assume that (X, Y) is a couple of random variables so that there exists a (smooth) function f^* such that

$$Y = f^*(X) + \varepsilon \tag{3.1}$$

with $\mathbb{E}[\varepsilon|X] = 0$. Different research communities have worked on this general problem. We find on the one hand the community of functional data analysis, an important reference being the book of Ramsay [28] and for the nonparametric case the book of Ferraty [10]. We will review their main model, the functional linear model, in section 3.1.2. On the other hand, the community of pattern recognition has also worked on this problem, focusing on times series representation. We will review some of their methods in section 3.1.1

3.1 Review of existing methods

3.1.1 Similarity measures

Time series modelling is an old statistical problem which has been revisited by the pattern recognition community. Indeed, because of the technological progress, time series are now highly dimensional, sampled at really small intervals, and traditional methods don't work in this context. A lot of work has thus been done on finding good similarity measures between curves and good high level representations. A review can be found in [27] and a systematic comparison of their predictive power has been undertaken in [8]. We describe here some of them and present in more details the Dynamic Time Wrapping (DTW) method, which shows the most promising results in [8].

The first and simplest measure is the Euclidean measure. A time series with n points is seen as a point in an n-dimensional Euclidean space and the distance between two time series is just their Euclidean distance. Similarly, one can define their L_p distance from the L_p norm in $\mathbb{R}^n$. A first problem is that two times series which are very similar but have been translated will have a big Euclidean distance. This can be solved by normalising

the time series before computing any distance. Still, it does not allow for acceleration or deceleration in time series, which is why the dynamic time warping measure has been introduced in [1]. The goal is to find a mapping of two time series along the time axis, so that the most similar parts of the curve are mapped together even if they take place at different time and have been stretched, see figure 3.1.

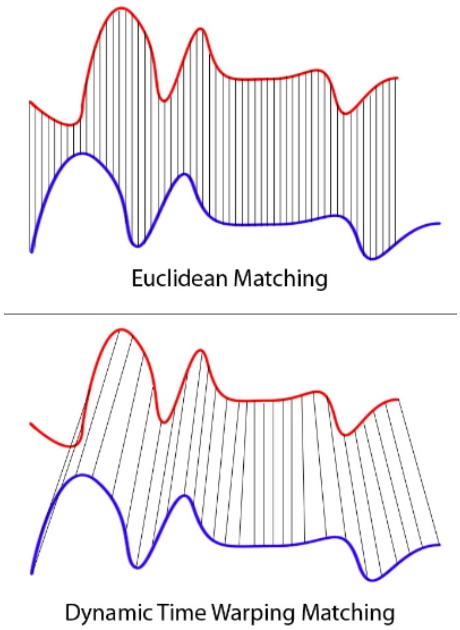


Figure 3.1: Dynamic time warping matching

The pseudo code of DTW is written in Algorithm 1. The input are two time series s and t of length n and m. A table of size $n \times m$ is constructed and the last entry is the DTW distance. The algorithm requires to have chosen a metric between two points of the series. A common choice is the Euclidean distance. We can see that the complexity is $O(nm)$, which is the main drawback of this algorithm. Some extensions have been proposed, which reduce the complexity to $O(n)$ by imposing a window when finding the best wrapping path.

Some other metrics have been constructed with the same kind of ideas. For example the, longest common subsequence similarity measure (LCSS) looks for the longest sequence of similar elements of the two series, while allowing to ignore some elements. Then a similarity measure is defined as $\frac{m+n-2\ell}{m+n}$ where ℓ is the longest common subsequence. This creates a metric more robust to outliers. See [27] for more details on these metrics.

When we have chosen a metric, any algorithm based on a metric can be used to classify or predict some output. The most commonly used is nearest neighbours but one can also construct kernels adapted to time series. In case of multivariate time series, these methods work as we can use the Euclidean metric in a higher dimensional space, but they do not contain any information about interaction between the different dimensions of the series. Moreover, they become computationally very intense. Still,, we will compare the DTW method with other methods in Table 4.3.

3.1.2 The functional linear model

We present here the fundamental model used in [28]. Contrary to the lasts section, this is a parametric model. Most of the work of the authors focus on estimating in this set-

Algorithm 1 Dtw algorithm

Require: s : array$[1 \cdots n]$, t : array$[1 \cdots m]$
  $DTW \leftarrow$ array$[0 \cdots n; 0 \cdots m]$
  for $i = 1$ to n **do**
    $DTW[i; 0] \leftarrow \infty$
  end for
  for $j = 1$ to m **do**
    $DTW[0; j] \leftarrow \infty$
  end for
  $DTW[0, 0] \leftarrow 0$
  for $i = 1$ to n **do**
    for $j = 1$ to m **do**
      cost $\leftarrow \|s[i] - t[j]\|$
      $DTW[i, j] \leftarrow$ cost + min$(DTW[i-1, j], DTW[i, j-1], DTW[i-1, j-1])$
    end for
  end for
  return $DTW[n, m]$

ting and reducing the dimension so that we are back in a classical finite dimensional regression setting. Assume we want to regress a scalar variable y on a functional predictor $x(t)$ for $t \in [0, 1]$ and that we have n data points. We assume the following model, first introduced in [29]. For $i = 1, \ldots, n$,

$$y_i = b_0 + \int_0^1 x_i(t)b(t)dt + \varepsilon_i, \tag{3.2}$$

with $b(t)$ a continuous coefficient, b_0 an intercept and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ the residual error.

We cannot use least squares method and minimize

$$\hat{b} = \underset{b}{\operatorname{argmin}} \sum_i \left(y_i - b_0 - \int_0^1 x_i(t)b(t)dt \right)^2$$

because b is infinite dimensional so we will always achieve a perfect fit, which is not desirable as we will strongly overfit the data. There are two main solutions to this dimension issue. Either we use a penalty term

$$\hat{b} = \underset{b}{\operatorname{argmin}} \sum_i \left(y_i - b_0 - \int x_i(t)b(t)dt \right)^2 + \lambda \int (Lb(t))^2 dt$$

with λ a smoothing parameter and L a linear operator (for example $Lb(t) = \frac{\partial^2 b(t)}{\partial t}$, called the total curvature penalty). The second solution is to reduce the class of possible functions for b, which is done with basis expansion, presented below. Both achieve regularization and can also be used together. A traditional approach is to use basis functions (see [28] section 15). Assume that b can be written as $b(t) = \sum_{j=1}^K b_j \varphi_j(t)$, we can rewrite model (3.2) as

$$y_i = b_0 + \sum_{j=1}^K b_j \int x_i(t)\varphi_j(t)dt + \varepsilon_i = b_0 + z_i^T b + \varepsilon_i$$

with $b = (b_1, \dots, b_K)^T$,$\varphi(t) = (\varphi_1(t), \dots, \varphi_K(t))^T$ and $z_i = \int x_i(t)\varphi(t)dt$. This is a linear model of dimension K. Then in matrix form the model is

$$y = Zb + \varepsilon$$

with the ith row of Z being z_i^T. The least squares solution is

$$\hat{b} = \text{argmin}(y - Zb)^T(y - Zb) = (Z^TZ)^{-1}Z^Ty.$$

If we add some penalty on b of the form $\lambda \int (Lb(t))^2 dt$, we get the solution

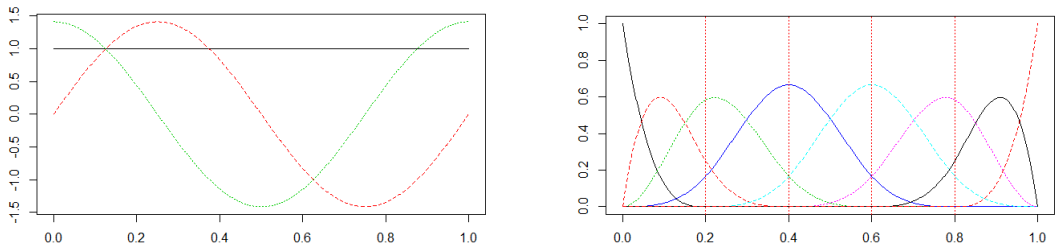
$$\hat{b} = \text{argmin}_b(y - Zb)^T(y - Zb) + b^TR_Lb = (Z^TZ + \lambda R_L)^{-1}Z^Ty$$

with R_L the penalization matrix $R_L = (\int L(\varphi_i(t)\varphi_j(t))_{1 \leq i,j \leq K}$.

Two commonly used basis functions are the Fourier basis and the B-spline basis, plotted in figure 3.2. The Fourier basis is defined as follows: if $K = 1 + 2m$, then

$$\varphi_0(t) = 1, \quad \varphi_{2j}(t) = \cos(2\pi mt), \quad \varphi_{2j-1} = \sin(2\pi mt) \text{ for } j = 1, \cdots, m.$$

B-splines are a really common set of basis functions: splines of order n are piecewise polynomial of degree $n - 1$, such that its first $n - 1$ derivatives are continuous at each knots. B-splines are defined such that any possible spline is equal to a unique linear combination of B-splines.



(a) Fourier basis for $K = 3$

(b) B-spline basis of order 4 with 4 knots

Figure 3.2: Basis functions

3.1.3 Functional principal component analysis

To conclude, we also present functional principal component analysis (fPCA), which has been a popular tool for modelling functional data, see [28] Chapter 8. Recall that in the last section, we have expanded b on a set of basis functions. An alternative is to use a basis for x, which is the method presented here. Let us recall how classical PCA is defined. Assume we have n data points $(x_1, \dots, x_n) \in \mathbb{R}^p$ and X is the design matrix in $\mathbb{R}^{n \times p}$. We assume that X has been centered : for any $1 \leq j \leq p$, $\sum_i x_{ij} = 0$.

- We define the first principal component as $f_1 = X\xi_1 \in \mathbb{R}^n$ with $\xi_1 \in \mathbb{R}^p$ the loadings such that the variance of f_1 is maximal and ξ_1 is normalized, i.e. we maximize $\sum_i f_{i1}^2$ such that $\|\xi_1\|^2 = \sum_j \xi_{1j}^2 = 1$. The f_{i1} are called the scores on the first PC and they are by definition $f_{i1} = \sum_j \xi_{1j}x_{ij}$, which is a linear combination of the predictors.

- We define similarly ξ_2 and $f_2 = \xi_2^T X$ maximizing $\sum_i f_{i2}^2$ such that f_2 is of norm 1 and uncorrelated with $f_1 : \sum_j \xi_{1j}\xi_{2j} = 0$.
- We define similarly $\xi_3, \dots, \xi_p$.

We observe that this problem is equivalent to finding the eigendecomposition of the covariance matrix $X^T X$. Indeed, we have $\sum_i f_{ik}^2 = f_k^T f_k = \xi_k^T X^T X \xi_k$ and for any matrix V the problem $\max_{\xi^T\xi=1} \xi^T V \xi$ is solved by finding the eigenvector corresponding to the largest eigenvalue of V. In our case, $V = X^T X$. It is a symmetric matrix so it has a spectral decomposition of the form $V = QDQ^T$ and ξ_k is the kth column of Q.

In the functional case, we define functional principal components similarly, mainly by replacing sums by integrals. More precisely, we define the weight functions $\xi_k : [0,1] \to \mathbb{R}$ and the principal component scores $f_{ik} = \int \xi_k(t) x_i(t) dt$. ξ_1 is chosen such that $\sum_i f_{i1}^2$ is maximal and $\int \xi_1(t)^2 dt = 1$. Similarly, ξ_2 maximizes $\sum_i f_{i2}^2$ such that $\int \xi_2(t)^2 dt = 1$ and $\int \xi_1(t)\xi_2(t)dt = 0$... Functional PCA can be expressed as an eigenvalue problem exactly like classical PCA. Indeed, let

$$v(s,t) = \frac{1}{N} \sum_i x_i(s) x_i(t)$$

be the covariance function. One can show that the weight functions are solution of

$$\int v(s,t)\xi(t)dt = \rho\xi(s)$$

for an appropriate value of ρ. If we define the operator $V : \xi \to \int v(\cdot,t)\xi(t)dt$, this can be written as $V\xi = \rho\xi$, which has the form of an eigenvalue problem. The functions ξ_k are also referred to as empirical orthonormal functions and one can show that if we define $\tilde{x_i}(t) = \sum_{k=1}^K f_{ik}\xi_k(t)$, then the ξ_k minimize the integrated squared error

$$\|x_i - \tilde{x_i}\|^2 = \int (x_i(t) - \tilde{x_i}(t))^2 dt.$$

Thus they are a good representation of x_i and can be used as a basis function for x_i.

Back to our regression problem, we then let

$$x_i(t) = \sum_j f_{ij}\xi_j(t)$$

If we plug it in equation (3.2), we get

$$y_i = b_0 + \int b(t) \sum_k f_{ij}\xi_j(t)dt = b_0 + \sum_j f_{ij} \int b(t)\xi_j(t)dt,$$

which gives

$$y_i = b_0 + \sum_j b_j f_{ij}$$

if $b_j = \int b(t)\xi_j(t)dt$. This is again a standard finite dimensional linear regression problem.

Note that both these methods are defined when there is only one predictor x which values in $\mathbb{R}$. Like the simple linear model, they can be extended to the multivariate case by adding explicit terms in (3.2). This supposes a parametric choice of the interaction between different terms. Moreover, these are linear models and thus they won't be able to model complex relationships.

3.2 Nonparametric regression with the signature features

We present now the signature method, which contrary to basis expansion and fPCA does not rely on any assumption concerning the relationship between the predictor function X and the output Y. Moreover, this method takes into account interaction between different dimensions of X, contrary to methods in Section 3.1.1, and linearises the problem so that it is computationally efficient. This method has been introduced in [21].

In light of theorem 2.3.5, we model Y as a linear function of the signature of the input X. The signature being an infinite object, we truncate it up to an order m. The best order is selected by cross validation and the regression coefficients are fitted with L_2 penalized least squares. Moreover, we have seen that the log signature is isomorphic to the signature and lives in a vector space, the free Lie algebra, which is of much smaller dimension than the full tensor space. Even if theorem 2.3.5 has been proven with the signature we will thus use the log signature features in our algorithms. To compute the signature, the data is linearly interpolated so that the formula of example 2.3.1 is valid. More details about the computations is given in section 4.2. To summarize, we can describe our approach by the following process:

$$\text{Path } X \longrightarrow \text{Features } S^m(X) \longrightarrow \text{Ridge regression} \longrightarrow \text{Prediction.}$$

3.2.1 Regression case

Let us first present our model and estimator in the regression case. Let D be a compact subset of $BV(\mathbb{R}^d)$ and $\mathcal{C}$ be defined by

$$\mathcal{C} = \{f : D \to \mathbb{R} | f \text{ continuous for } \|.\|_{1-var}\}.$$

We assume that (X, Y) is a random vector, $X \in BV(\mathbb{R}^d)$, $Y \in \mathbb{R}$. We make the assumption that $X \in D$ almost surely and that there exists a function $f^* \in \mathcal{C}$ such that

$$Y = f^*(X) + \varepsilon,$$

with $\mathbb{E}[\varepsilon|X] = 0$, $\mathbb{E}[\varepsilon^2|X] = \sigma^2$. We are thus in a random design setting. We estimate f^* by a linear function on the signature features, truncated up to a certain order m. Let us define these function spaces more precisely:

$$\begin{aligned}\mathcal{S}_m &= \{f : BV(\mathbb{R}^d) \to \mathbb{R} | \exists \beta \in T^m(\mathbb{R}^d), \forall X \in BV(\mathbb{R}^d), f(X) = \langle \beta, S^m(X) \rangle\} \\ &= \text{Span}\{\pi_I : X \to S^I(X) | I \subset \{1, \cdots, d\}^k, k \leq m\}\end{aligned}$$

with the notations introduced in section 2.2.1. In other words, we approximate the unknown function f^* by a function of the form

$$f(.) = \sum_{k=1}^{m} \sum_{I \subset \{1, \cdots, d\}^k} \beta_I S^I(.) \in \mathcal{S}_m.$$

We let the space of any finite linear form on the signature be $\mathcal{S} = \bigcup_{m \in \mathbb{N}} \mathcal{S}_m$. Note that each $\mathcal{S}_m$ is a linear subspace of $\mathcal{C}$, spanned by the functions π_I. Theorem 2.3.5 tells us that $\mathcal{S}$ is dense in $\mathcal{C}$ so that f^* can be approximated arbitrarily good by a function of $\mathcal{S}$.

The least squares estimator, denoted by $\hat{f}_{m,n}$, is defined to be

$$\hat{f}_{m,n} = \text{argmin}_{f \in \mathcal{S}_m} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

In practical applications, we will use a L_2 regularization, so that $\hat{f}_{m,n} = \langle \hat{\beta}_{m,n}, S^m(.) \rangle$ and

$$\hat{\beta}_{m,n} = \underset{\beta}{\operatorname{argin}} (Y_i - \langle \beta, S^m(X_i) \rangle)^2 + \lambda \|\beta\|_2^2$$

3.2.2 Classification case

For the classification case, we use a logistic regression model. Let us restate it more precisely. We consider a random vector (X, Y) so that Y takes only two values: $Y \in \{0, 1\}$, $X \in D$ a.s., and we assume that there exists a function $f^* \in \mathcal{C}$ such that

$$\mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x) = \frac{\exp(f^*(x))}{1 + \exp(f^*(x))}.$$

Then, we estimate f^* by $\hat{f}_{m,n}$ that maximises the likelihood of the data D_n in the class $\mathcal{S}_m$. More precisely, if $p_i = \mathbb{P}(Y = 1|X = X_i)$, $f \in \mathcal{S}_m$, the likelihood is equal to

$$\mathcal{L}(Y|X, f) = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1-Y_i}$$

and the log likelihood is then

$$\begin{aligned} \ell(f) = \log \mathcal{L}(Y|X, f) &= \sum_{i=1}^{n} Y_i \log \left(\frac{p_i}{1 - p_i} \right) + \log(1 - p_i) \\ &= \sum_{i=1}^{n} Y_i f(X_i) - \log(1 + \exp(f(X_i))). \end{aligned}$$

Therefore, we define

$$\hat{f}_{m,n} = \underset{f \in \mathcal{S}_m}{\operatorname{argmax}}\ \ell(f).$$

In applications, we will use some regularization on β, that is

$$\hat{\beta}_{m,n} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n} Y_i \langle \beta, S^m(X_i) \rangle - \log(1 + \exp(\langle \beta, S^m(X_i) \rangle)) - \lambda \|\beta\|_2^2.$$

3.2.3 Least squares rate of convergence

In this section, we present a theoretical bound on convergence of the signature regression method presented in Section 3.2.1, without regularization. We keep assumptions of Section 3.2.1 but for technical reasons, we also need to assume that f^* is uniformly bounded: there exists $L > 0$ such that $\|f^*\|_\infty = \sup_{X \in BV(\mathbb{R}^d)} |f^*(X)| \leq L$. Classically, we define the L^2 risk for a function $f \in \mathcal{C}$ by

$$\mathcal{R}(f) = \mathbb{E}(Y - f(X))^2$$

and, for an iid sample $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, the empirical risk is defined by

$$\mathcal{R}_n(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

We also define the truncated estimator $\hat{f}^L_{m,n}$ by the following: for any $x \in \mathbb{R}^d$,

$$\hat{f}^L_{m,n}(x) = \begin{cases} \hat{f}_{m,n}(x) \text{ if } |x| \leq L. \\ L \operatorname{sign}(x) \text{otherwise.} \end{cases}$$

The following theorem gives a rate of convergence of the L^2 risk of $\hat{f}^L_{m,n}$ to f^*, following the methodology of [13]. For simplicity of notation, we write $\|X\|$ for its 1-variation $\|X\|_{1-var}$.

Theorem 3.2.1. *Let $p(m)$ be the dimension of $\mathcal{S}_m$. For any $\delta > 0$, there exists constants M_δ and c such that*

$$\begin{aligned}\mathcal{R}(\hat{f}^L_{m,n}) - \mathcal{R}(f^*) \leq & c \max(\sigma^2, L^2)\frac{(\log n + 1)p(m)}{n} + \\ & \frac{M_\delta}{(m+1)!}\mathbb{E}\left[\|X\|^{m+1}e^{\|X\|}\left(\delta + M_\delta \frac{\|X\|^{m+1}}{(m+1)!}e^{\|X\|}\right)\right] + \delta^2.\end{aligned} \tag{3.3}$$

Proof. We can decompose the left hand side into estimation and approximation errors, which we will treat separately:

$$\mathcal{R}(\hat{f}_{n,m}) - \mathcal{R}(f^*) = \mathcal{R}(\hat{f}_{n,m}) - \inf_{f\in\mathcal{S}_m}\mathcal{R}(f) + \inf_{f\in\mathcal{S}_m}\mathcal{R}(f) - \mathcal{R}(f^*).$$

Approximation error: By density of $\mathcal{S}$ (see theorem 2.3.5) we know that, for any $\delta > 0$ small, there exists $m^* \in \mathbb{N}$ and $\beta^* \in T^{m^*}(\mathbb{R}^d)$ such that, for any $X \in D$,

$$|f^*(X) - \langle \beta^*, S^{m^*}(X)\rangle| \leq \delta. \tag{3.4}$$

We assume m^* is larger than m (otherwise we add zeros to β^*) so that we can decompose the scalar product as

$$\langle \beta^*, S^{m^*}(X)\rangle = \langle \beta^*_{1:m}, S^m(X)\rangle + \langle \beta^*_{m:m^*}, S^{m:m^*}(X)\rangle$$

with $S^{m:m^*}(X)$ the vector containing the signature terms from order $m+1$ to m^*. We can decompose the approximation error into two terms

$$\inf_{f\in\mathcal{S}_m}\mathcal{R}(f) - \mathcal{R}(f^*) = \inf_{f\in\mathcal{S}_m}\mathcal{R}(f) - \mathcal{R}(\langle \beta^*, S^{m^*}(.)\rangle) + \mathcal{R}(\langle \beta^*, S^{m^*}(.)\rangle) - \mathcal{R}(f^*). \tag{3.5}$$

First note that

$$\mathcal{R}(f^*) = \mathbb{E}(Y - f^*(X))^2 = \mathbb{E}(\varepsilon^2) = \sigma^2.$$

Then, the second term of (3.5) is bounded by

$$\begin{aligned}\mathcal{R}(\langle \beta^*, S^{m^*}(.)\rangle) - \mathcal{R}(f^*) &= \mathbb{E}(Y - \langle \beta^*, S^{m^*}(X)\rangle)^2 - \sigma^2 \\ &= \mathbb{E}(f^*(X) - \langle \beta^*, S^{m^*}(X)\rangle)^2 + \mathbb{E}\varepsilon^2 - \sigma^2 \\ &\leq \delta^2\end{aligned}$$

To deal with the other term, we decompose $\mathcal{R}(\langle \beta^*, S^{m^*}(.)\rangle)$ as follows:

$$\begin{aligned}\mathcal{R}(\langle \beta^*, S^{m^*}(.)\rangle) &= \mathbb{E}\left(Y - \langle \beta^*_{1:m}, S^m(X)\rangle - \langle \beta^*_{m:m^*}, S^{m:m^*}(X)\rangle\right)^2 \\ &= \mathbb{E}\left(Y - \langle \beta^*_{1:m}, S^m(X)\rangle\right)^2 + \mathbb{E}\left(\langle \beta^*_{m:m^*}, S^{m:m^*}(X)\rangle\right)^2 \\ &\quad - 2\mathbb{E}\left(\langle \beta^*_{m:m^*}, S^{m:m^*}(X)\rangle(Y - \langle \beta^*_{1:m}, S^m(X)\rangle)\right).\end{aligned}$$

First, we see that

$$\mathbb{E}\left(Y - \langle \beta^*_{1:m}, S^m(X)\rangle\right)^2 = \mathcal{R}(\langle \beta^*_{1:m}, S^m(.)\rangle) \geq \inf_{f\in\mathcal{S}_m}\mathcal{R}(f),$$

so that

$$\begin{aligned}
&\inf_{f\in\mathcal{S}_m}\mathcal{R}(f)-\mathcal{R}(f^*)\\
&\leq -\mathbb{E}\left(\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle\right)^2+2\mathbb{E}\left(\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle(Y-\langle\beta^*_{1:m},S^m(X)\rangle)\right)\\
&\leq \mathbb{E}\left(\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle\left(2\left(Y-\langle\beta^*_{1:m},S^m(X)\rangle\right)-\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle\right)\right)\\
&\leq \mathbb{E}\left(\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle\left(2\left(f^*(X)-\langle\beta^*,S^{m^*}(X)\rangle\right)+\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle+2\varepsilon\right)\right)\\
&\leq \mathbb{E}\left(\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle\left(2\delta+\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle+2\varepsilon\right)\right)
\end{aligned}$$

where we use (3.4) for the last inequality. Given that $\mathbb{E}[\varepsilon|X]=0$, the term with ε cancels out. Moreover, by Cauchy-Schwartz inequality, for any $X\in D$,

$$|\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle|\leq\|\beta^*_{m:m^*}\|\|S^{m:m^*}(X)\|.$$

By Proposition 2.3.2, the signature coefficients have an exponential decay, so that

$$\|S^{m:m^*}(X)\|=\sum_{k=m+1}^{m^*}\|\mathbf{X}^k\|_{E^{\otimes k}}\leq\sum_{k=m+1}^{\infty}\frac{\|X\|^k}{k!}=\frac{\|X\|^{m+1}}{(m+1)!}\exp(\|X\|).$$

and

$$|\langle\beta^*_{m:m^*},S^{m:m^*}(X)\rangle|\leq M_\delta\frac{\|X\|^{m+1}}{(m+1)!}\exp(\|X\|)$$

where $M_\delta=\|\beta^*_{m:m^*}\|$ is a constant depending on δ. Wrapping things up, we can bound the approximation error by

$$\inf_{f\in\mathcal{S}_m}\mathcal{R}(f)-\mathcal{R}(f^*)\leq\frac{M_\delta}{(m+1)!}\mathbb{E}\left[\|X\|^{m+1}e^{\|X\|}\left(\delta+M_\delta\frac{\|X\|^{m+1}}{(m+1)!}e^{\|X\|}\right)\right]+\delta^2.$$

Estimation error: We can now turn to the estimation error term

$$\mathcal{R}(\hat{f}_{n,m})-\inf_{f\in\mathcal{S}_m}\mathcal{R}(f).$$

The proof is a direct application of Theorem 11.3 from [13], we refer the reader to Appendix C.3 for more details. We need to reformulate the problem so that it has the same form as Theorem C.3.2. For any measurable function f, we have

$$\mathcal{R}(f)=\mathbb{E}(Y-f(X))^2=\mathbb{E}\left[(f^*(X)-f(X))^2\right]+\sigma^2$$

and for any estimator $\hat{f}_n$ depending on D_n, independent of (X,Y), we have,

$$\mathcal{R}(\hat{f}_n)=\mathbb{E}(f^*(X)-\hat{f}_n(X)+\varepsilon)^2=\mathbb{E}(f^*(X)-\hat{f}_n(X))^2+\sigma^2$$

because $\mathbb{E}(\varepsilon(f^*(X)-\hat{f}_n(X)))=\mathbb{E}((f^*(X)-\hat{f}_n(X))\mathbb{E}(\varepsilon|X))=0$

So

$$\mathcal{R}(\hat{f}^L_{m,n})-\inf_{f\in\mathcal{S}_m}\mathcal{R}(f)=\mathbb{E}(f^*(X)-\hat{f}^L_{m,n}(X))^2-\inf_{f\in\mathcal{S}_m}(f^*(X)-f(X))^2$$

and from Theorem C.3.2, one has

$$\mathcal{R}(\hat{f}^L_{m,n})-\inf_{f\in\mathcal{S}_m}\mathcal{R}(f)\leq c\max(\sigma^2,L^2)\frac{(\log n+1)p(m)}{n},$$

which completes the proof. □

This inequality is a bias-variance trade off: when the complexity of the class $\mathcal{S}_m$ increases, the first term increases as $p(m)$ and the second one decreases like $\frac{1}{(m+1)!}$. The first term is thus a variance term and the second one a bias term. One can see this behaviour in experiments on cross-validation curves (see figure 4.5). The dimension of the vector space $\mathcal{S}_m$ is the same as the dimension of the free Lie algebra $\mathcal{L}^{(m)}$, defined in section 2.3.3. The exact form of $p(m)$ will be given in section 4.2 and is quite involved. To conclude, this ridge regression algorithm is totally non-parametric, it reduces the analysis to a linear problem (when the signature features have been computed) and decreases strongly the dimensionality of the problem.

3.2.4 Conclusion and future work

Now that we have obtained inequality (3.3), the next step is to look for an adaptive estimate, minimizing the risk for several class of functions (i.e., several m) simultaneously. It should also be relevant to compute (3.3) with some assumptions on f^* (Lipschitz continuous, Hölder...) and compare the resulting bound with the ones obtained with other representations (wavelet, spline basis...). Furthermore, one needs to extend the result 3.2.1 to the classification case. Finally, it will be interesting to investigate theoretical properties of the signature of paths of unbounded variation. It would extend the method to highly oscillatory signals, which is useful in many applications. We would also like to investigate other algorithms combined with the signature features. Indeed, the major success of the signature has been achieved with a convolutional neural network (see [34]) and it would be interesting to investigate the properties of the signature features in this setting.

Chapter 4

Experimental results

We can now proceed to a description of the experimental results, which show the potentiality of the signature in real-world applications.

4.1 Data

4.1.1 Real world datasets

Let us first describe the different data sets we have used. We have tried to use data sets of different nature. The first one is a one-dimensional data set from [5], called ECG200 dataset, see Figure 4.1. It consists of a set of electrocardiogram records of normal and Myocardial infarction heartbeats. There are 100 observations in the training set and 100 in the test set.

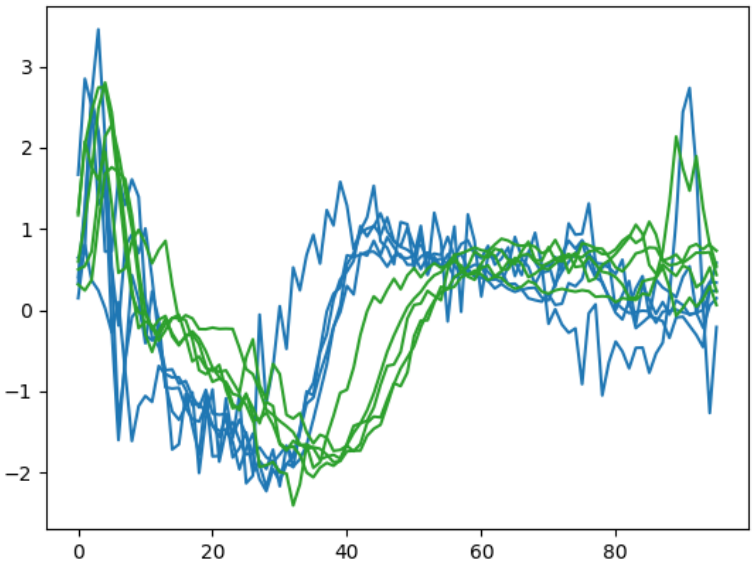


Figure 4.1: 5 samples of each class of the ECG dataset

The second dataset is a character trajectories dataset from the UCI Machine learning

repository [7], see Figure 4.2. It consists of 2858 samples of a 3-dimensional time series: position coordinates x and y and pen tip force. The data has been preprocessed with first order differentiation and Gaussian smoothing.

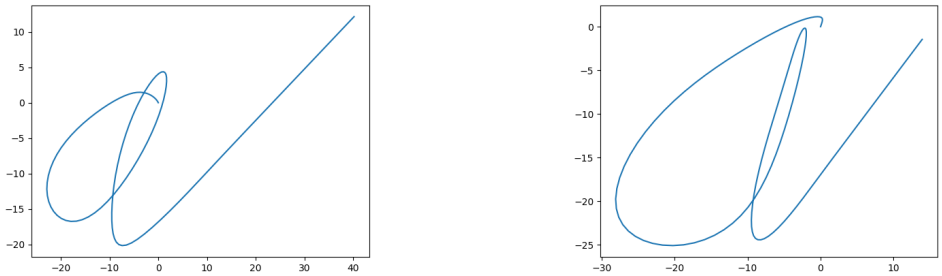


Figure 4.2: Two samples of the letter a in the Character trajectory dataset.

The last dataset is called Japanese Vowels and consists of 640 time series of 12 LPC cepstrum coefficients (see [18]) taken from 9 male speakers uttering the vowel /ae/. The goal is to determine who is speaking from the 12-dimensional time series. 9 of the 12 dimensions of one curve are plotted in Figure 4.3.

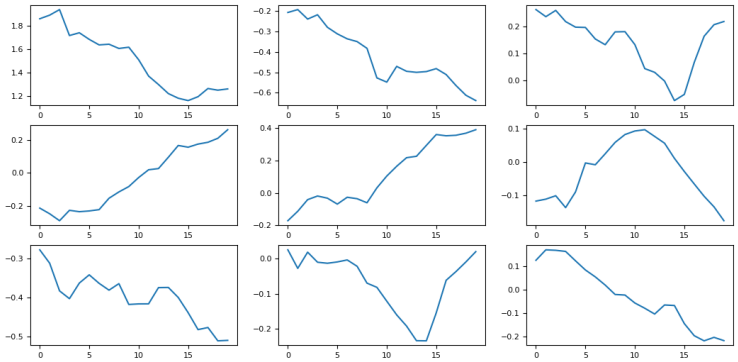


Figure 4.3: 9 coordinates of a sample of the Japanese Vowels dataset

4.1.2 Simulated datasets

It is also useful to use data for which we know the underlying regression function. We thus create the following data: $X : [0,1] \to \mathbb{R}^d$, $X_t = (X_t^1, \cdots, X_t^d)$ and the kth component of X is the function $X_t^k = \alpha_{1k} + 10\alpha_{2k}\sin(\frac{2\pi t}{\alpha_{3k}}) + \alpha_{4k}(t - \alpha_{5k})^3$, with all the parameters α sampled uniformly on $[0,1]$, see figure 4.4. Thus we have a dataset for any possible dimension d.

Then we create different responses to these inputs. For regression, we define

- Linear relationship:

$$Y = \frac{1}{d}\sum_{k=1}^{d}\int_0^1 \cos(2\pi t) X_t^k dt + \varepsilon,$$

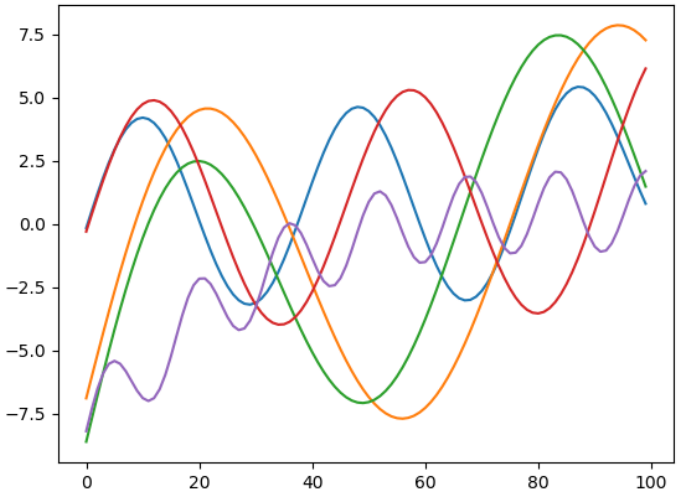


Figure 4.4: 5 different realizations of X_t

- Nonlinear relationship:

$$Y = \frac{1}{d}\sum_{k=1}^{d}\int_0^1 \cos(2\pi t)(X_t^k)^2 dt + \varepsilon,$$

- Nonlinear relationship with interaction:

$$Y = \frac{1}{d-1}\sum_{k=1}^{d-1}\int_0^1 (X_t^k)^2 \log(|X_t^{k+1}|)dt + \varepsilon,$$

- Sparse relationship:

$$Y = \int_0^1 \cos(2\pi t)(X_t^1)^2 dt + \varepsilon.$$

For classification, we can them transform Y into a binary response : $Y^{\text{new}} = \mathbb{1}_{Y \geq M}$ with M a constant properly chosen so that the dataset is as balanced as possible.

4.2 Computing the signature

For our approach to be practically relevant, one needs to have an efficient way of calculating signature features. To our current knowledge, two Python packages are available: the `esig` package from CoRoPa [9] and the `iisignature` package. We have used the latter, which is more recent and more efficient for our data. Indeed, the CoRoPa project focuses on sparse signature arrays, when a lot of signature coefficients are null, whereas `iisignature` is more efficient for dense signature arrays, see [31] for a performance benchmark of these two implementations.

We have defined in section 2.3.3 the space of the homogeneous Lie polynomials of degree m, $\mathcal{L}^{(m)}(E)$ and stated that the log-signature truncated at order m belongs to this space. For computations, we still need to construct a basis of this vector space, so that the log signature is described with a minimum number of coefficients. Several basis exist,

one of them being the Lyndon basis, described in [30] and it is the default option in the `iisignature` package. Lyndon basis is obtained by constructing a bijection between Lyndon words of length m on an alphabet $\mathbf{1}, \ldots, \mathbf{d}$ and $\mathcal{L}^{(m)}(E)$. A Lyndon word of length m is such that it is strictly smaller for the lexicographic than any of its rotations.

Example 4.2.1 (Lyndon words)**.** We present some examples of Lyndon words.

- The word **1223** is a Lyndon word, but its permutations **2231**, **2312** and **3122** are not.
- The empty word, every single-letter world and every word with a repeating pattern (e.g. **123123**) are not Lyndon words.

The number of possible Lyndon words gives the dimension of $\mathcal{L}^{(m)}(E)$, denoted $p(m)$ in section 3.2.3, and is equal to

$$p(m) = \frac{1}{m} \sum_{q|m} \mu(\frac{m}{q}) d^q, \tag{4.1}$$

with the sum being over all possible divisors q of m and μ being the Möbius function, see [32] and the references given there. We give some of these dimensions in Table 4.1 for comparison with the dimension of the entire tensor space.

m	0	1	2	3	4	5	6	7
$T^m(\mathbb{R}^2)$	1	3	7	15	31	63	127	255
$\mathcal{L}^{(m)}$	0	2	3	5	8	14	23	41

Table 4.1: Comparison of dimensions of $T^m(\mathbb{R}^2)$ and $\mathcal{L}^{(m)}(\mathbb{R}^2)$

Now that we have a basis on which we can represent the log signature, we need to be able to compute it. In `iisignature`, the path is linearly interpolated so the task is reduced to the case of piecewise linear paths. We have seen that there exists a closed form formula for the signature (see example 2.3.1) thanks to Chen's identity. It is a little more complicated to compute the log-signature directly. Indeed, we know the log signature of a linear path (it is just its displacement, see Example 2.3.4) but Chen's formula 2.3.1 is not valid any more so we cannot compute the log signature of the concatenation of two paths directly from the log signature of each segment. However, the log signature of the concatenation of two paths can be computed as the Baker-Campbell-Hausdorff (BCH) product of the two log-signatures. In its more general form, the BCH formula gives the expression in a Lie algebra of $\log(e^x e^y)$ when x and y are elements of a Lie algebra. In other words, it expresses the logarithm of the product of two Lie group elements as a Lie algebra element. In our case, let $\ell(X)$ and $\ell(Y)$ be the log signatures of two paths X and Y, then the signature of the concatenation $X * Y$ is $e^{\ell(X)} \otimes e^{\ell(Y)}$ and its log signature is $\log(e^{\ell(X)} \otimes e^{\ell(Y)})$. The first terms of the BCH formula are given below:

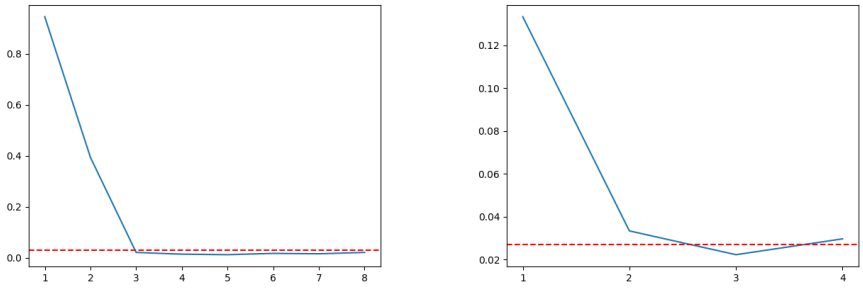
$$\log(e^X \otimes e^Y) = X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}([X, [X, Y]] + [Y, [Y, X]]) + ...$$

This gives a method for computing directly the log signature of a piecewise linear path.

4.3 Procedure

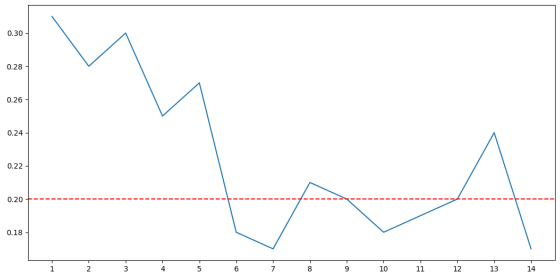
Let us now describe the algorithm we have implemented which we call later "Ridge with signature". First, we split the data into a training set and a test set. On the training

set, we perform a 5-folds cross validation to determine the best truncation order of the signature. More precisely, we split the data into 5 folds, we iteratively leave one out, fit a ridge (logistic) regression with log-signature features up to order m and test the prediction error on the fold left apart. For classification, the error metric used is the misclassification rate : the proportion of samples which have not been classified in the right class. For regression, it is the relative mean squared error: $RMSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$ with $\bar{Y}$ the sample average of Y.



(a) Character trajectories

(b) Japanese vowels



(c) ECG

Figure 4.5: Error rate as a function of truncation order, obtained by 5-folds cross validation for our 3 datasets. The dotted red line is the error rate of a naive regression with raw features.

We plot in figure 4.5 the cross-validation curves, that is the prediction error as a function of the truncation order. We can see that the optimal truncation order is 7 for the ECG dataset, 5 for the Character trajectories and 3 for the Japanese vowels. It achieves a better error rate than a naive regression where the positions are stacked together, while it also reduces the dimension from 615 to 294 features for the character trajectories and from 96 to41 for ECG. The dimension for the Japanese Vowels actually increases because the initial trajectories are sampled with only 29 points so that the raw vector is of dimension 348 and the signature one of size 819.

4.4 Results

We describe in this section some of the experiments we have conducted in order to better understand properties of the signature. For this, we have used the real data described in section 4.1 as well as some simulated datasets.

4.4.1 Study of the log transformation

A first question we have asked ourselves is whether considering the log signature instead of the signature changes the results. Indeed, we have seen in previous sections that the log signature is a lower dimensional representation of the signature, thus a better one from a statistical point of view. We have compared experimentally the results of regression on the signature and on the log signature for several datasets. We show for example the result for the simulated dataset with a non linear response and $d = 1$. We can see that using the signature features give slightly better results for small truncation orders but that they both converge to the same error so that from order 8 there is hardly any difference. On the x-axis, the dimensions of the coefficients corresponding to this truncation order are shown. We can see that at order 8 the error rates are similar and taking the log decreases the dimension from 510 to 71.

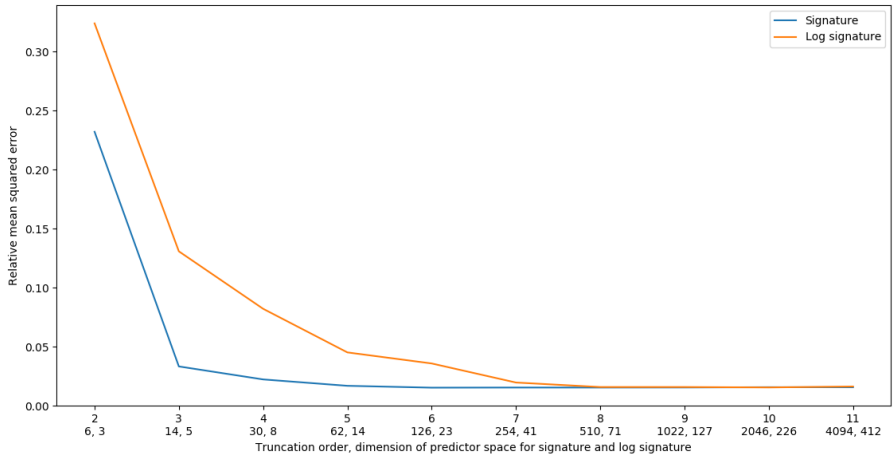


Figure 4.6: Comparison of relative mean squared error for the 1-dimensional simulated data set with non linear response.

We also report in table 4.2 the error rate differences for the 3 experimental datasets. For each dataset we report the error rate for the truncation order minimizing the cross validation error. We can see that the signature yields only slightly better results.

Dataset	Truncation order	Log signature	Signature
Character trajectories	5	0.013	0.013
ECG	8	0.18	0.16
Japanese vowels	3	0.024	0.016

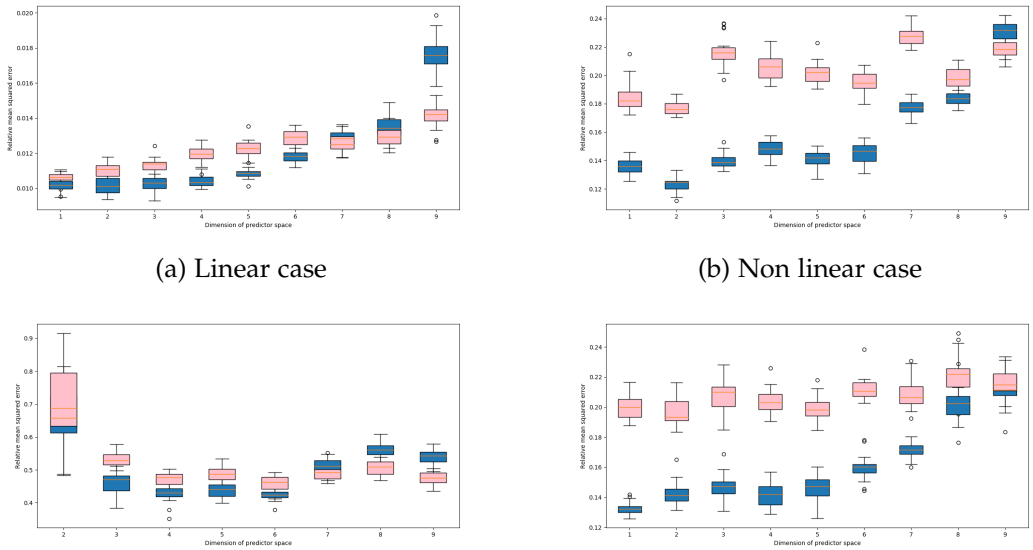
Table 4.2: Comparison signature and log signature results

4.4.2 Study of influence of dimension of X

We have also studied how the signature behaves when the dimension of X increases. More precisely, we have a process $X : [0,1] \to \mathbb{R}^d$ and a response $Y = f(X) + \varepsilon$ and we want to know what happens when d increases. For this we study the evolution of the error rate when we increase the dimension of the simulated dataset. We compare the performance of the signature truncated to order 3 and the raw regression when all dimensions are stacked in one vector.

We use the following procedure :

1. We simulate a data set of 2000 samples (X_i, Y_i) according to one of the models described in section 4.1.2.
2. We randomly split it into a training set and a test set (half of the observations in each). We fit a regression model with raw features and signature truncated at order 3. We compute the error on the test set.
3. Repeat the last 2 steps 20 times to obtain a boxplot accounting for the randomness in the splitting process.



(a) Linear case

(b) Non linear case

(c) Interaction case

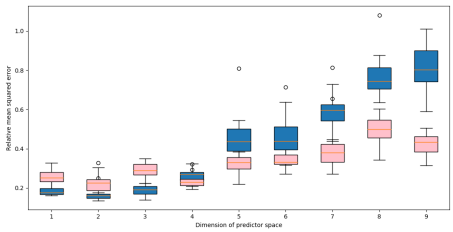
(d) Sparse case

Figure 4.7: Boxplot of the *RMSE* for 2000 samples of different simulated datasets. Blue : ridge with signature features. Pink : ridge with raw features.

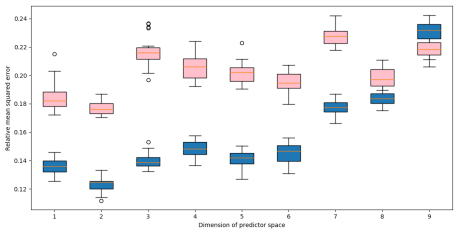
Figure 4.7 shows the results. One can observe that when signature features are used, the error decreases in almost all settings. It decreases most in the non-linear and sparse case, which suggests that the signature is able to model any complex, non linear, relationship between X and Y. In the linear case, the model with raw features is the true model so it should perform really well but even in this case the signature performs a little better. Concerning the behaviour with regard to the dimension of X, we can see that the error always increases with the dimension of X. There seems to be a threshold d after which raw features show a smaller error than signature features. To understand this better, we investigate the evolution of this threshold with sample size. We observe in figure 4.8 that this threshold increases the sample size. This suggests that asymptotically, the signature has better properties. Finally, for the interaction case, both algorithms have quite poor results, the *RMSE* is between 0.5 and 1.

4.4.3 Study of noise influence

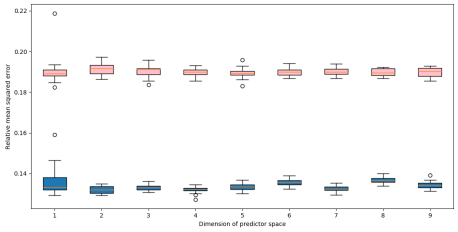
With the simulated dataset, we can also vary the variance of the noise ε. We plot in figure 4.9 the results for a non linear response. We can observe that when the signature features are used, the error increases with noise whereas the error stays similar when raw data is used.



(a) 200 samples



(b) 2000 samples



(c) 20000 samples

Figure 4.8: *RMSE* as a function of dimension for a non linear response. Blue : ridge with signature features. Pink : ridge with raw features.

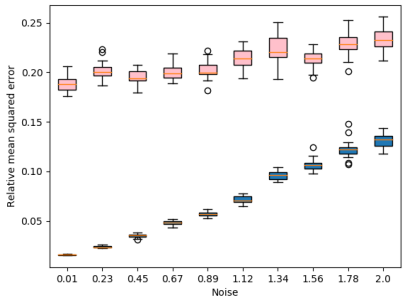


Figure 4.9: Relative MSE as a function of the variance of the noise for a non linear response. Blue : ridge with signature features. Pink: ridge with raw features.

4.4.4 Comparison of prediction performances

We compare the performance of the ridge regression with signature features with some other classical algorithms. The results are summarized in table 4.3. For Ridge, Random Forest, 1-NN and XGBoost, we use as features a vector of discretization of the path. When it is multidimensional, we just stack them into a unique, long vector. The algorithms are fitted with the Python library `scikit-learn` [25] except Nearest Neighbours which are fitted with `tslearn`. The regularization parameter in Ridge is chosen by the default efficient Leave-One-Out cross validation of the function RidgeCV in `scikit-learn`. For the Random Forest, we have used 20 trees and 30% of features at each split. For XGBoost, we have used the default parameters.

One can see in table 4.3 that the signature achieves the best or close to the best prediction accuracy, while being computationally less intense. Indeed, we have seen in figures 4.5 that a small truncation order is usually selected. Another interesting observation is that the best error reduction is achieved on the Character trajectories dataset. This agrees with the fact that the signature coefficients encode information about geometry of the path, as these are the important aspects for recognising character trajectories. Moreover,

Algo	Character trajectories	ECG	Japanese Vowels
Ridge with signature	**0.012**	0.18	**0.024**
Ridge	0.028	0.2	0.027
Random forest	0.029	0.18	0.059
1-NN	0.028	**0.1**	0.075
1-NN dtw	0.020	0.23	0.051
XGBoost	0.027	0.21	0.083

Table 4.3: Comparison of results

it improves accuracy for the two multidimensional datasets whereas for the simpler data ECG, a simple nearest neighbours algorithm is better. The signature captures interaction between different dimensions and thus captures information that other algorithms do not.

Appendix A

Tensor product space

A.1 Constructive definition

We present here a constructive definition of the tensor product space as well as the universal property (see [26]). Let us recall the definition of a tensor product.

Definition A.1.1 (Tensor product of vector spaces)**.** Let V and W be two vector spaces over a field K. A tensor product of V and W denoted by $V \otimes W$ is a vector space over the same field K with a bilinear map $\varphi : V \times W \to V \otimes W$ such that for any basis $e = (e_i)_{i \in I}$ of V and $f = (f_j)_{j \in J}$ then

$$\varphi(e \times f) = \{\varphi(e_i, f_j) | e_i \in e, f_j \in f\}$$

is a basis of $V \otimes W$.

Definition A.1.2 (Free vector space)**.** Let A be a set and K a field. Then the free vector space over K generated by A is the space of all formal finite linear combinations of elements of A, denoted by $F(A)$.

Thus A is a basis of $F(A)$.

Definition A.1.3 (Quotient space)**.** Let V be a vector space over a field K and N a subspace of V. We define an equivalence relation $\sim$ on V : for any $x, y \in V$,

$$x \sim y \quad \iff x - y \in N,$$

and we denote its equivalence classes as $[x]$. Then, the quotient space V/N is defined as the set of all equivalence classes of V for $\sim$, with addition and scalar multiplications defined as

$$[x] + [y] = [x + y],$$
$$\alpha[x] = [\alpha x],$$

for any $x, y \in V$, $\alpha \in K$. The map $x \in V \to [x] \in V/N$ is called the quotient map.

The tensor space of two vector spaces V and W can be constructed as the quotient of the free vector space of the Cartesian product of V and W, $F(V \times W)$, by the subspace N consisting of all formal series of $F(V \times W)$ that are equal to 0 by bilinearity. More precisely, we define N as the subspace of $F(V \times W)$ spanned by vectors of the form

$$(u + kv, w + lx) - (u, w) - k(v, w) - l(u, x) - kl(v, x),$$

for $k, l \in$, $u, v \in V$ and $w, x \in W$. We will prove in theorem A.1.2 that $F(V \times W)/N$ satisfies Definition A.1.1. Let $\pi : F(V \times W) \to F(V \times W)/N$ be the quotient map, we define for $v \in V, w \in W$

$$\varphi(v, w) = \pi((v, w)).$$

Proposition A.1.1. *φ defined above is bilinear.*

Proof. This comes directly from the definition of N. Indeed, we need to show that for any $k, l \in$, $u, v \in V$ and $w, x \in W$, we have

$$\varphi(u + kv, w + lx) = \varphi(u, w) + k\varphi(v, w) + l\varphi(u, x) + kl\varphi(v, x).$$

We know that for any $z \in N$, $\pi(z) = 0$ (by definition of the quotient map) so $\varphi(z) = 0$ and one has $(u + kv, w + lx) - (u, w) - k(v, w) - l(u, x) - kl(v, x) \in N$ hence the result by bilinearity of π. □

We need the following lemma to prove the universal property and the fact that $F(V \times W)/N$ is indeed a tensor product space.

Lemma A.1.1. *Let V, W vector spaces over $\mathbb{R}$, $T : V \to W$ linear and S subspace of V. Then there exists $\bar{T} : V/S \to W$ linear such that $\bar{T}(x + S) = T(x)$ for all $x \in V$ if and only if $T(s) = 0$ for all $s \in S$.*

Proof. Let us assume that $\bar{T}$ exists, then, for any $s \in S$, $T(s) = \bar{T}(s + S) = \bar{T}(0) = 0$. Conversely, if $T(s) = 0$ for all s then $\bar{T}$ is well defined : if $x, y \in V$ are such that $x + S = y + S$, then $T(x) = T(y)$ (because $x - y \in S$) and $\bar{T}(x + S) = \bar{T}(y + S)$. □

Theorem A.1.1 (Universal property)**.** *Let V and W be two vector spaces over a field K, $V \otimes W = F(V \times W)/N$ and φ as defined above. Then, for every vector space X and bilinear map $f : V \times W \to X$, there exists a unique linear map $T : V \otimes W \to X$ such that $f = T \circ \varphi$.*

In other words, giving a bilinear map from $E \times F$ is the same as giving a linear map from $E \otimes F$.

Proof. This is a consequence of the previous lemma. Indeed, as $V \times W$ is a basis of $F(V \times W)$ we can extend by linearity f to a mapping from $F(V \times W)$ to X. Then, as f is bilinear, we have, for any $z \in N$, $f(z) = 0$ (by a similar argument as in propostion A.1.1) and thus we can apply the lemma : there exists a linear $T : F(V \times W)/N \to X$ such that $T(\varphi(v, w)) = f(v, w)$. □

This universal property is often taken as a definition of the tensor product.

Theorem A.1.2. *$F(V \times W)/N$ with the bilinear map φ is a product space, that is for any basis $e = (e_i)_{i \in I}$ of V and $f = (f_j)_{j \in J}$ then*

$$\varphi(e \times f) = \{\varphi(e_i, f_j) | e_i \in e, f_j \in f\}$$

is a basis of $F(V \times W)/N$.

Proof. Show that it is spanning by directly writing what is an element of $F(V \times W)/N$. Show linear independence by using the universal theorem for a particular function. See [26] for details. □

Proposition A.1.2 (Uniqueness of the tensor product)**.** *Let $((V \otimes W)_1, \mu)$ a tensor product satisfying definition A.1.1. Then there exists a isomorphism $T : F(V \times W) \to (V \otimes W)_1$ such that*

$$\mu = F \circ \varphi.$$

In other words, the tensor product is unique up to isomorphisms.

A.2 Norm on tensor product

Definition A.2.1 (Admissible norm)**.** Let V a Banach space. We say that its tensor powers are endowed with admissible norms if

1. For any $n \geq 1$, for any permutation S of elements of $v \in V^{\otimes n}$, then

$$\|Sv\| = \|v\|.$$

2. For all $n, m \geq 1$, $v \in V^{\otimes n}$, $w \in V^{\otimes m}$,

$$\|v \otimes w\| \leq \|v\| \|w\|.$$

Appendix B

Lie group

This section is based on the lecture notes [15] and from the textbook [20].

B.1 Definition of a Lie group

Definition B.1.1 (Topology)**.** Let X be a set and τ a family of subsets of X. τ is a topology if

- $\emptyset$ and X are in τ.
- Any union of elements of τ is in τ.
- Any finite intersection of elements of τ is in τ.

If τ is a topology, then (X, τ) is called a topological space.

A map $f : X \to Y$ between two topological spaces X and Y is called an homeomorphism if it is continuous, invertible and its inverse is continuous.

Definition B.1.2 (Topological group)**.** Let G be a group. A topology $\tau \subset \mathcal{P}(G)$ endows G with a structure of a topological group if the product map $G \times G \to G$ and the inverse map $G \to G$ are continuous.

Example B.1.1. We give two basic examples of topological groups.

- $(\mathbb{R}^n, +)$ endowed with the Euclidean topology is a topological group.
- $GL(n, \mathbb{R}) = \{X \in M_{n,n}(\mathbb{R}) | \det(X) \neq 0\}$ equipped with the natural topology (via the identification of $M_{n,n}$ with $\Re^{n.n}$).

Definition B.1.3 (Locally compact Hausdorff space)**.** A topological space X is Hausdorff if any two distinct points have disjoint neighbourhoods. It is locally compact if for all $x \in X$ and for any V neighborhood of x, there is a compact neighbourhood W of x such that $x \in W \subseteq V$.

A topological space is called second countable if its topology has a countable base (B is a base of a topology τ if any element of τ can be written as a union of elements of B). We also recall that a function $F : U \to V$ for U and V open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ is called smooth if all its coordinates have continuous partial derivatives of all orders. If in addition it is invertible and its inverse is mooth, F is called a differomorphism.

Definition B.1.4 (Topological n-manifold)**.** A topological n-manifold M is a Hausdorff, second countable topological space such that every point of M has a neighbourhood which is homeomorphic to an open subset of $\mathbb{R}^n$. A pair (U, φ) consisting of an open subset $U \subseteq M$ and a map $\varphi : U \to \varphi(U) \subseteq \mathbb{R}^n$ which is an homeomorphism onto its image is a (coordinate) chart at any point of U.

A smooth structure on M is an atlas $\mathcal{A} = \{(U_\alpha, \varphi_\alpha) | \alpha \in A\}$ whose domain covers M and such that for all $\alpha, \beta \in A$, the map $\tau : \varphi_\alpha(U_\alpha \cap U_\beta) \to \varphi_\beta(U_\alpha \cap U_\beta)$ is smooth (that is $\mathcal{C}^\infty$). τ is called the transition map from φ_α to φ_β. In this case, M is a smooth n-manifold.

Example B.1.2. Let E be a real vector space of dimension $d < \infty$. Then it is a topological d-manyfold, for the topology induced by a norm on E. Indeed, let $(e_1, ..., e_d)$ be a basis of E, and $\pi : \mathbb{R}^d \to E$ the isomorphism

$$\pi : x \mapsto \sum_{i=1}^{d} x_i e_i.$$

It is an homeomorphism so (E, π^{-1}) is a coordinate chart.
We can also define a smooth structure on E by defining the atlas of all such coordinate charts for any basis of E. Then let $(e_1, ..., e_d)$ and $(\tilde{e}_1, ..., \tilde{e}_n)$ and other basis of E, φ and ψ their associated maps, there exists an invertible matrix $A \in \mathbb{R}^{d \times d}$ such that

$$e_i = \sum_{j=1}^{d} A_{ij} \tilde{e}_j,$$

so the transition map between the two charts is given by $\tau = \psi^{-1} \circ \varphi$ so that if $x \in \mathbb{R}^d$, $\tau(x) = \tilde{x} \in \mathbb{R}^d$ we have

$$\psi(\tau(x)) = \sum_{j=1}^{d} \tilde{x}_j \tilde{e}_j = \varphi(x) = \sum_{i=1}^{d} x_i e_i = \sum_{i,j=1}^{d} x_i A_{ij} \tilde{e}_j.$$

So $\tilde{x}_j = \sum_{i=1}^{d} A_{ij} x_i$, $\tau(x) = Ax$ is an invertible linear map and hence a diffeomorphism.

Definition B.1.5 (Smooth function)**.** Suppose M is a smooth n-manifold, $k > 0$ and $f : M \to \mathbb{R}^k$ any function. We say that f is a smooth function if for any $p \in M$, there exists a chart (U, φ) whose domain contains p and such that $f \circ \varphi^{-1} : \varphi(U) \subseteq \mathbb{R}^n \to \mathbb{R}^k$ is smooth.

The set of all smooth real-valued functions is denoted $\mathcal{C}^\infty(M)$.

We can extend this definition for functions between manifolds.

Definition B.1.6. Let M, N be smooth manifolds. $F : M \to N$. F is said to be smooth if, for any $p \in M$, there exists a chart (U, φ) with $p \in U$ and (V, ψ) with $F(p) \in V$ such that $\psi \circ F \circ \varphi^{-1} : \varphi(U) \to \psi(V)$ is smooth.

Definition B.1.7 (Real Lie group)**.** A Lie group G is a group endowed with the structure of a smooth manifold, in which the operations of multiplication and inversion are smooth maps.

It is in particular a topological group (smooth maps are continuous).

Example B.1.3. $GL(n, \mathbb{R})$ is a smooth n^2-manifold. Product is a polynomial map and inversion is a rational map so they are smooth maps. Thus it is a Lie group.

B.2 Vector fields and Lie algebra

Definition B.2.1. Let M be a smooth manifold and $p \in M$. A linear map $v : \mathcal{C}^\infty(M) \to \mathbb{R}$ is called a derivation at p if it satisfies

$$v(fg) = f(p)v(g) + g(p)v(f) \quad \text{for all } f, g \in \mathcal{C}^\infty(M).$$

The set of all derivations at p, denoted by T_pM is a vector space called tangent space to M at p. An element of T_pM is called a tangent vector at p. We also define the tangent bundle of M, denoted TM, which is the set of all tangent spaces at all points in M

$$TM = \coprod_{p \in M} T_pM.$$

(Note that the union is disjoint).

Definition B.2.2 (Vector field)**.** Let M be a smooth manifold. A vector field on M is a continuous map $X : M \to TM$, $p \mapsto X_p$ such that $X_p \in T_pM$ for each $p \in M$. It is smooth if for any $f \in \mathcal{C}^\infty(M)$, the map $M \to, p \mapsto X_p(f)$ is smooth. We denote $\text{Vect}^\infty(M)$ the set of all smooth vector fields on M.

A vector field should be visualized as we do for vector fields in an Euclidean space: a set of arrows attached to points of M and tangent to it. Let M, N smooth manifolds and $F : M \to N$ a smooth map. Then, we define the differential of F at $p \in M$ as the map $dF_p : T_pM \to T_{F(p)}N$ such that for any $v \in T_pM$, $f \in \mathcal{C}^\infty(N)$

$$dF_p(v)(f) = v(f \circ F).$$

The function $dF_p(v) : \mathcal{C}^\infty(N) \to \mathbb{R}$ is linear (because v is) and it is a derivation at $F(p)$. Let G a Lie group and $L_g : G \to G, h \mapsto gh$ the left translation on G. Then, a vector field X on G is said to be left-invariant if for any $g, g' \in G$

$$d(L_g)_{g'}(X_{g'}) = X_{gg'}.$$

Definition B.2.3 (Lie algebra)**.** A Lie algebra over a vector field K is a vector space V endowed with a bilinear map $V \times V \to V, (x, y) \mapsto [x, y]$ which is antisymmetric and satisfies the Jacobi identity, that is

1. (Antisymmetry) $\forall x, y \in V \ [x, y] + [y, x] = 0$,
2. (Jacobi identity) $\forall x, y, z \in V \ [x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0$.

A subset $W \subseteq V$ of a Lie algebra V is a Lie subalgebra of V if it is closed under brackets. In this case it is also a Lie algebra.

Proposition B.2.1.
- *$\text{Vect}^\infty(M)$ for M smooth manifold is a Lie algebra.*
- *The set of all left-invariant smooth vector fields over a Lie group G is a Lie subalgebra of $\text{Vect}^\infty(G)$. It is called the Lie algebra of G, denoted by Lie(G).*

One can prove that Lie(G) is finite dimensional and has the same dimension of G. The Lie algebra provides a "linear model" for the Lie group.

Theorem B.2.1. *Let G be a Lie group with identity element e. The evaluation map*

$$\varepsilon : \begin{array}{ll} Lie(G) & \to T_e \\ X & \mapsto X_e \end{array}$$

is a vector space isomorphism. Thus, Lie(G) is finite dimensional and has dimension dim(G).

Appendix C

Tools

C.1 Controlled differential equations

Theorem C.1.1 (Picard's theorem)**.** *Let V and W Banach spaces, $\mathcal{L}(V,W)$ the set of linear mappings from V to W. We consider $X : [0,1] \to V$ of finite p-variation with $1 \leq p < 2$, $Y : [0,1] \to W$ and $f : W \to \mathcal{L}(V,W)$ Lipschitz(γ) with $p < \gamma$. The for every $\xi \in W$, the controlled differential equation*

$$dY_t = f(Y_t)dX_t \qquad Y_0 = \xi$$

has a unique solution. The solution mapping $I_f : BV^p(V) \times W \to BV^p(W)$ is continuous (with $I_f(X,\xi)$ the solution).

C.2 Stone - Weierstrass theorem

Theorem C.2.1 (Stone-Weierstrass)**.** *Suppose D is a compact Hausdorff space, $C(D,)$ the set of continuous real-valued functions on D and A is a sub-algebra of $C(D,)$ which contains a non-zero constant function. Then A is dense in $C(D,)$ if and only if it separates points (i.e. $\forall x, y \in D \quad \exists f \in A$ s.t. $f(x) \neq f(y)$).*

C.3 Least squares rate of convergence

We state here the main theorems from [13] used to prove theorem 3.2.1. We refer the reader to [13] for the proofs. We consider a random vector (X,Y) where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ and we want to find a measurable function f such that $|f(X) - Y|$ is "small" in some sense. We let $f* = \mathbb{E}[Y|X = x]$ our target function. We denote μ the density of X.

Proposition C.3.1. *Assume $\mathcal{F}$ is a vector space of functions from $\mathbb{R}^d \to \mathbb{R}$ of dimension k finite. Then if $f_1, ..., f_k$ is a basis of $\mathcal{F}$, the least squares estimator*

$$\hat{f}_n = \text{argmin}_{f \in \mathcal{F}} \mathcal{R}_n(f)$$

can be written as $\hat{f}_n = \sum_{j=1}^k \hat{\beta}_j f_j$ with $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_k)$ solution of

$$B^T B\hat{\beta} = B^T Y, \tag{C.1}$$

with $B = (f_j(X_i))_{i=1,...,n, j=1,...,k}$.

Proof. Any $f \in \mathcal{F}$ can be rewritten as $f = \sum_{j=1}^{k} \beta_j f_j$ for some $\beta = (\beta_1, ..., \beta_k) \in \mathbb{R}^k$. Then

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{k} a_j f_j(X_i))^2 = \frac{1}{n}\|Y - B\beta\|_2^2$$

with $\|.\|$ the Euclidean norm in $\mathbb{R}^k$. We are thus looking for

$$\hat{\beta} = \text{argmin}_{\beta \in R^k}\|Y - B\beta\|_2^2$$

This is the classical regression problem, and equation C.1 is well known. □

Theorem C.3.1. *Assume*

$$\sigma^2 = \sup_{x \in \mathbb{R}^d} Var(Y|X = x) < \infty$$

Let $\hat{f}_n = \text{argmin}_{f \in \mathcal{F}}\mathcal{R}_n(f)$ *with* $\mathcal{F}$ *a linear vector space of functions* $f : \mathbb{R}^d \to \mathbb{R}$, k *its dimension,* $f^*(x) = \mathbb{E}[Y|X = x]$ *and for any function* f,

$$\|f\|_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)^2.$$

Then

$$\mathbb{E}\left[\|f^* - \hat{f}_n\|_n^2 | X_1, ..., X_n\right] \leq \sigma^2\frac{k}{n} + \min_{f \in \mathcal{F}}\|f^* - f\|. \tag{C.2}$$

Theorem C.3.2. *Assume*

$$\sigma^2 = \sup_{x \in \mathbb{R}^d} Var(Y|X = x) < \infty$$

and

$$\|f^*\|_\infty = sup_{x \in \mathbb{R}^d}|f(x|) \leq L$$

for some $L > 0$. *Let* $\mathcal{F}$ *be a linear space of functions* $f : \mathbb{R}^d \to R$ *of dimension* k. *We define* $\hat{f}_n = arg\ min_{f \in \mathcal{F}}\mathcal{R}_n(f)$ *and* $\hat{f}_n^L$ *such that for any* $x \in \mathbb{R}^d$

$$\hat{f}_n^L(x) = \begin{cases} \hat{f}_n(x) \text{ if } |x| \leq L \\ L\text{sign}(x) \text{ else} \end{cases}$$

Then, there exists a constant c *such that*

$$\begin{aligned}\mathbb{E}\left(\int(\hat{f}_n^L(x) - f^*(x))^2\mu(dx)\right) \leq & c\max(\sigma^2, L^2)\frac{(\log(n)+1)k}{n} \\ & + 8\inf_{f \in \mathcal{F}}\int(f(x) - f^*(x))^2\mu(dx).\end{aligned} \tag{C.3}$$

Bibliography

[1] Donald J Berndt. Finding patterns in time series: a dynamic programming approach. *Advances in knowledge discovery and data mining*, 1996.

[2] Horatio Boedihardjo and Xi Geng. The uniqueness of signature problem in the non-markov setting. *Stochastic Processes and their Applications*, 125(12):4674–4701, 2015.

[3] Horatio Boedihardjo, Xi Geng, Terry Lyons, and Danyu Yang. The signature of a rough path: uniqueness. *Advances in Mathematics*, 293:720–737, 2016.

[4] Kuo-sai Chen. Integration of paths—a faithful representation of paths by noncommutative formal power series. *Transactions of the American Mathematical Society*, 89(2):395–407, 1958.

[5] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.

[6] Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.

[7] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml.

[8] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[9] Terry Lyons et al. Coropa computational rough paths (software library), 2010. http://coropa.sourceforge.net/.

[10] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[11] Peter K Friz and Nicolas B Victoir. *Multidimensional stochastic processes as rough paths: theory and applications*, volume 120. Cambridge University Press, 2010.

[12] Benjamin Graham. Sparse arrays of signatures for online character recognition. *arXiv preprint arXiv:1308.0371*, 2013.

[13] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[14] Ben Hambly and Terry Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, pages 109–167, 2010.

[15] Alessandra Iozzi and Robert Zimmer. Introduction to lie groups. 2016.

[16] Franz J Király and Harald Oberhauser. Kernels for sequentially ordered data. *arXiv preprint arXiv:1601.08169*, 2016.

[17] AB Kormilitzin, KEA Saunders, PJ Harrison, JR Geddes, and TJ Lyons. Application of the signature method to pattern recognition in the cequel clinical trial. *arXiv preprint arXiv:1606.02074*, 2016.

[18] Mineichi Kudo, Jun Toyama, and Masaru Shimbo. Multidimensional curve classification using passing-through regions. *Pattern Recognition Letters*, 20(11-13):1103–1111, 1999.

[19] Yves Le Jan and Zhongmin Qian. Stratonovich's signatures of brownian motion determine brownian sample paths. *Probability Theory and Related Fields*, 157(1-2):209–223, 2013.

[20] John M Lee. *Introduction to Smooth Manifolds*. Springer-Verlag New York, 2012.

[21] Daniel Levin, Terry Lyons, and Hao Ni. Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*, 2013.

[22] Terry Lyons, Mark Kontkowski, Jonathan Field, et al. Extracting information from the signature of a financial data stream. Technical report, arXiv. org, 2014.

[23] Terry J Lyons, Michael Caruana, and Thierry Lévy. *Differential equations driven by rough paths*. Springer, 2007.

[24] Anastasia Papavasiliou, Christophe Ladroue, et al. Parameter estimation for rough differential equations. *The Annals of Statistics*, 39(4):2047–2073, 2011.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] K Purbhoo. Notes on tensor products and the exterior algebra. 2012.

[27] Chotirat Ann Ralanamahatana, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos, and Gautam Das. Mining time series data. In *Data mining and knowledge discovery handbook*, pages 1069–1103. Springer, 2005.

[28] James O Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

[29] James O Ramsay and CJ Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 539–572, 1991.

[30] Jeremy Reizenstein. Calculation of iterated-integral signatures and log signatures. *arXiv preprint arXiv:1712.02757*, 2017.

[31] Jeremy Reizenstein and Benjamin Graham. The iisignature library: efficient calculation of iterated-integral signatures and log signatures. *arXiv preprint arXiv:1802.08252*, 2018.

[32] Christophe Reutenauer. Free lie algebras. In *Handbook of algebra*, volume 3, pages 887–903. Elsevier, 2003.

[33] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

[34] 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.

[35] Weixin Yang, Terry Lyons, Hao Ni, Cordelia Schmid, Lianwen Jin, and Jiawei Chang. Leveraging the path signature for skeleton-based human action recognition. *arXiv preprint arXiv:1707.03993*, 2017.

[36] Laurence C Young. An inequality of the hölder type, connected with stieltjes integration. *Acta Mathematica*, 67(1):251, 1936.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

SIGNATURE AND STATISTICAL LEARNING

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
FERHANIAN	Adeline

With my signature I confirm that

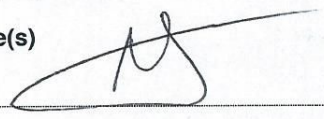
- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date

Paris, 26/10/2018

Signature(s)



For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.