A statistical point of view on signatures

Conference Pathwise Stochastic Analysis and Applications

CIRM, Marseille

Adeline Fermanian

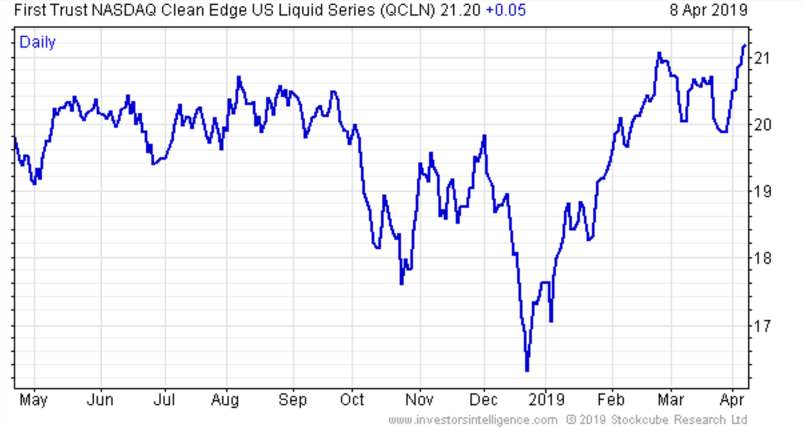
March 12th 2021

Joint work with

Benoît Cadre
University Rennes 2

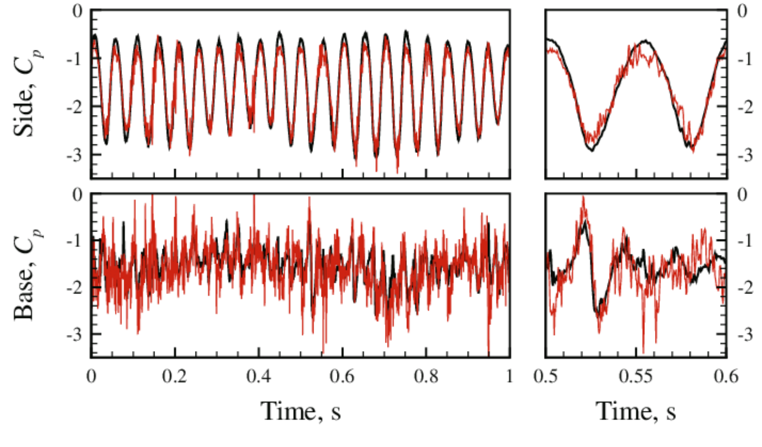
Gérard Biau
Sorbonne University

Learning from a data stream



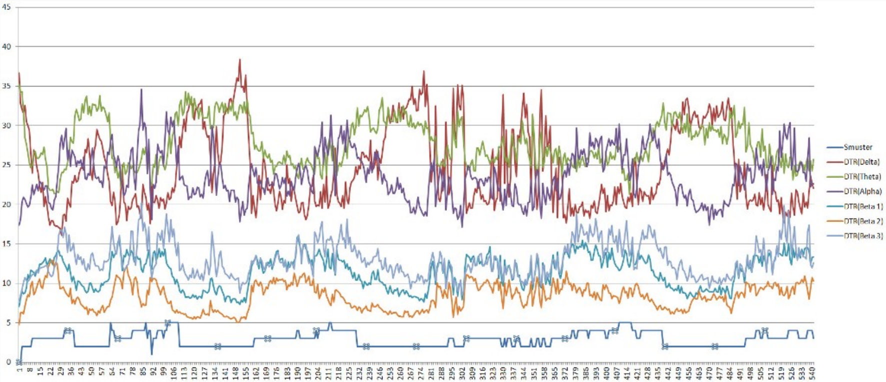
Time series prediction

Learning from a data stream



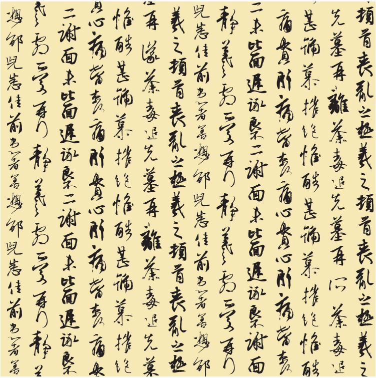
Stereo sound recognition

Learning from a data stream



Automated medical diagnosis from sensor data

Learning from a data stream

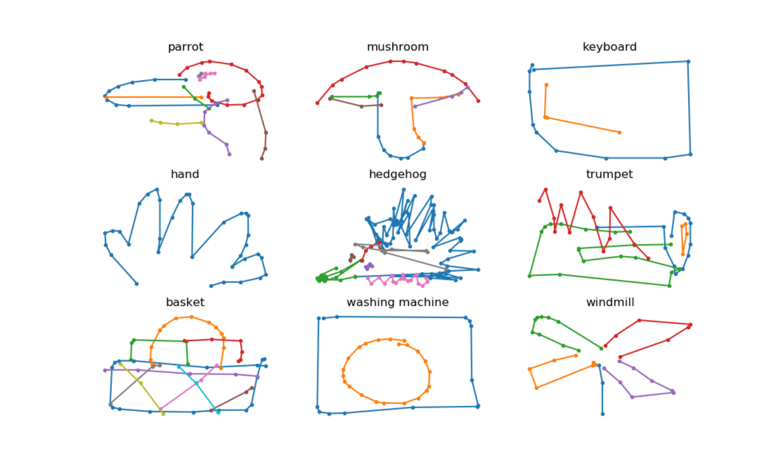


Recognition of characters or handwriting

Common feature

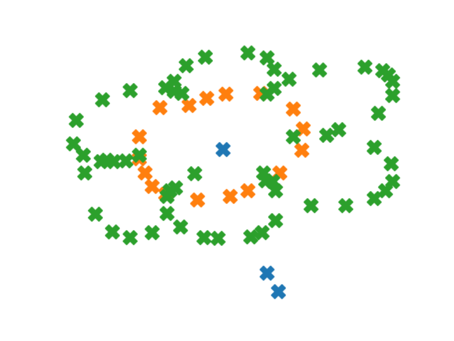
The predictor is a path $X : [a, b] \to \mathbb{R}^d$.

Google "Quick, Draw!" dataset



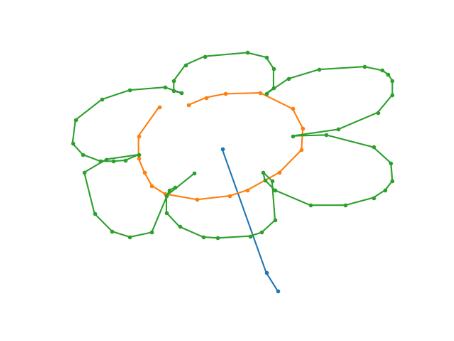
50 million drawings, 340 classes

Data representation



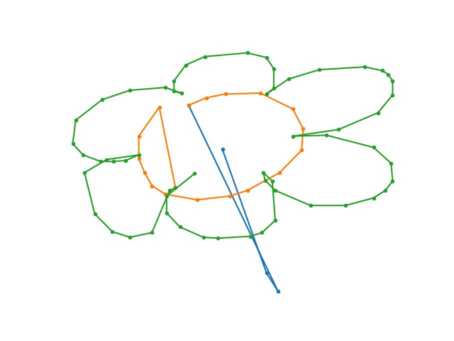
A sample from the class flower

Data representation



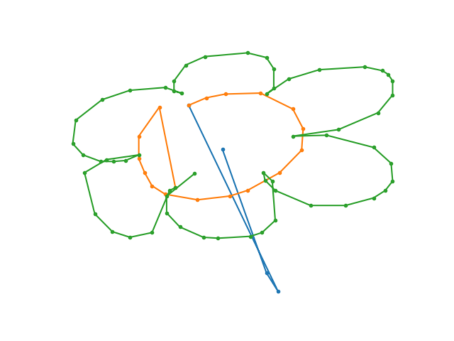
A sample from the class flower

Data representation

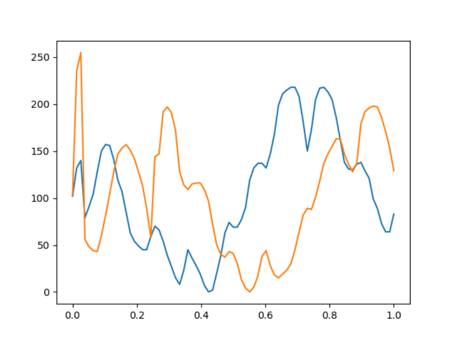


A sample from the class flower

Data representation

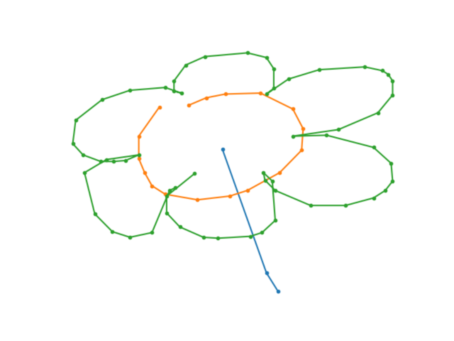


A sample from the class flower

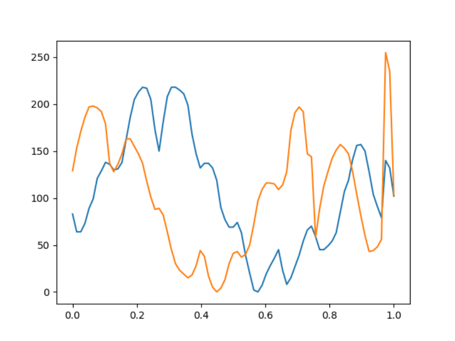


x and y coordinates

Data representation

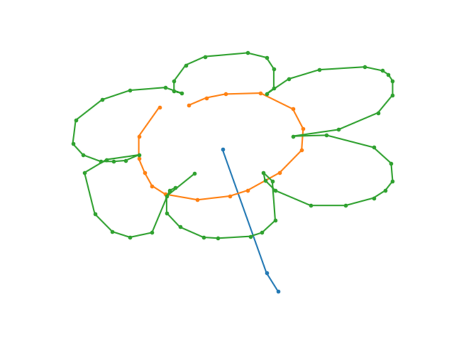


A sample from the class flower

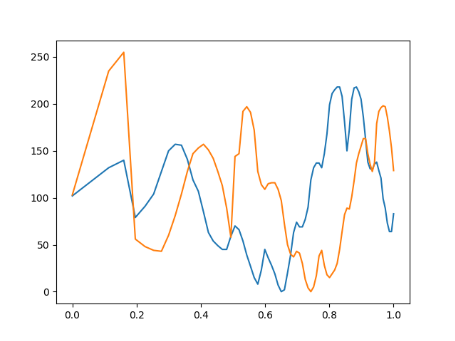


Time reversed

Data representation



A sample from the class flower



x and y at a different speed

The signature will overcome some of these problems.

The signature will overcome some of these problems.

- It is a transformation from a path to a sequence of coefficients.

The signature will overcome some of these problems.

- It is a transformation from a path to a sequence of coefficients.
- Independent of time parameterization.

The signature will overcome some of these problems.

- It is a transformation from a path to a sequence of coefficients.
- Independent of time parameterization.
- Encodes geometric properties of the path.

The signature will overcome some of these problems.

- It is a transformation from a path to a sequence of coefficients.
- Independent of time parameterization.
- Encodes geometric properties of the path.
- No loss of information.

Table of contents

Definition and basic properties

A brief history

Chen's work for piecewise smooth paths.

ANNALS OF MATHEMATICS
Vol. 65, No. 1, January, 1957
Printed in U.S.A.

INTEGRATION OF PATHS, GEOMETRIC INVARIANTS AND A GENERALIZED BAKER-HAUSDORFF FORMULA

BY KUO-TSAI CHEN

(Received October 17, 1955)

(Revised May 28, 1956)

Let $\alpha: \langle \alpha_1(t), \cdots, \alpha_m(t) \rangle$, $a \leqq t \leqq b$, be a path in the affine m-space R^m. Starting from the line integral $\int_\alpha dx_i$, we define inductively, for $p \geqq 2$,

$$\int_\alpha dx_{i_1} \cdots dx_{i_p} = \int_a^b \left(\int_{\alpha^t} dx_{i_1} \cdots dx_{i_{p-1}} \right) d\alpha_{i_p}(t),$$

where α^t denotes the portion of α with the parameter ranging from a to t. It is observed that $\int_\alpha dx_{i_1} \cdots dx_{i_p}$ acts as a p^{th} order contravariant tensor associated with the path α when R^m undergoes a linear transformation. Some affine and euclidean invariants of α are derived from these tensors. Moreover, we associate to the path α the formal power series

$$\theta(\alpha) = 1 + \sum_{p=1}^{\infty} \sum \left(\int_\alpha dx_{i_1} \cdots dx_{i_p} \right) X_{i_1} \cdots X_{i_p}$$

where $X_1, \cdots, X_m$ are noncommutative indeterminates. Theorem 4.2 asserts that $\log \theta(\alpha)$ is a Lie element, i.e., a formal power series $u_1 + \cdots + u_p + \cdots$, where each u_p is a form of degree p generated by $X_1, \cdots, X_m$ through taking bracket products and forming linear combinations. We obtain, as a corollary, the Baker-Hausdorff formula which states that, if X and Y are noncommutative indeterminates, then $\log (\exp X \cdot \exp Y)$ is a Lie element.

Section 1 supplies first some basic knowledge about non-commutative formal power series and then some preparatory definitions and formulas for Theorems 4.1 and 4.2. In Section 2, the iterated integration of paths is defined; and, in Section 3, its geometric applications are indicated. Section 4 contains mainly the proof of the generalized Baker-Hausdorff formula which is further extended, in Section 5, to the case where the affine space R^m is replaced by a differentiable mainfold. For those who are only interested in the geometric aspect of this paper, Sections 2 and 3 may be easily read without Section 1.

This paper is a continuation of the author's work in [Chen, (3)] and is somewhat related to the paper [Chen, (2)]. The proof of Lemma 1.2 is essentially Hausdorff's, in which Lemma 1.1 is implicitly used. Its proof, not an obvious one, is furnished in this paper. Though borrowing some of Hausdorff's technique, Theorem 4.2 is proved in a simpler way and offers a stronger result than the Baker-Hausdorff formula.

A brief history

Lyons' extension to rough paths.

A brief history

Machine learning applications are ↗.

DeepWriterID: An End-to-end Online Text-independent Writer Identification System

Weixin Yang, Lianwen Jin*, Manfei Liu
College of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
wxy1290@163.com, *lianwen.jin@gmail.com

***Abstract*—Owing to the rapid growth of touchscreen mobile terminals and pen-based interfaces, handwriting-based writer identification systems are attracting increasing attention for personal authentication and digital forensics. However, most studies on writer identification have not been satisfying because of the insufficiency of data and the difficulty of designing good features for various conditions of handwriting samples. Hence, we introduce an end-to-end system called DeepWriterID that employs a deep convolutional neural network (CNN) to address these problems. A key feature of DeepWriterID is a new method we are proposing, called DropSegment. It is designed to achieve data augmentation and to improve the generalized applicability of CNN. For sufficient feature representation, we further introduce path-signature feature maps to improve performance. Experiments were conducted on the NLPR handwriting database. Even though we only use pen-position information in the pen-down state of the given handwriting samples, we achieved new state-of-the-art identification rates of 95.72% for Chinese text and 98.51% for English text.**

***Keywords*—Online text-independent writer identification; convolutional neural network; deep learning; DropSegment; path-signature feature maps.**

1. Introduction

Writer identification is a task of determining a list of candidate writers according to the degree of similarity between their handwriting and a sample of unknown authorship [1]. Currently, it is popular owing to the development and commercialization of touchscreen or pen-enabled electronic devices such as smartphones, and tablet PCs. Its wide range of downstream uses include distinguishing forensic trace evidence, performing mobile bank transactions, and authenticating access to networks. Since most of these applications are closely related to the purpose of assuring personal and property security, handwriting identification merits more attention from academia and industry.

Identifying the handwriting of a writer is one of the highly challenging problems in the fields of artificial intelligence and pattern recognition. Conventionally, handwriting identification systems follow a sequence of data acquisition, data preprocessing, feature extraction, and classification [2]. Research into handwriting identification has been focused on two categories: offline and online. Offline handwritten materials are considered more general but harder to identify, as they contain merely scanned image information. In contrast, systems

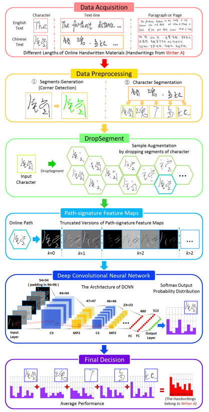


Figure 1. Illustration of DeepWriterID for online handwriting-based writer identification.

Mathematical setting

- A path $X : [0,1] \to \mathbb{R}^d$. Notation: X_t.

Mathematical setting

- A path $X : [0, 1] \to \mathbb{R}^d$. Notation: X_t.
- Assumption: $\|X\|_{\text{1-var}} < \infty$.

Mathematical setting

- A path $X: [0,1] \to \mathbb{R}^d$. Notation: X_t.
- Assumption: $\|X\|_{1\text{-var}} < \infty$.
- $Y: [0,1] \to \mathbb{R}$ a continuous path.

Mathematical setting

- A path $X : [0,1] \to \mathbb{R}^d$. Notation: X_t.
- Assumption: $\|X\|_{\text{1-var}} < \infty$.
- $Y : [0,1] \to \mathbb{R}$ a continuous path.
- Riemann-Stieljes integral of Y against X is well-defined. Notation:

$$\int_0^1 Y_t dX_t.$$

Mathematical setting

- A path $X : [0,1] \to \mathbb{R}^d$. Notation: X_t.
- Assumption: $\|X\|_{1\text{-var}} < \infty$.
- $Y : [0,1] \to \mathbb{R}$ a continuous path.
- Riemann-Stieljes integral of Y against X is well-defined. Notation:

$$\int_0^1 Y_t dX_t.$$

Example :

- X_t continuously differentiable:

$$\int_0^1 Y_t dX_t = \int_0^1 Y_t \dot{X}_t dt$$

Mathematical setting

- A path $X : [0,1] \to \mathbb{R}^d$. Notation: X_t.
- Assumption: $\|X\|_{1\text{-var}} < \infty$.
- $Y : [0,1] \to \mathbb{R}$ a continuous path.
- Riemann-Stieljes integral of Y against X is well-defined. Notation:

$$\int_0^1 Y_t dX_t.$$

Example :

- $Y_t = 1$ for all $t \in [0,1]$:

$$\int_0^1 Y_t dX_t = \int_0^1 dX_t = X_1 - X_0.$$

Iterated integrals

- $X : [0,1] \to \mathbb{R}^d$, $X = (X^1, \ldots, X^d)$.

Iterated integrals

- $X : [0, 1] \to \mathbb{R}^d$, $X = (X^1, \dots, X^d)$.
- For $i \in \{1, \dots, d\}$,

$$S^{(i)}(X)_{[0,t]} = \int_{0<s<t} dX_s^i = X_t^i - X_0^i$$

Iterated integrals

- $X : [0,1] \to \mathbb{R}^d$, $X = (X^1, \dots, X^d)$.
- For $i \in \{1, \dots, d\}$,

$$S^{(i)}(X)_{[0,t]} = \int_{0<s<t} dX_s^i = X_t^i - X_0^i \quad \to \text{a path!}$$

Iterated integrals

- $X : [0,1] \to \mathbb{R}^d$, $X = (X^1, \ldots, X^d)$.
- For $i \in \{1, \ldots, d\}$,

$$S^{(i)}(X)_{[0,t]} = \int_{0<s<t} dX_s^i = X_t^i - X_0^i \quad \to \text{a path!}$$

- For $(i,j) \in \{1, \ldots, d\}^2$,

$$S^{(i,j)}(X)_{[0,t]} = \int_{0<s<t} S^{(i)}(X)_{[0,s]} dX_s^j = \int_{0<r<s<t} dX_r^i dX_s^j$$

Iterated integrals

- $X : [0,1] \to \mathbb{R}^d$, $X = (X^1, \dots, X^d)$.
- For $i \in \{1, \dots, d\}$,

$$S^{(i)}(X)_{[0,t]} = \int_{0<s<t} dX_s^i = X_t^i - X_0^i \quad \to \text{a path!}$$

- For $(i, j) \in \{1, \dots, d\}^2$,

$$S^{(i,j)}(X)_{[0,t]} = \int_{0<s<t} S^{(i)}(X)_{[0,s]} dX_s^j = \int_{0<r<s<t} dX_r^i dX_s^j \quad \to \text{a path!}$$

Iterated integrals

- $X : [0,1] \to \mathbb{R}^d$, $X = (X^1, \dots, X^d)$.
- For $i \in \{1, \dots, d\}$,

$$S^{(i)}(X)_{[0,t]} = \int_{0<s<t} dX_s^i = X_t^i - X_0^i \quad \to \text{a path!}$$

- For $(i,j) \in \{1, \dots, d\}^2$,

$$S^{(i,j)}(X)_{[0,t]} = \int_{0<s<t} S^{(i)}(X)_{[0,s]} dX_s^j = \int_{0<r<s<t} dX_r^i dX_s^j \quad \to \text{a path!}$$

- Recursively, for $(i_1, \dots, i_k) \in \{1, \dots, d\}^k$,

$$S^{(i_1, \dots, i_k)}(X)_{[0,t]} = \int_{0<t_1<t_2<\cdots<t_k<t} dX_{t_1}^{i_1} \dots dX_{t_k}^{i_k}.$$

Iterated integrals

- $X : [0,1] \to \mathbb{R}^d$, $X = (X^1, \dots, X^d)$.
- For $i \in \{1, \dots, d\}$,

$$S^{(i)}(X)_{[0,t]} = \int_{0<s<t} dX^i_s = X^i_t - X^i_0 \quad \to \text{a path!}$$

- For $(i,j) \in \{1, \dots, d\}^2$,

$$S^{(i,j)}(X)_{[0,t]} = \int_{0<s<t} S^{(i)}(X)_{[0,s]} dX^j_s = \int_{0<r<s<t} dX^i_r dX^j_s \quad \to \text{a path!}$$

- Recursively, for $(i_1, \dots, i_k) \in \{1, \dots, d\}^k$,

$$S^{(i_1,\dots,i_k)}(X)_{[0,t]} = \int_{0<t_1<t_2<\dots<t_k<t} dX^{i_1}_{t_1} \dots dX^{i_k}_{t_k}.$$

- $S^{(i_1,\dots,i_k)}(X)_{[0,1]}$ is the k-fold iterated integral of X along $i_1, \dots, i_k$.

Signature

Definition
The signature of X is the sequence of real numbers

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

Signature

Definition
The signature of X is the sequence of real numbers

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

- $d = 3 \to (1, 2, 3, 11, 12, 13, 21, 22, 23, 31, 32, 33, 111, 112, 113, \ldots)$

Signature

Definition
The signature of X is the sequence of real numbers

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

- $d = 3 \rightarrow (1, 2, 3, 11, 12, 13, 21, 22, 23, 31, 32, 33, 111, 112, 113, \ldots)$
- Tensor notation:

$$\mathbf{X^k} = \sum_{(i_1,\ldots,i_k)\subset\{1,\ldots,d\}^k} S^{(i_1,\ldots,i_k)}(X) e_{i_1} \otimes \cdots \otimes e_{i_k}.$$

Signature

Definition
The signature of X is the sequence of real numbers

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

- $d = 3 \to (1, 2, 3, 11, 12, 13, 21, 22, 23, 31, 32, 33, 111, 112, 113, \ldots)$
- Tensor notation:

$$\mathbf{X}^{\mathbf{k}} = \sum_{(i_1,\ldots,i_k)\subset\{1,\ldots,d\}^k} S^{(i_1,\ldots,i_k)}(X) e_{i_1} \otimes \cdots \otimes e_{i_k}.$$

- Signature:

$$S(X) = (1, \mathbf{X}^{\mathbf{1}}, \mathbf{X}^{\mathbf{2}}, \ldots, \mathbf{X}^{\mathbf{k}}, \ldots) \in T(\mathbb{R}^d),$$

Signature

Definition
The signature of X is the sequence of real numbers

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

- $d = 3 \rightarrow (1, 2, 3, 11, 12, 13, 21, 22, 23, 31, 32, 33, 111, 112, 113, \dots)$
- Tensor notation:

$$\mathbf{X^k} = \sum_{(i_1,\dots,i_k)\subset\{1,\dots,d\}^k} S^{(i_1,\dots,i_k)}(X) e_{i_1} \otimes \cdots \otimes e_{i_k}.$$

- Signature:

$$S(X) = (1, \mathbf{X^1}, \mathbf{X^2}, \dots, \mathbf{X^k}, \dots) \in T(\mathbb{R}^d),$$

  where

$$T(\mathbb{R}^d) = 1 \oplus \mathbb{R}^d \oplus (\mathbb{R}^d)^{\otimes 2} \oplus \cdots \oplus (\mathbb{R}^d)^{\otimes k} \oplus \cdots$$

Example

For $X_t = (X_t^1, X_t^2)$,

$$\mathbf{X}^1 = \begin{pmatrix} \int_0^1 dX_t^1 & \int_0^1 dX_t^2 \end{pmatrix} = \begin{pmatrix} X_1^1 - X_0^1 & X_1^2 - X_0^2 \end{pmatrix}$$

Example

For $X_t = (X_t^1, X_t^2)$,

$$\mathbf{X}^1 = \begin{pmatrix} \int_0^1 dX_t^1 & \int_0^1 dX_t^2 \end{pmatrix} = \begin{pmatrix} X_1^1 - X_0^1 & X_1^2 - X_0^2 \end{pmatrix}$$

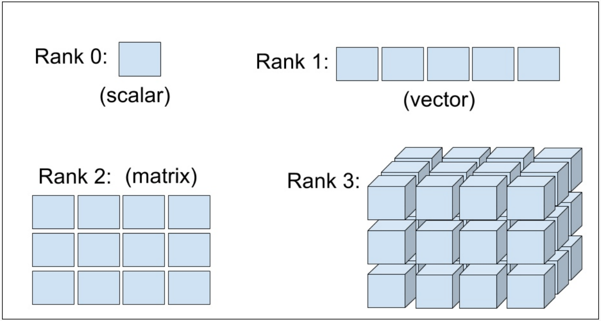
$$\mathbf{X}^2 = \begin{pmatrix} \int_0^1 \int_0^t dX_s^1 dX_t^1 & \int_0^1 \int_0^t dX_s^1 dX_t^2 \\ \int_0^1 \int_0^t dX_s^2 dX_t^1 & \int_0^1 \int_0^t dX_s^2 dX_t^2 \end{pmatrix}$$

Example

For $X_t = (X_t^1, X_t^2)$,

$$\mathbf{X}^1 = \begin{pmatrix} \int_0^1 dX_t^1 & \int_0^1 dX_t^2 \end{pmatrix} = \begin{pmatrix} X_1^1 - X_0^1 & X_1^2 - X_0^2 \end{pmatrix}$$

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



Truncated signature

- Truncated signature at order m:

$$S^m(X) = (1, \mathbf{X^1}, \mathbf{X^2}, \dots, \mathbf{X^m}).$$

Truncated signature

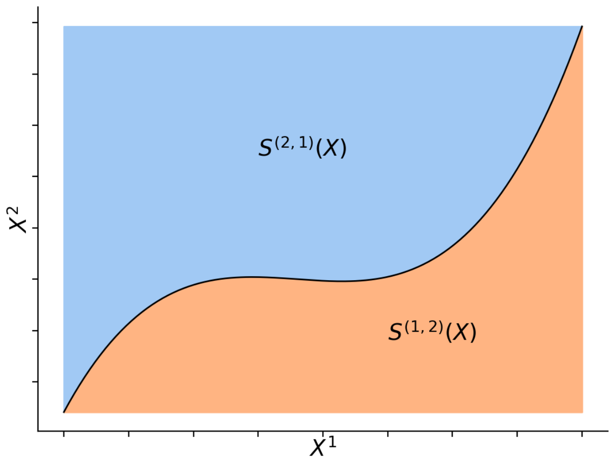
- Truncated signature at order m:

$$S^m(X) = (1, \mathbf{X^1}, \mathbf{X^2}, \dots, \mathbf{X^m}).$$

- Dimension:

$$s_d(m) = \sum_{k=0}^{m} d^k = \frac{d^{m+1} - 1}{d - 1}.$$

Geometric interpretation



Important example

Linear path

- $X : [0, 1] \to \mathbb{R}^d$ a linear path.

Important example

Linear path

- $X : [0, 1] \to \mathbb{R}^d$ a linear path.
- $X_t = X_0 + (X_1 - X_0)t$.

Important example

Linear path

- $X : [0,1] \to \mathbb{R}^d$ a linear path.
- $X_t = X_0 + (X_1 - X_0)t$.
- For any $I = (i_1, \ldots, i_k)$,

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

Important example

Linear path

- $X : [0,1] \to \mathbb{R}^d$ a linear path.
- $X_t = X_0 + (X_1 - X_0)t$.
- For any $I = (i_1, \dots, i_k)$,

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

▷ Very useful: in practice, we always deal with piecewise linear paths.

▷ Needed: concatenation operations.

Properties 1

Chen's identity

- $X : [a, b] \to \mathbb{R}^d$ and $Y : [b, c] \to \mathbb{R}^d$ paths.

Properties 1

Chen's identity

- $X: [a, b] \to \mathbb{R}^d$ and $Y: [b, c] \to \mathbb{R}^d$ paths.
- $X * Y: [a, c] \to \mathbb{R}^d$ the concatenation.

Properties 1

Chen's identity

- $X\colon [a,b] \to \mathbb{R}^d$ and $Y\colon [b,c] \to \mathbb{R}^d$ paths.
- $X * Y\colon [a,c] \to \mathbb{R}^d$ the concatenation.
- Then

$$S(X * Y) = S(X) \otimes S(Y).$$

Properties 1

Chen's identity

- $X : [a, b] \to \mathbb{R}^d$ and $Y : [b, c] \to \mathbb{R}^d$ paths.
- $X * Y : [a, c] \to \mathbb{R}^d$ the concatenation.
- Then
$$S(X * Y) = S(X) \otimes S(Y).$$

▷ We can compute the signature of piecewise linear paths!

▷ Data stream of p points and truncation at m: $O(pd^m)$ operations.

▷ Fast packages and libraries available in `C++` and `Python`.

Properties 2

Invariance under time reparametrization

- $X : [0, 1] \to \mathbb{R}^d$ a path.

Properties 2

Invariance under time reparametrization

- $X : [0, 1] \to \mathbb{R}^d$ a path.
- $\psi : [0, 1] \to [0, 1]$ a reparametrization

Properties 2

Invariance under time reparametrization

- $X : [0,1] \to \mathbb{R}^d$ a path.
- $\psi : [0,1] \to [0,1]$ a reparametrization
- If $\tilde{X}_t = X_{\psi(t)}$, then

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

Properties 2

Invariance under time reparametrization

- $X : [0,1] \to \mathbb{R}^d$ a path.
- $\psi : [0,1] \to [0,1]$ a reparametrization
- If $\tilde{X}_t = X_{\psi(t)}$, then

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

▷ A key advantage of the signature modeling.

▷ Encoding of the geometric properties of paths.

Properties 3

Time reversal

- $X : [0,1] \to \mathbb{R}^d$ a path.

Properties 3

Time reversal

- $X : [0,1] \to \mathbb{R}^d$ a path.
- $\overleftarrow{X}$ time-reversal of X: $\overleftarrow{X}_t = X_{1-t}$.
- If $\mathbf{1} = (1, 0, \ldots, 0, \ldots) \in T(\mathbb{R}^d)$, then

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

Properties 3

Time reversal

- $X : [0,1] \to \mathbb{R}^d$ a path.
- $\overleftarrow{X}$ time-reversal of X: $\overleftarrow{X}_t = X_{1-t}$.
- If $\mathbf{1} = (1, 0, \ldots, 0, \ldots) \in T(\mathbb{R}^d)$, then

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

▷ Think "$S(X)^{-1} = S(\overleftarrow{X})$".

Properties 3

Time reversal

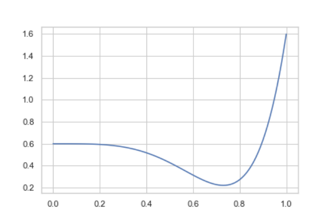
- $X : [0,1] \to \mathbb{R}^d$ a path.
- $\overleftarrow{X}$ time-reversal of X: $\overleftarrow{X}_t = X_{1-t}$.
- If $\mathbf{1} = (1, 0, \dots, 0, \dots) \in T(\mathbb{R}^d)$, then

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

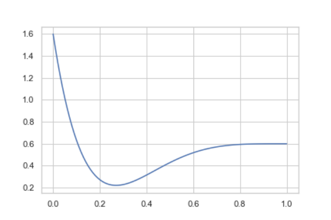
▷ Think "$S(X)^{-1} = S(\overleftarrow{X})$".

▷ Signature not unique: $S(X) \otimes S(\overleftarrow{X}) = S(X * \overleftarrow{X}) = \mathbf{1}$.

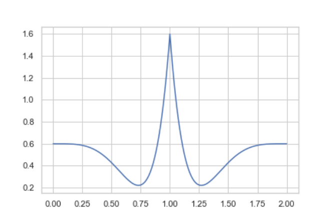
Properties 3



X



$\overleftarrow{X}$



$X * \overleftarrow{X}$

Properties 4

Tree-like paths

- Definition of an equivalence relation on paths such that

$$X \sim Y \Leftrightarrow S(X) = S(Y).$$

Properties 4

Tree-like paths

- Definition of an equivalence relation on paths such that

$$X \sim Y \Leftrightarrow S(X) = S(Y).$$

- $X \sim Y$ if $X * \overleftarrow{Y}$ is tree-like.

Properties 4

Tree-like paths

- Definition of an equivalence relation on paths such that

$$X \sim Y \Leftrightarrow S(X) = S(Y).$$

- $X \sim Y$ if $X * \overleftarrow{Y}$ is tree-like.
- $S(X) = \mathbf{1} \Leftrightarrow X$ tree-like.

Properties 4

Tree-like paths

- Definition of an equivalence relation on paths such that

$$X \sim Y \Leftrightarrow S(X) = S(Y).$$

- $X \sim Y$ if $X * \overleftarrow{Y}$ is tree-like.
- $S(X) = \mathbf{1} \Leftrightarrow X$ tree-like.
- Examples of tree-like paths:
  - $X * \overleftarrow{X}$,

Properties 4

Tree-like paths

- Definition of an equivalence relation on paths such that

$$X \sim Y \Leftrightarrow S(X) = S(Y).$$

- $X \sim Y$ if $X * \overleftarrow{Y}$ is tree-like.
- $S(X) = \mathbf{1} \Leftrightarrow X$ tree-like.
- Examples of tree-like paths:
  - $X * \overleftarrow{X}$,
  - $X * \overleftarrow{X} * \overleftarrow{Y} * Y$,

Properties 4

Tree-like paths

- Definition of an equivalence relation on paths such that

$$X \sim Y \Leftrightarrow S(X) = S(Y).$$

- $X \sim Y$ if $X * \overleftarrow{Y}$ is tree-like.
- $S(X) = \mathbf{1} \Leftrightarrow X$ tree-like.
- Examples of tree-like paths:
  - $X * \overleftarrow{X}$,
  - $X * \overleftarrow{X} * \overleftarrow{Y} * Y$,
  - $X * Y * \overleftarrow{Z} * Z * \overleftarrow{Y} * \overleftarrow{X}$.

Properties 4

Uniqueness

- For any X, there exists a unique path of minimal length in its equivalence class, denoted by $\overline{X}$ and called the reduced path.

Properties 4

Uniqueness

- For any X, there exists a unique path of minimal length in its equivalence class, denoted by $\overline{X}$ and called the reduced path.
- If X has at least one monotonic coordinate, then $S(X)$ determines X uniquely, up to translation and reparametrization.

Properties 4

Uniqueness

- For any X, there exists a unique path of minimal length in its equivalence class, denoted by $\overline{X}$ and called the reduced path.
- If X has at least one monotonic coordinate, then $S(X)$ determines X uniquely, up to translation and reparametrization.

▷ The signature characterizes paths.

▷ Trick: add a dummy monotonic component to X.

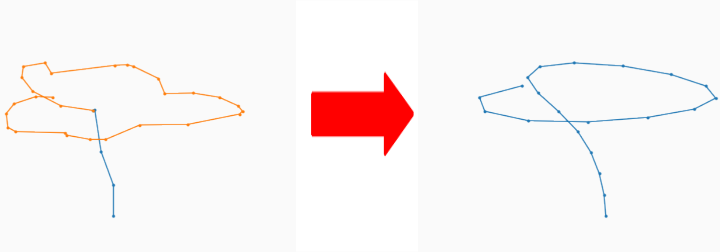
▷ Important concept of augmentation.

Can we reconstruct the path from its signature?

- Currently a lot of work in this direction;
- Efficient algorithm for piecewise linear paths (Chang and Lyons, 2019) → Python implementation.
- Applications in signal processing, e.g., sound compression, time series smoothing...

Can we reconstruct the path from its signature?

- Currently a lot of work in this direction;
- Efficient algorithm for piecewise linear paths (`Chang and Lyons, 2019`) → `Python` implementation.
- Applications in signal processing, e.g., sound compression, time series smoothing...



Properties 5

Signature approximation

- D compact subset of paths from $[0, 1]$ to $\mathbb{R}^d$ that are not tree-like equivalent.

Properties 5

Signature approximation

- D compact subset of paths from $[0,1]$ to $\mathbb{R}^d$ that are not tree-like equivalent.
- $f: D \to \mathbb{R}$ continuous.

Properties 5

Signature approximation

- D compact subset of paths from $[0,1]$ to $\mathbb{R}^d$ that are not tree-like equivalent.
- $f : D \to \mathbb{R}$ continuous.
- Then, for every $\varepsilon > 0$, there exists $w \in T(\mathbb{R}^d)$ such that, for any $X \in D$,

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

Properties 5

Signature approximation

- D compact subset of paths from $[0,1]$ to $\mathbb{R}^d$ that are not tree-like equivalent.
- $f: D \to \mathbb{R}$ continuous.
- Then, for every $\varepsilon > 0$, there exists $w \in T(\mathbb{R}^d)$ such that, for any $X \in D$,
$$\big| f(X) - \langle w, S(X) \rangle \big| \leq \varepsilon.$$

▷ Signature and linear model are happy together!

▷ This raises many interesting statistical issues.

Properties 6

Exponential decay of signature coefficients

- $X : [0, 1] \to \mathbb{R}^d$ a path.

Properties 6

Exponential decay of signature coefficients

- $X : [0,1] \to \mathbb{R}^d$ a path.
- Then, for any $k \geq 0$, $I \subset \{1, \dots d\}^k$,

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

Properties 6

Exponential decay of signature coefficients

- $X : [0,1] \to \mathbb{R}^d$ a path.
- Then, for any $k \geq 0$, $I \subset \{1, \dots d\}^k$,

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

▷ Useful for approximation properties.

Learning with signatures

Supervised learning

- Goal: understand the relationship between $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$.

Supervised learning

- Goal: understand the relationship between $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$.
- Regression: $\mathscr{Y} = \mathbb{R}$ Classification: $\mathscr{Y} = \{1, \ldots, q\}$.

Supervised learning

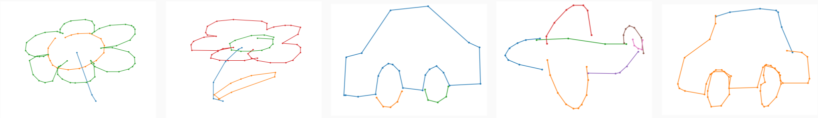
- Goal: understand the relationship between $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$.
- Regression: $\mathscr{Y} = \mathbb{R}$ Classification: $\mathscr{Y} = \{1, \dots, q\}$.
- Data: $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathscr{X} \times \mathscr{Y}$, i.i.d. $\sim (X, Y)$.

Supervised learning

- Goal: understand the relationship between $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$.
- Regression: $\mathscr{Y} = \mathbb{R}$ Classification: $\mathscr{Y} = \{1, \dots, q\}$.
- Data: $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathscr{X} \times \mathscr{Y}$, i.i.d. $\sim (X, Y)$.
- Prediction function: $f_\theta(X) \approx Y$, $\theta \in \mathbb{R}^p$.

Supervised learning

- Goal: understand the relationship between $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$.
- Regression: $\mathscr{Y} = \mathbb{R}$ Classification: $\mathscr{Y} = \{1, \ldots, q\}$.
- Data: $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathscr{X} \times \mathscr{Y}$, i.i.d. $\sim (X, Y)$.
- Prediction function: $f_\theta(X) \approx Y$, $\theta \in \mathbb{R}^p$.



$y_1 = 1$ $y_2 = 1$ $y_3 = 2$ $y_4 = 3$ $y_5 = 2$

Supervised learning

- Loss function $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}^+$.

Supervised learning

- Loss function $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}^+$.
- Empirical risk minimization: choose

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)).$$

Supervised learning

- Loss function $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}^+$.
- Empirical risk minimization: choose

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)).$$

- **Least squares regression**: $\mathscr{Y} = \mathbb{R}$ and $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.

Supervised learning

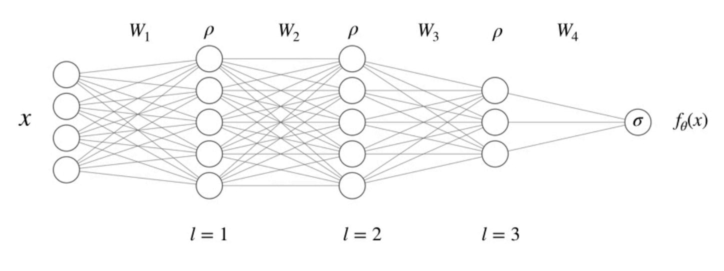
- Loss function $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}^+$.
- Empirical risk minimization: choose

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^p}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)).$$

- **Least squares regression**: $\mathscr{Y} = \mathbb{R}$ and $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.
- **Binary classification**: $\mathscr{Y} = \{0, 1\}$ and $\ell(y, f_\theta(x)) = \mathbb{1}_{[f_\theta(x) \neq y]}$.

Feedforward neural network

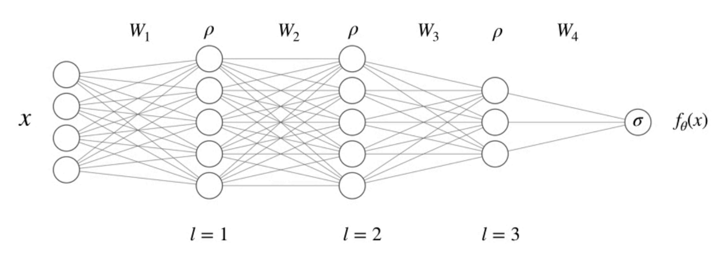
$$f_\theta(x) = \sigma(T_L\rho(T_{L-1}\rho(\cdots\rho(T_1x))))$$



Feedforward neural network

$$f_\theta(x) = \sigma(T_L \rho(T_{L-1} \rho(\cdots \rho(T_1 x))))$$

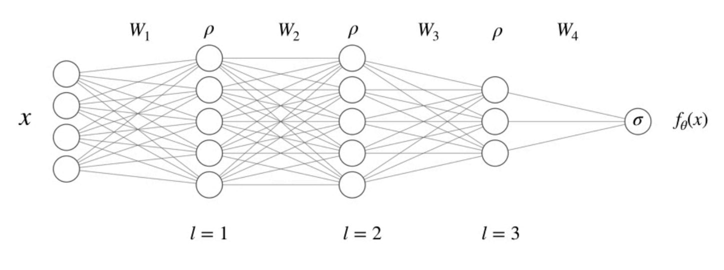
- $L - 1$ hidden layers.



Feedforward neural network

$$f_\theta(x) = \sigma(T_L\rho(T_{L-1}\rho(\cdots\rho(T_1x))))$$

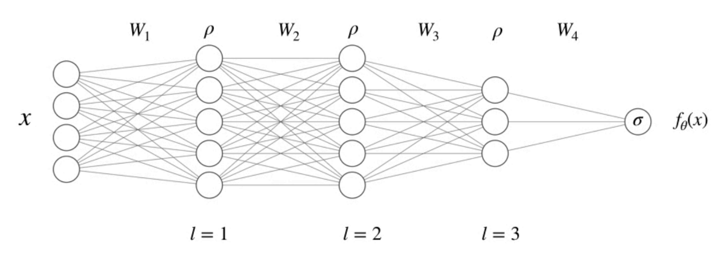
- $L-1$ hidden layers.
- $T_\ell x = W_\ell x + b_\ell$, $\ell = 1, \ldots, L$.



Feedforward neural network

$$f_\theta(x) = \sigma(T_L \rho(T_{L-1} \rho(\cdots \rho(T_1 x))))$$

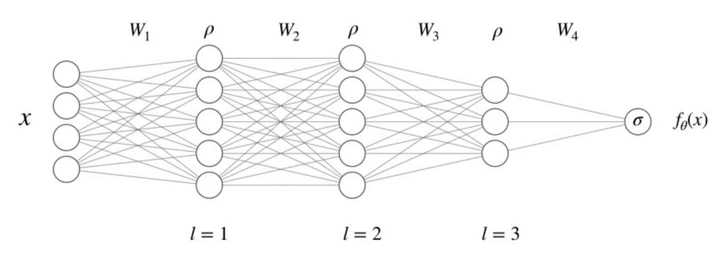
- $L-1$ hidden layers.
- $T_\ell x = W_\ell x + b_\ell$, $\ell = 1, \dots, L$.
- ρ = activation function (ReLU $\rho(x) = \max(x, 0)$).



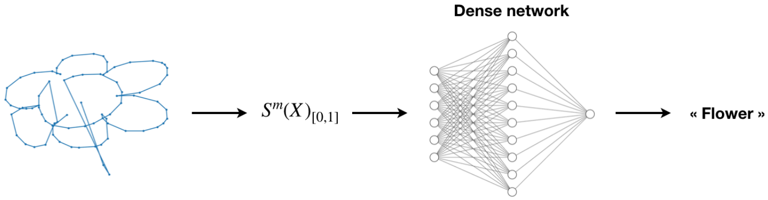
Feedforward neural network

$$f_\theta(x) = \sigma(T_L\rho(T_{L-1}\rho(\cdots\rho(T_1x))))$$

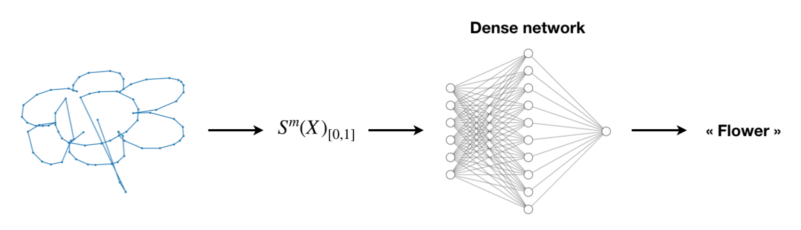
- $L-1$ hidden layers.
- $T_\ell x = W_\ell x + b_\ell$, $\ell = 1, \ldots, L$.
- ρ = activation function (ReLU $\rho(x) = \max(x, 0)$).
- σ = output function.



Signature + learning algorithm

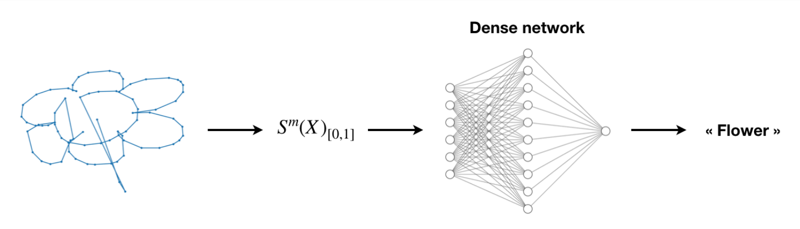


Signature + learning algorithm



▷ Yang et al. (2017): skeleton-based human action recognition.

Signature + learning algorithm



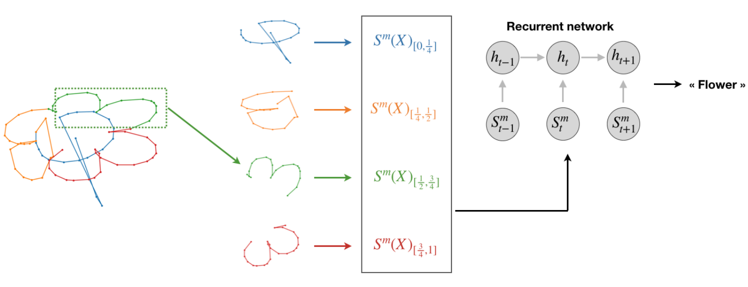
- Yang et al. (2017): skeleton-based human action recognition.
- Sequence of positions of human joints → high dimensional signature coefficients → small dense network.

Temporal approaches

- Idea: construct a path of signature coefficients.

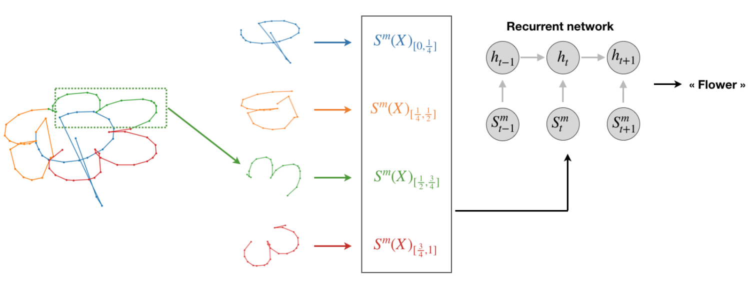
Temporal approaches

- Idea: construct a path of signature coefficients.



Temporal approaches

- Idea: construct a path of signature coefficients.



▷ Lai et al. (2017) and Liu et al. (2017): writer recognition.

Questions

- How should we choose the order of truncation?

Questions

- How should we choose the order of truncation?
- How does it perform compared to traditional functional linear models ?

Questions

- How should we choose the order of truncation?
- How does it perform compared to traditional functional linear models ?
- Could we find a canonical signature pipeline that would be a domain-agnostic starting point for practitioners?

The signature linear model

Regression model on the signature

- $X : [0,1] \to \mathbb{R}^d$ random path, $Y \in \mathbb{R}$ random variable.

Regression model on the signature

- $X : [0,1] \to \mathbb{R}^d$ random path, $Y \in \mathbb{R}$ random variable.
- Assumption: there exists $m^* \in \mathbb{N}$, $\beta^* \in \mathbb{R}^{s_d(m^*)}$ such that

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

Regression model on the signature

- $X : [0,1] \to \mathbb{R}^d$ random path, $Y \in \mathbb{R}$ random variable.
- Assumption: there exists $m^* \in \mathbb{N}$, $\beta^* \in \mathbb{R}^{s_d(m^*)}$ such that

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

- Goal: estimate m^* and β^*.

Regression model on the signature

→ m^* is a key quantity! Recall that

$$s_d(m) = \sum_{k=0}^{m} d^k = \frac{d^{m+1} - 1}{d - 1}.$$

Regression model on the signature

→ m^* is a key quantity! Recall that

$$s_d(m) = \sum_{k=0}^{m} d^k = \frac{d^{m+1}-1}{d-1}.$$

Typical values of $s_d(m)$.

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

Estimation of m^*

- Data: $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d.

Estimation of m^*

- Data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
- For any $m \in \mathbb{N}$, $\alpha > 0$,

$$B_{m,\alpha} = \left\{ \beta \in \mathbb{R}^{s_d(m)} : \|\beta\|_2 \leq \alpha \right\}.$$

Estimation of m^*

- Data: $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d.
- For any $m \in \mathbb{N}$, $\alpha > 0$,

$$B_{m,\alpha} = \left\{ \beta \in \mathbb{R}^{s_d(m)} : \|\beta\|_2 \leq \alpha \right\}.$$

- For any $m \in \mathbb{N}$, $\beta \in B_{m,\alpha}$,

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

Estimation of m^*

- Data: $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d.
- For any $m \in \mathbb{N}$, $\alpha > 0$,

$$B_{m,\alpha} = \left\{ \beta \in \mathbb{R}^{s_d(m)} : \|\beta\|_2 \leq \alpha \right\}.$$

- For any $m \in \mathbb{N}$, $\beta \in B_{m,\alpha}$,

$$\mathcal{R}_{m,n}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \langle \beta, S^m(X_i) \rangle \big)^2.$$

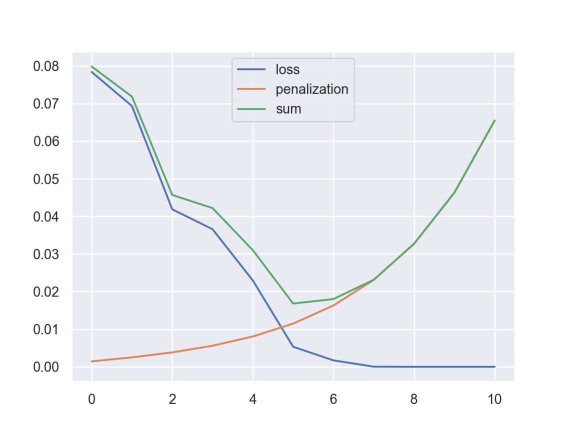
- For any $m \in \mathbb{N}$,

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

Estimation of m^*

Estimator:

$$\widehat{m} = \min\Big(\underset{m}{\operatorname{argmin}}\big(\widehat{L}_n(m) + \operatorname{pen}_n(m)\big)\Big).$$



Result

Additional assumptions:

(H_α) $\beta^* \in B_{m^*,\alpha}$.

(H_K) There exists $K_Y > 0$ and $K_X > 0$ such that almost surely

$$|Y| \leq K_Y \quad \text{and} \quad \|X\|_{1\text{-var}} \leq K_X.$$

Result

Theorem

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

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

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

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

where n_0, C_1 and C_2 are explicit constants.

Result

Theorem

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

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

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

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

where n_0, C_1 and C_2 are explicit constants.

Corollary
$\widehat{m}$ converges almost surely towards m^*.

Result

We can then estimate β^* by

$$\widehat{\beta} = \underset{\beta \in B_{\widehat{m},\alpha}}{\operatorname{argmin}}\ \mathcal{R}_{\widehat{m},n}(\beta),$$

Result

We can then estimate β^* by

$$\widehat{\beta} = \underset{\beta \in B_{\widehat{m},\alpha}}{\operatorname{argmin}} \mathcal{R}_{\widehat{m},n}(\beta),$$

and show that

$$\mathbb{E}\Big(\big\langle \widehat{\beta}, S^{\widehat{m}}(X)\big\rangle - \big\langle \beta^*, S^{m^*}(X)\big\rangle\Big)^2 = \mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big).$$

Functional linear model

- In the case $d = 1$,

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

Functional linear model

- In the case $d = 1$,

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

- Basis expansion:

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

Functional linear model

- In the case $d = 1$,

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

- Basis expansion:

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

- Back to the multivariate case: estimate the b_ks.

Functional linear model

- In the case $d = 1$,

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

- Basis expansion:

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

- Back to the multivariate case: estimate the b_ks.

▷ Choice for $\phi_1, \ldots, \phi_K$? Splines, monomials, Fourier basis... or functional principal components of the X_is.

Functional linear model

- In the case $d = 1$,

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

- Basis expansion:

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

- Back to the multivariate case: estimate the b_ks.

▷ Choice for $\phi_1, \ldots, \phi_K$? Splines, monomials, Fourier basis... or functional principal components of the X_is.

▷ If $d > 2$? Treat each coordinate independently.

Dimension study

- Gaussian processes covariates: or any $t \in [0,1]$, $1 \leq i \leq n$, $1 \leq k \leq d$,

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

Dimension study

- Gaussian processes covariates: or any $t \in [0,1]$, $1 \leq i \leq n$, $1 \leq k \leq d$,

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

- ξ_i^k is a Gaussian process with exponential covariance matrix.

Dimension study

- Gaussian processes covariates: or any $t \in [0,1]$, $1 \leq i \leq n$, $1 \leq k \leq d$,

$$X^k_{i,t} = \alpha^k_i t + \xi^k_{i,t}, \quad 1 \leq k \leq d, \quad t \in [0,1],$$

- ξ^k_i is a Gaussian process with exponential covariance matrix.
- Response is the norm of the trend: $Y_i = \|\alpha_i\|$.

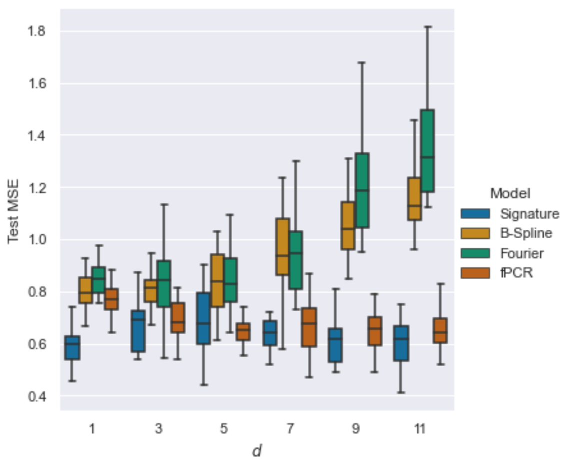
Dimension study

- Gaussian processes covariates: or any $t \in [0,1]$, $1 \le i \le n$, $1 \le k \le d$,

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

- ξ_i^k is a Gaussian process with exponential covariance matrix.
- Response is the norm of the trend: $Y_i = \|\alpha_i\|$.

Dimension study



Electricity consumption

- Electricity consumption of 370 clients, recorded every 15min from 2011 to 2014.

Electricity consumption

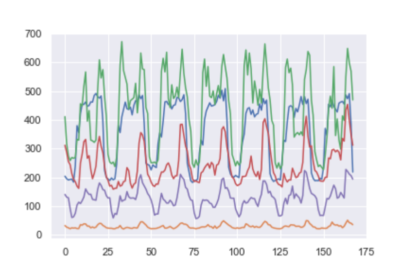
- Electricity consumption of 370 clients, recorded every 15min from 2011 to 2014.
- Observe a subset of clients during a week and predict the consumption peak of the following week: maximal consumption summed over all clients.

Electricity consumption

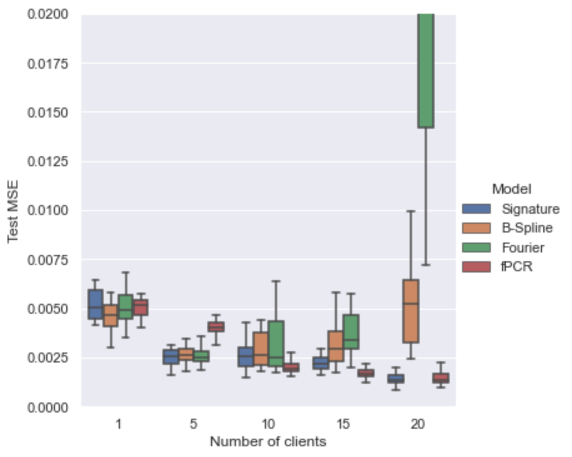
- Electricity consumption of 370 clients, recorded every 15min from 2011 to 2014.
- Observe a subset of clients during a week and predict the consumption peak of the following week: maximal consumption summed over all clients.
- Vary the size of the subset: the more clients the more information!

Electricity consumption

- Electricity consumption of 370 clients, recorded every 15min from 2011 to 2014.
- Observe a subset of clients during a week and predict the consumption peak of the following week: maximal consumption summed over all clients.
- Vary the size of the subset: the more clients the more information!

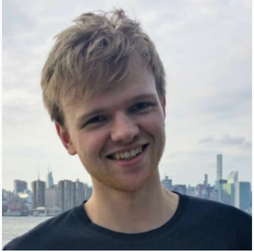


Electricity consumption



A generalized signature method for multivariate time series classification

Joint work with



James Morrill
University of Oxford

Patrick Kidger
University of Oxford

Terry Lyons
University of Oxford

Overview

- Goal: systematic comparison of the different variations of the signature method.

Overview

- Goal: systematic comparison of the different variations of the signature method.
- Empirical study over 26 datasets of time series classification.

Overview

- Goal: systematic comparison of the different variations of the signature method.
- Empirical study over 26 datasets of time series classification.
- Define a generalised signature method as a framework to capture all these variations.

Overview

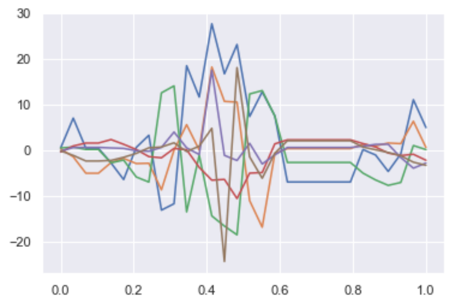
- Goal: systematic comparison of the different variations of the signature method.
- Empirical study over 26 datasets of time series classification.
- Define a generalised signature method as a framework to capture all these variations.
- Give practitioners some simple, domain-agnostic guidelines for a first signature algorithm.

Framework

- Input: a sequence $\mathbf{x} \in \mathcal{S}(\mathbb{R}^d)$, where

$$\mathcal{S}(\mathbb{R}^d) = \{(x_1, \ldots, x_n) \,|\, x_i \in \mathbb{R}^d, n \in \mathbb{N}\}.$$

Racketsports dataset



A sample $\mathbf{x}$ with $d = 6$, $n = 30$

Framework

- Input: a sequence $\mathbf{x} \in \mathcal{S}(\mathbb{R}^d)$, where

$$\mathcal{S}(\mathbb{R}^d) = \{(x_1, \dots, x_n) \,|\, x_i \in \mathbb{R}^d, n \in \mathbb{N}\}.$$

- Output: a label $y \in \{1, \dots, q\}$.

Framework

- For some $e, p \in \mathbb{N}$, an augmentation is a map

$$\phi = (\phi^1, \dots, \phi^p) \colon \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^e)^p.$$

- For some $q \in \mathbb{N}$, a window is a map

$$W \colon \mathcal{S}(\mathbb{R}^e) \to \mathcal{S}(\mathbb{R}^e)^w.$$

- Signature or logsignature transform: S^m.
- Rescaling operation ρ_{post} or ρ_{pre}.

Feature set

$$\mathbf{y}_{i,j} = (\rho_{\mathrm{post}} \circ S^m \circ \rho_{\mathrm{pre}} \circ W^j \circ \phi^i)(\mathbf{x}).$$

Framework

- For some $e, p \in \mathbb{N}$, an augmentation is a map

$$\phi = (\phi^1, \ldots, \phi^p)\colon \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^e)^p.$$

- For some $q \in \mathbb{N}$, a window is a map

$$W\colon \mathcal{S}(\mathbb{R}^e) \to \mathcal{S}(\mathbb{R}^e)^w.$$

- Signature or logsignature transform: S^m.
- Rescaling operation ρ_{post} or ρ_{pre}.

Feature set

$$\mathbf{y}_{i,j} = (\rho_{\mathrm{post}} \circ S^m \circ \rho_{\mathrm{pre}} \circ W^j \circ \phi^i)(\mathbf{x}).$$

Augmentations

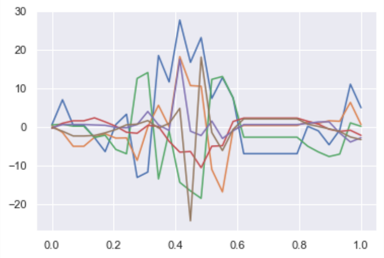
- Time augmentation

$$\phi_{\mathbf{t}}(\mathbf{x}) = \big((t_1, x_1), \ldots, (t_n, x_n)\big) \in \mathcal{S}(\mathbb{R}^{d+1}).$$

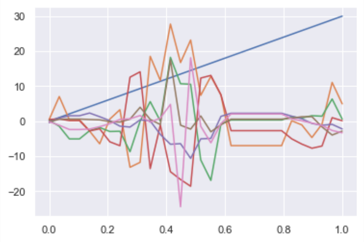
Augmentations

- Time augmentation

$$\phi_{\mathbf{t}}(\mathbf{x}) = \big((t_1, x_1), \ldots, (t_n, x_n)\big) \in \mathcal{S}(\mathbb{R}^{d+1}).$$



Sample $\mathbf{x} \in \mathcal{S}(\mathbb{R}^6)$

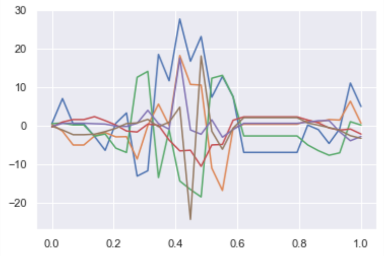


Augmented path $\phi(\mathbf{x}) \in \mathcal{S}(\mathbb{R}^7)$

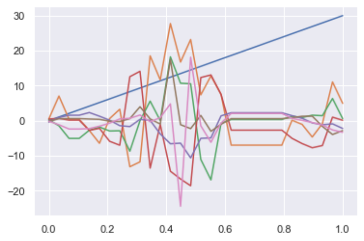
Augmentations

- Time augmentation

$$\phi_{\mathbf{t}}(\mathbf{x}) = \big((t_1, x_1), \ldots, (t_n, x_n)\big) \in \mathcal{S}(\mathbb{R}^{d+1}).$$



Sample $\mathbf{x} \in \mathcal{S}(\mathbb{R}^6)$



Augmented path $\phi(\mathbf{x}) \in \mathcal{S}(\mathbb{R}^7)$

▷ Sensitivity to parametrization and ensures signature uniqueness.

Augmentations

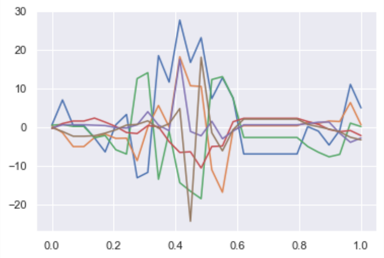
- Lead-lag augmentation

$$\phi(\mathbf{x}) = ((x_1, x_1), (x_2, x_1), (x_2, x_2), \ldots, (x_n, x_n)) \in \mathcal{S}(\mathbb{R}^{2d}).$$

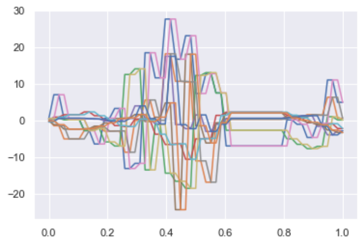
Augmentations

- Lead-lag augmentation

$$\phi(\mathbf{x}) = ((x_1, x_1), (x_2, x_1), (x_2, x_2), \ldots, (x_n, x_n)) \in \mathcal{S}(\mathbb{R}^{2d}).$$



Sample $\mathbf{x} \in \mathcal{S}(\mathbb{R}^6)$

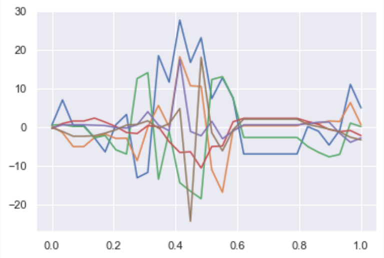


Augmented path $\phi(\mathbf{x}) \in \mathcal{S}(\mathbb{R}^{12})$

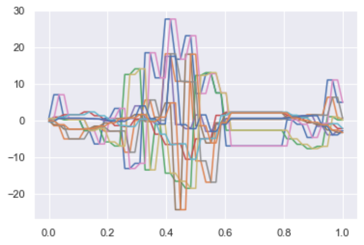
Augmentations

- Lead-lag augmentation

$$\phi(\mathbf{x}) = ((x_1, x_1), (x_2, x_1), (x_2, x_2), \ldots, (x_n, x_n)) \in \mathcal{S}(\mathbb{R}^{2d}).$$



Sample $\mathbf{x} \in \mathcal{S}(\mathbb{R}^6)$



Augmented path $\phi(\mathbf{x}) \in \mathcal{S}(\mathbb{R}^{12})$

▷ Captures the quadratic variation of a process.

Augmentations

- Basepoint augmentation

$$\phi(\mathbf{x}) = (0, x_1, \ldots, x_n) \in \mathcal{S}(\mathbb{R}^d).$$

Augmentations

- Basepoint augmentation

$$\phi(\mathbf{x}) = (0, x_1, \ldots, x_n) \in \mathcal{S}(\mathbb{R}^d).$$

- Invisibility-reset augmentation

$$\phi(\mathbf{x}) = \big((1, x_1), \ldots, (1, x_{n-1}), (1, x_n), (0, x_n), (0, 0)\big) \in \mathcal{S}(\mathbb{R}^{d+1}).$$

Augmentations

- Basepoint augmentation

$$\phi(\mathbf{x}) = (0, x_1, \ldots, x_n) \in \mathcal{S}(\mathbb{R}^d).$$

- Invisibility-reset augmentation

$$\phi(\mathbf{x}) = \big((1, x_1), \ldots, (1, x_{n-1}), (1, x_n), (0, x_n), (0, 0)\big) \in \mathcal{S}(\mathbb{R}^{d+1}).$$

▷ Sensitivity to translations.

Framework

- For some $e, p \in \mathbb{N}$, an augmentation is a map

$$\phi = (\phi^1, \dots, \phi^p) \colon \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^e)^p.$$

- For some $q \in \mathbb{N}$, a window is a map

$$W \colon \mathcal{S}(\mathbb{R}^e) \to \mathcal{S}(\mathbb{R}^e)^w.$$

- Signature or logsignature transform: S^m.
- Rescaling operation ρ_{post} or ρ_{pre}.

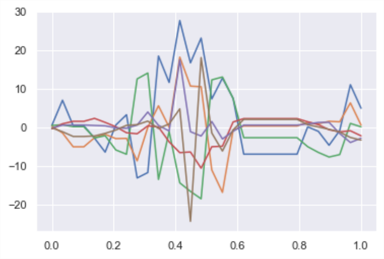
Feature set

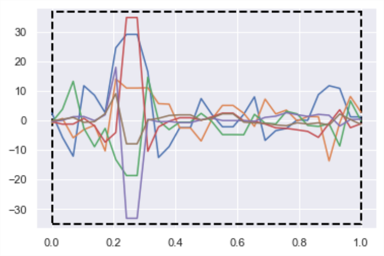
$$\mathbf{y}_{i,j} = (\rho_{\mathrm{post}} \circ S^m \circ \rho_{\mathrm{pre}} \circ W^j \circ \phi^i)(\mathbf{x}).$$

Windows

- Global window

$$W(\mathbf{x}) = (\mathbf{x}) \in \mathcal{S}(\mathbb{R}^e),$$

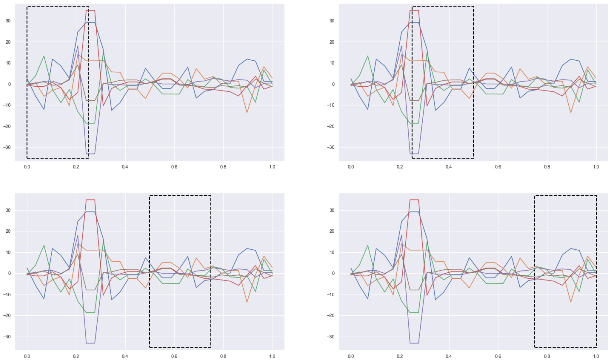




Windows

- Sliding window

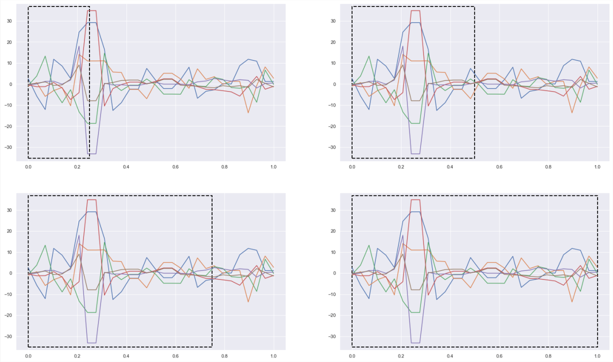
$$W(\mathbf{x}) = (\mathbf{x}_{1,\ell}, \mathbf{x}_{l+1,l+\ell}, \mathbf{x}_{2l+1,2l+\ell}, \ldots) \in \mathcal{S}(\mathcal{S}(\mathbb{R}^e)),$$



Windows

- Expanding window

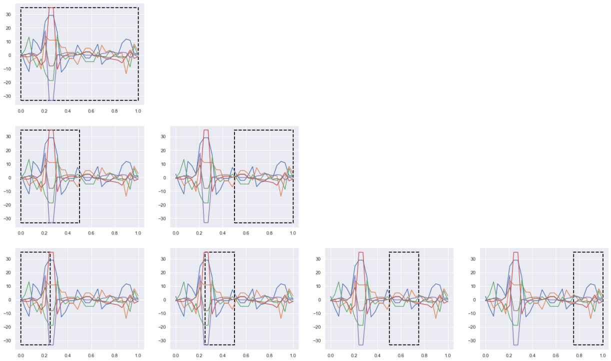
$$W(\mathbf{x}) = (\mathbf{x}_{1,\ell}, \mathbf{x}_{1,l+\ell}, \mathbf{x}_{1,2l+\ell}, \ldots) \in \mathcal{S}(\mathcal{S}(\mathbb{R}^e)).$$



Windows

- Dyadic window

$$W(\mathbf{x}) = (W^1(\mathbf{x}), \ldots, W^q(\mathbf{x})) \in \mathcal{S}(\mathcal{S}(\mathbb{R}^e))^q.$$



Framework

- For some $e, p \in \mathbb{N}$, an augmentation is a map

$$\phi = (\phi^1, \dots, \phi^p) \colon \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^e)^p.$$

- For some $q \in \mathbb{N}$, a window is a map

$$W \colon \mathcal{S}(\mathbb{R}^e) \to \mathcal{S}(\mathbb{R}^e)^w.$$

- Signature or logsignature transform: S^m.
- Rescaling operation ρ_{post} or ρ_{pre}.

Feature set

$$\mathbf{y}_{i,j} = (\rho_{\mathrm{post}} \circ S^m \circ \rho_{\mathrm{pre}} \circ W^j \circ \phi^i)(\mathbf{x}).$$

Framework

- Signature transform

$$S^m(\mathbf{x}) = (1, \mathbf{X^1}, \mathbf{X^2}, \ldots, \mathbf{X^m}).$$

Framework

- Signature transform

$$S^m(\mathbf{x}) = (1, \mathbf{X^1}, \mathbf{X^2}, \dots, \mathbf{X^m}).$$

- Logsignature transform $\log(S^m(\mathbf{x}))$, where for any $a \in T((\mathbb{R}^d))$,

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

Framework

- Signature transform

$$S^m(\mathbf{x}) = (1, \mathbf{X^1}, \mathbf{X^2}, \dots, \mathbf{X^m}).$$

- Logsignature transform $\log(S^m(\mathbf{x}))$, where for any $a \in T((\mathbb{R}^d))$,

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

▷ Same information and logsignature less dimensional but no linear approximation property.

Signature versus logsignature

Table 1: Typical dimensions of $S^m(\mathbf{x})$ and $\log(S^m(\mathbf{x}))$.

	$d = 2$	$d = 3$	$d = 6$
$m = 1$	2 / 2	3 / 3	6 / 6
$m = 2$	6 / 3	12 / 6	42 / 21
$m = 5$	62 / 14	363 / 80	9330 / 1960
$m = 7$	254 / 41	3279 / 508	335922 / 49685

Empirical study methodology

- 26 datasets: Human Activities and Postural Transitions, Speech Commands and 24 datasets from the UEA archive.

Empirical study methodology

- 26 datasets: Human Activities and Postural Transitions, Speech Commands and 24 datasets from the UEA archive.
- Definition of a baseline:

Empirical study methodology

- 26 datasets: Human Activities and Postural Transitions, Speech Commands and 24 datasets from the UEA archive.
- Definition of a baseline: time augmentation + global window + signature of depth 3 + pre-signature scaling

Empirical study methodology

- 26 datasets: Human Activities and Postural Transitions, Speech Commands and 24 datasets from the UEA archive.
- Definition of a baseline: time augmentation + global window + signature of depth 3 + pre-signature scaling

$$(S^3 \circ \rho_{\mathrm{pre}} \circ \phi_{\mathbf{t}})(\mathbf{x}).$$

Empirical study methodology

- 26 datasets: Human Activities and Postural Transitions, Speech Commands and 24 datasets from the UEA archive.
- Definition of a baseline: time augmentation + global window + signature of depth 3 + pre-signature scaling

$$(S^3 \circ \rho_{\mathrm{pre}} \circ \phi_{\mathbf{t}})(\mathbf{x}).$$

- Vary each group of options with regards to this baseline.

Empirical study methodology

- 26 datasets: Human Activities and Postural Transitions, Speech Commands and 24 datasets from the UEA archive.
- Definition of a baseline: time augmentation + global window + signature of depth 3 + pre-signature scaling

$$(S^3 \circ \rho_{\mathrm{pre}} \circ \phi_{\mathbf{t}})(\mathbf{x}).$$

- Vary each group of options with regards to this baseline.
- 4 classifiers: logistic regression, random forest, GRU, CNN.

Empirical study methodology

- 26 datasets: Human Activities and Postural Transitions, Speech Commands and 24 datasets from the UEA archive.
- Definition of a baseline: time augmentation + global window + signature of depth 3 + pre-signature scaling

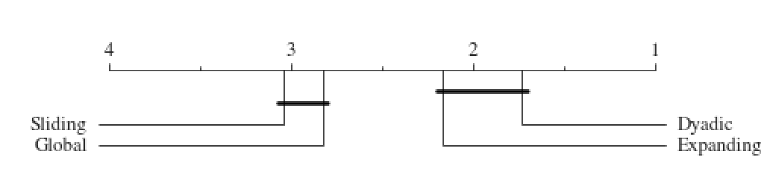
$$(S^3 \circ \rho_{\mathrm{pre}} \circ \phi_{\mathbf{t}})(\mathbf{x}).$$

- Vary each group of options with regards to this baseline.
- 4 classifiers: logistic regression, random forest, GRU, CNN.

$\rightarrow$ 9984 combinations.

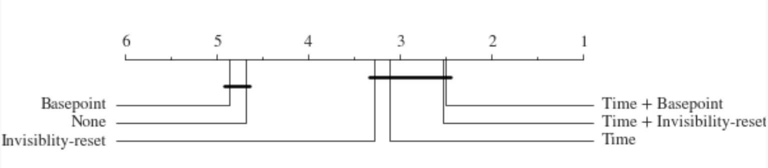
Results

- Windows:



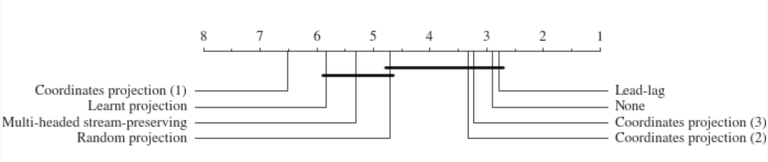
Results

▷ Invariance-removing augmentations:



Results

- Other augmentations:

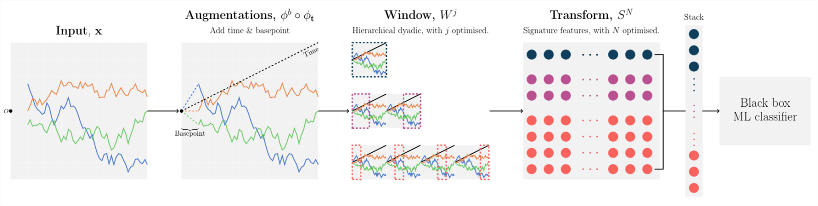


Results

- Signature versus logsignature transform:

	Signature	Logsignature
Average ranks	**1.25**	1.75
p-value		0.01

Canonical signature pipeline

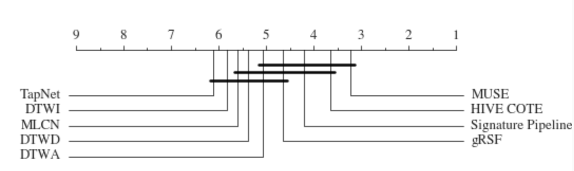


Canonical signature pipeline

Implement this pipeline on the 30 datasets from the UEA archive, with a random forest classifier, and compare it to benchmark classifiers.

Canonical signature pipeline

Implement this pipeline on the 30 datasets from the UEA archive, with a random forest classifier, and compare it to benchmark classifiers.



- Competitive with ensemble methods (MUSE and HIVE COTE) and deep neural networks (MLCN and TapNet).

Conclusion

- Signatures are a flexible tool.

Conclusion

- Signatures are a flexible tool.
- The combination "signature + generic algorithm" $\approx$ state-of-the-art.

Conclusion

- Signatures are a flexible tool.
- The combination "signature + generic algorithm" $\approx$ state-of-the-art.
- Few computing resources and no domain-specific knowledge.

Conclusion

- Signatures are a flexible tool.
- The combination "signature + generic algorithm" $\approx$ state-of-the-art.
- Few computing resources and no domain-specific knowledge.
- A lot of open questions and potential applications.

Thank you!