Rough paths, signature and statistical learning

Adeline Fermanian [1]
Supervisors: Gérard Biau [1] and Benoît Cadre [2]

[1]LPSM, Sorbonne Université

[2]IRMAR, ENS Rennes

51[es] Journées de Statistique, Nancy, June 2019.

Table of Contents

Table of Contents

Learning from a data stream.

We are interested in predicting from a data stream, for example:

Learning from a data stream.

We are interested in predicting from a data stream, for example:

- Time series prediction,
- Online recognition of characters or handwriting,
- Sound recognition,
- Automated medical diagnosis from sensor data,

  ...

Learning from a data stream.

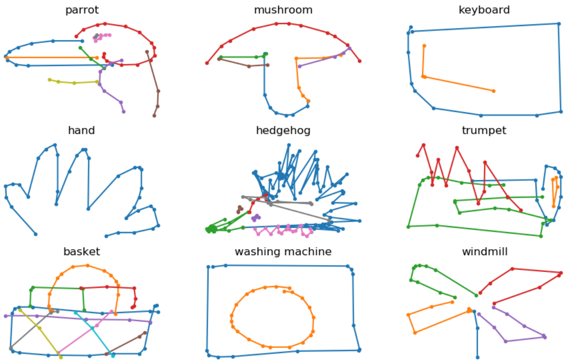
We are interested in predicting from a data stream, for example:

- Time series prediction,
- Online recognition of characters or handwriting,
- Sound recognition,
- Automated medical diagnosis from sensor data,

  ...

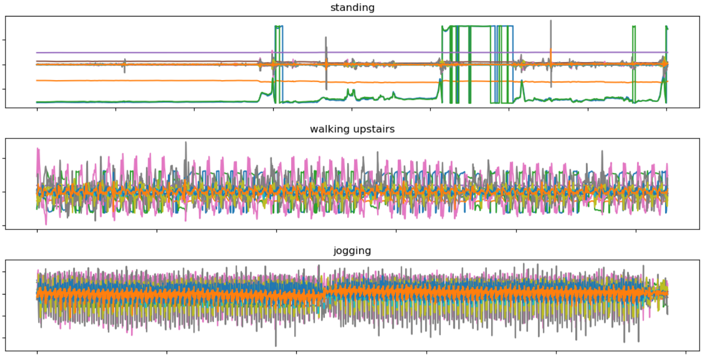
In all of these cases, the predictor can be seen as a path $X : [a, b] \to \mathbb{R}^d$.

Google "Quick, Draw!" dataset



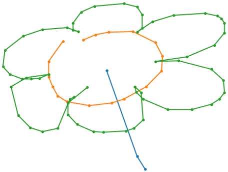
→ 50 million drawings, 430 classes.

Motion Sense dataset



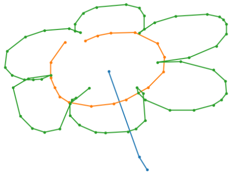
$\rightarrow$ Smartphone sensory data: time series in $\mathbb{R}^{12}$. 300 samples, 6 classes.

How can we represent such data?

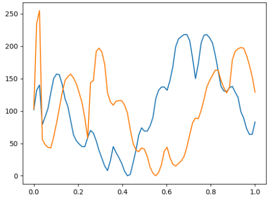


(a) Sample from the class "flower".

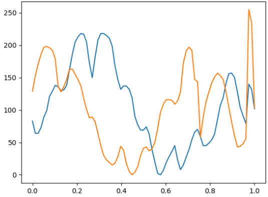
How can we represent such data?



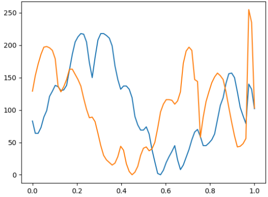
(a) Sample from the class "flower".



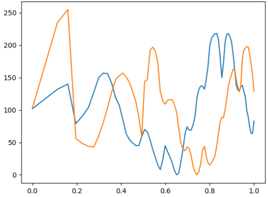
(b) x and y coordinates.



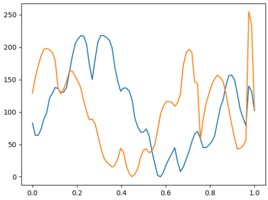
(a) Time reversed x and y coordinates.



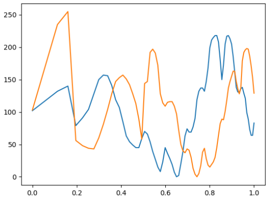
(a) Time reversed x and y coordinates.



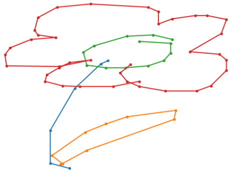
(b) x and y coordinates of the flower drawn at a different speed.



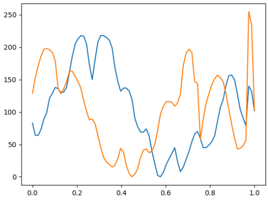
(a) Time reversed x and y coordinates.



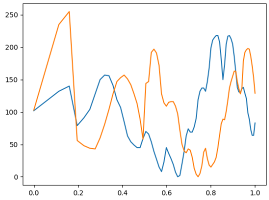
(b) x and y coordinates of the flower drawn at a different speed.



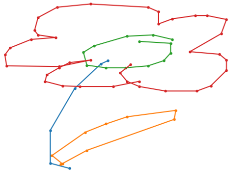
(c) Another sample from the class "flower".



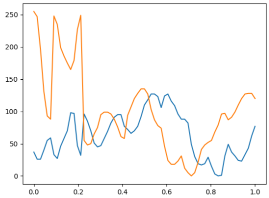
(a) Time reversed x and y coordinates.



(b) x and y coordinates of the flower drawn at a different speed.



(c) Another sample from the class "flower".



(d) x and y coordinates of the second flower drawing.

The signature will overcome some of these problems:

- It is a transformation from a path to a sequence of coefficients of unique length, regardless of the length of the initial time sequence.
- Independent of time parametrization.
- Encodes geometric properties of the path.
- Unique (under some assumptions).

Origin of signature ? Chen in the 60s, then Lyons' rough paths theory in the 90s.

Table of Contents

Setting

- X is a continuous path $X : [0, 1] \to \mathbb{R}^d$.
- $\|X\|_{\text{1-var}}$ is its total variation, assumed to be finite.
- One can then define the Riemann-Stieljes integral $\int_0^1 Y_t dX_t$, where Y is a continuous path.

If X is differentiable, it is just $\int_0^1 Y_t dX_t = \int_0^1 Y_t \dot{X}_t dt$.

First terms of the signature

Let $d = 2$, and $X_t = (X_t^1, X_t^2)$, then

$$\begin{pmatrix} S^{(1)}(X) \\ S^{(2)}(X) \end{pmatrix} = \begin{pmatrix} \int_0^1 dX_t^1 \\ \int_0^1 dX_t^2 \end{pmatrix} = \int_0^1 dX_t,$$

First terms of the signature

Let $d = 2$, and $X_t = (X_t^1, X_t^2)$, then

$$\begin{pmatrix} S^{(1)}(X) \\ S^{(2)}(X) \end{pmatrix} = \begin{pmatrix} \int_0^1 dX_t^1 \\ \int_0^1 dX_t^2 \end{pmatrix} = \int_0^1 dX_t,$$

$$\begin{pmatrix} S^{(1,1)}(X) & S^{(1,2)}(X) \\ S^{(2,1)}(X) & S^{(2,2)}(X) \end{pmatrix} = \begin{pmatrix} \iint\limits_{0 \le s < t \le 1} dX_s^1 dX_t^1 & \iint\limits_{0 \le s < t \le 1} dX_s^1 dX_t^2 \\ \iint\limits_{0 \le s < t \le 1} dX_s^2 dX_t^1 & \iint\limits_{0 \le s < t \le 1} dX_s^2 dX_t^2 \end{pmatrix}$$

First terms of the signature

Let $d = 2$, and $X_t = (X_t^1, X_t^2)$, then

$$\begin{pmatrix} S^{(1)}(X) \\ S^{(2)}(X) \end{pmatrix} = \begin{pmatrix} \int_0^1 dX_t^1 \\ \int_0^1 dX_t^2 \end{pmatrix} = \int_0^1 dX_t,$$

$$\begin{pmatrix} S^{(1,1)}(X) & S^{(1,2)}(X) \\ S^{(2,1)}(X) & S^{(2,2)}(X) \end{pmatrix} = \begin{pmatrix} \iint\limits_{0\le s<t\le 1} dX_s^1 dX_t^1 & \iint\limits_{0\le s<t\le 1} dX_s^1 dX_t^2 \\ \iint\limits_{0\le s<t\le 1} dX_s^2 dX_t^1 & \iint\limits_{0\le s<t\le 1} dX_s^2 dX_t^2 \end{pmatrix}$$

$$= \iint\limits_{0\le s<t\le 1} dX_s \otimes dX_t.$$

First terms of the signature

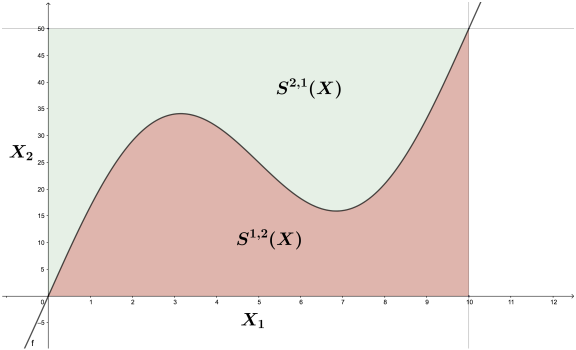
$$\boxed{\begin{array}{cc} S^{(1,1,1)} & S^{(1,2,1)} \\ S^{(2,1,1)} & S^{(2,2,1)} \end{array}} \quad \boxed{\begin{array}{cc} S^{(1,1,2)} & S^{(1,2,2)} \\ S^{(2,1,2)} & S^{(2,2,2)} \end{array}} = \iint_{0 \leq s < t < u \leq 1} dX_s \otimes dX_t \otimes dX_u .$$

For example,

$$S^{(1,2,2)} = \iint_{0 \leq s < t < u \leq 1} dX_s^1 dX_t^2 dX_u^2 .$$

Geometric interpretation

The signature contains geometric properties of the path.



Signature of a path

Definition

Let $X : [0,1] \to \mathbb{R}^d$ be a path of bounded variation and $I = (i_1, \ldots, i_k) \subset \{1, \ldots, d\}^k$ be a multi index. The signature coefficient corresponding to I is

$$S^{(i_1,\ldots,i_k)}(X) = \int \cdots \int_{0 \le u_1 < \cdots < u_k \le 1} dX^{i_1}_{u_1} \ldots dX^{i_k}_{u_k}.$$

Signature of a path

Definition

The signature of X is the vector containing all signature coefficients:

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

The signature of X truncated at order m is:

$$S^m(X) = \Big(1, S^{(1)}(X), S^{(2)}(X), \ldots, S^{\overbrace{(d, \ldots, d)}^{m}}(X)\Big).$$

Its dimension is $\sum_{k=0}^{m} d^k = \frac{d^{m+1}-1}{d-1}$.

Example

Example (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 $I = (i_1, \cdots, i_k) \in \{1, \cdots, d\}^k$,

$$S^{(i_1,\ldots,i_k)}(X) = \frac{1}{k!} \prod_{j=1}^{k} (X_1 - X_0)^{i_j}.$$

Properties (1)

Proposition (Invariance under time reparametrization)

Let $X : [0,1] \to \mathbb{R}^d$ be a path and $\psi : [0,1] \to [0,1]$ a reparametrization. Then, if $\tilde{X}_t = X_{\psi(t)}$,

$$S(\tilde{X}) = S(X).$$

Proposition (Uniqueness)

If X has at least one monotonous coordinate, then $S(X)$ determines X uniquely.

Properties (2)

Proposition (Signature approximation)

Let D be a compact subset of the space of paths from $[0,1]$ to $\mathbb{R}^d$ of bounded variation. Let $f : D \to \mathbb{R}$ continuous. Then, for every $\epsilon > 0$, there exists $N \in \mathbb{N}$, $w \in \mathbb{R}^N$ such that, for any $X \in D$,

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

Properties (3)

Proposition (Exponential decay of signature coefficients)

Let $X : [0,1] \to \mathbb{R}^d$ *be a path of bounded variation. Then, for any* $k \geq 0$, $I \subset \{1, \dots d\}^k$,

$$\left|S^I(X)\right| \leq \frac{\|X\|_{1-var}^k}{k!}.$$

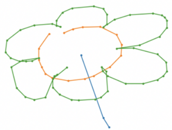
Table of Contents

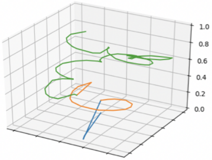
Procedure

Data $\longrightarrow$ Path embedding $\longrightarrow$

$$\Big[[(x_1^1, y_1^1), \ldots, (x_{p_1}^1, y_{p_1}^1)],\\ [(x_1^2, y_1^2), \ldots, (x_{p_2}^2, y_{p_2}^2)],\\ \ldots\\ [(x_1^K, y_1^K), \ldots, (x_{p_K}^K, y_{p_K}^K)]\Big]$$

$$X : [0,1] \to \mathbb{R}^3$$





Procedure

Signature transformation

$$S^m(X) = \begin{pmatrix} S^1(X) \\ \vdots \\ S^d(X) \\ S^{11}(X) \\ S^{12}(X) \\ \vdots \\ S^{d\ldots d}(X) \end{pmatrix} \in \mathbb{R}^{\frac{d^{m+1}-1}{d-1}}$$

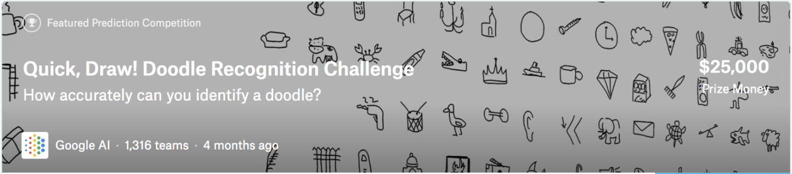
$$\begin{pmatrix} -19 \\ 26 \\ \vdots \\ 2929.3 \end{pmatrix}$$

Prediction with any learning algorithm

$$Y = \hat{f}(S^m(X))$$

« Flower »

Kaggle prediction competition



Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Late Submission**

Overview

Description

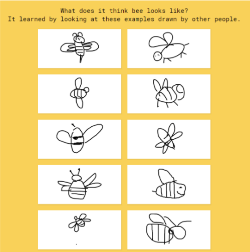
Evaluation

Prizes

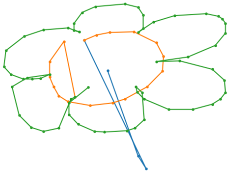
Timeline

"Quick, Draw!" was released as an experimental game to educate the public in a playful way about how AI works. The game prompts users to draw an image depicting a certain category, such as "banana," "table," etc. The game generated more than 1B drawings, of which a subset was publicly released as the basis for this competition's training set. That subset contains 50M drawings encompassing 340 label categories.

Sounds fun, right? Here's the challenge: since the training data comes from the game itself, drawings can be incomplete or may not match the label. You'll need to build a recognizer that can effectively learn from this noisy data and perform well on a manually-labeled test set from a different distribution.

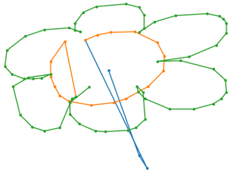


Data embedding as a path

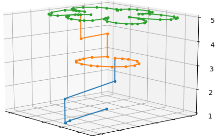


(a) 2-dimensional path.

Data embedding as a path

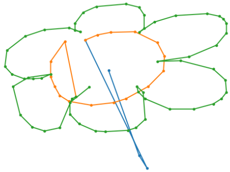


(a) 2-dimensional path.

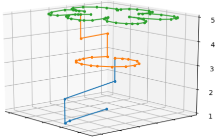


(b) 3-dimensional ”stroke path”.

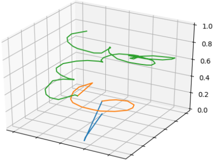
Data embedding as a path



(a) 2-dimensional path.

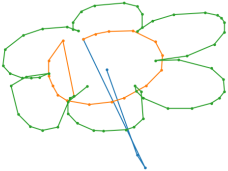


(b) 3-dimensional "stroke path".

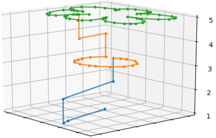


(c) 3-dimensional "time path".

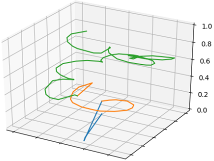
Data embedding as a path



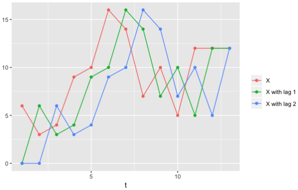
(a) 2-dimensional path.



(b) 3-dimensional "stroke path".



(c) 3-dimensional "time path".



(d) Lead-lag transformation.

Results: Quick, Draw! dataset

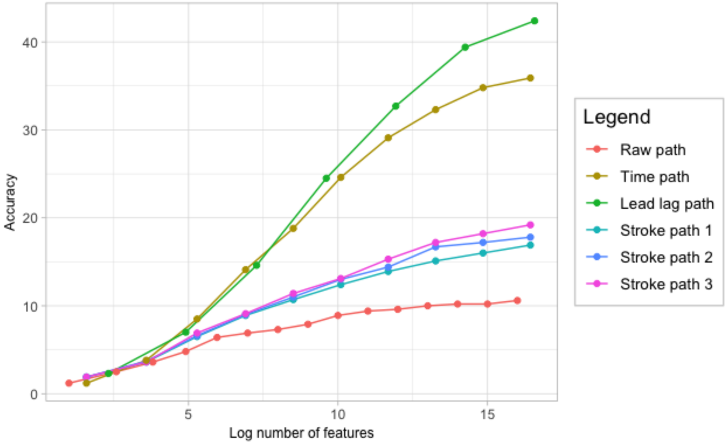


Figure: Prediction accuracy on the "Quick, Draw!" dataset with a linear neural network with one hidden layer.

Results: Motion sense dataset

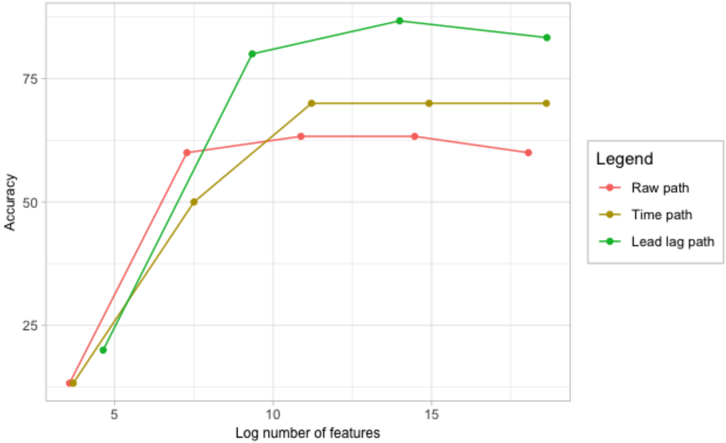


Figure: Prediction accuracy on the "Motions Sense" dataset with a linear neural network with one hidden layer.

Results: XGBoost algorithm

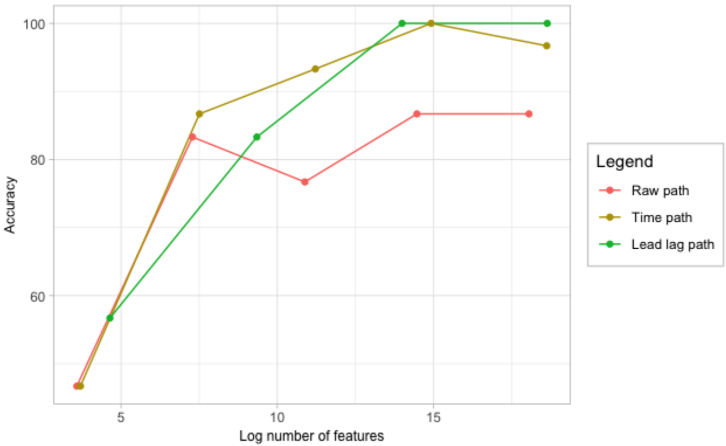


Figure: Prediction accuracy on the "Motions Sense" dataset with XGBoost algorithm.

Linear regression on the signature

Model: Let $(X_t)_{t\in[0,1]}$ be a stochastic process of bounded variation. We assume that there exists $m^* \in \mathbb{N}$, $\beta^* \in \mathbb{R}^{s_d(m^*)}$ such that

$$Y = \langle \beta^*, S^{m^*}(X) \rangle + \varepsilon,$$

where

$$\mathbb{E}[\varepsilon | X] = 0 \quad \text{and} \quad \mathrm{Var}(\varepsilon | X) = \sigma^2 < \infty.$$

Goal: Estimate β^* and m^*.

Conclusion

- The signature is a generic method that can be used for multidimensional sequential data.
- It encodes, in a fixed number of coefficients, geometric properties of the input path and it linearizes complex functions of the path.
- Data embedding has a huge influence on prediction performance.
- Further work:
  - Theoretical investigation of embedding properties.
  - Truncation order selection in regression models.
  - Extension to highly oscillatory paths.

Thank you for your attention. Questions?