The signature method for data streams.

Adeline Fermanian[1], supervised by Gérard Biau[1] and Benoît Cadre[2]

[1]LPSM, Sorbonne Université, Paris [2]IRMAR, ENS Rennes

Setting

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

Goal: find a good representation of these processes as vectors of finite dimension.

History

- 1960s Chen notices in [1] that a path can be represented by its iterated integrals.
- 1990s The signature is at the center of Lyons' rough paths theory.
- 2010s Combined with a deep learning algorithms, it achieves state of the art results for several applications, see e.g. [2] or [3].

Definition

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

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

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

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

The signature of X truncated at order m is:

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

Properties

- **Invariance under time reparametrization:** Let $\psi : [0, 1] \to [0, 1]$ be a reparametrization. Then, if $\tilde{X}_t = X_{\psi(t)}$, $S(\tilde{X}) = S(X)$.
- **Uniqueness:** If X has at least one monotonous coordinate, then $S(X)$ determines X uniquely.
- **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 $X \in D$,

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

Learning process

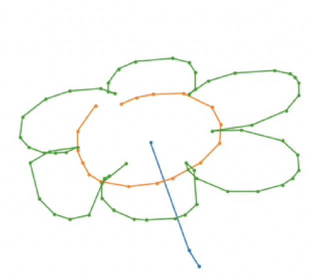
Data $\longrightarrow$ Path embedding $\longrightarrow$ Signature transformation $\longrightarrow$ Prediction with any learning algorithm

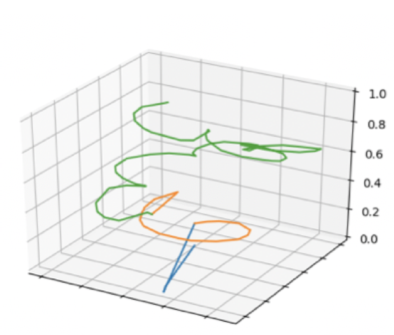
$$\left[[(x_1^1, y_1^1), \dots, (x_{p_1}^1, y_{p_1}^1)], [(x_1^2, y_1^2), \dots, (x_{p_2}^2, y_{p_2}^2)], \dots [(x_1^K, y_1^K), \dots, (x_{p_K}^K, y_{p_K}^K)]\right]$$

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

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

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





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

« Flower »

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.
- Data embedding has a huge influence on prediction performance.

Mathematical challenges

- Truncation order selection in a regression model: estimator, rate of convergence and simulation study.
- Theoretical understanding of embedding properties.
- Extension to paths of finite p-variation with $p \ge 2$.
- Sparsity in the signature vector.

Dataset

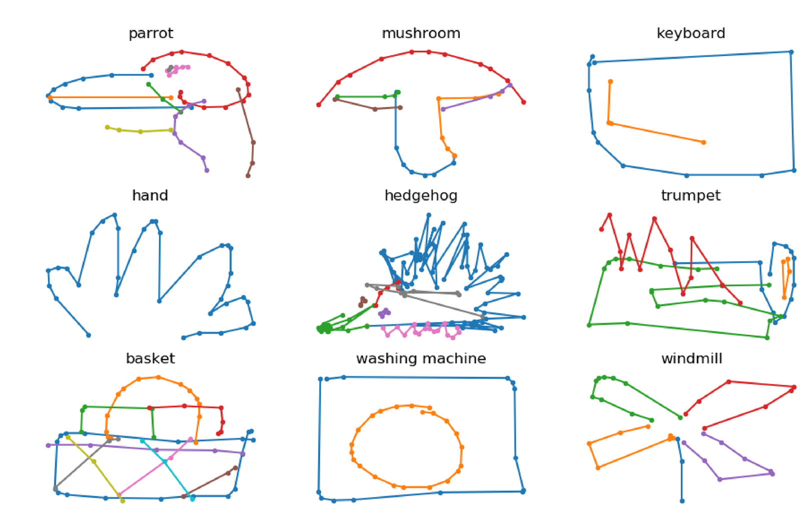
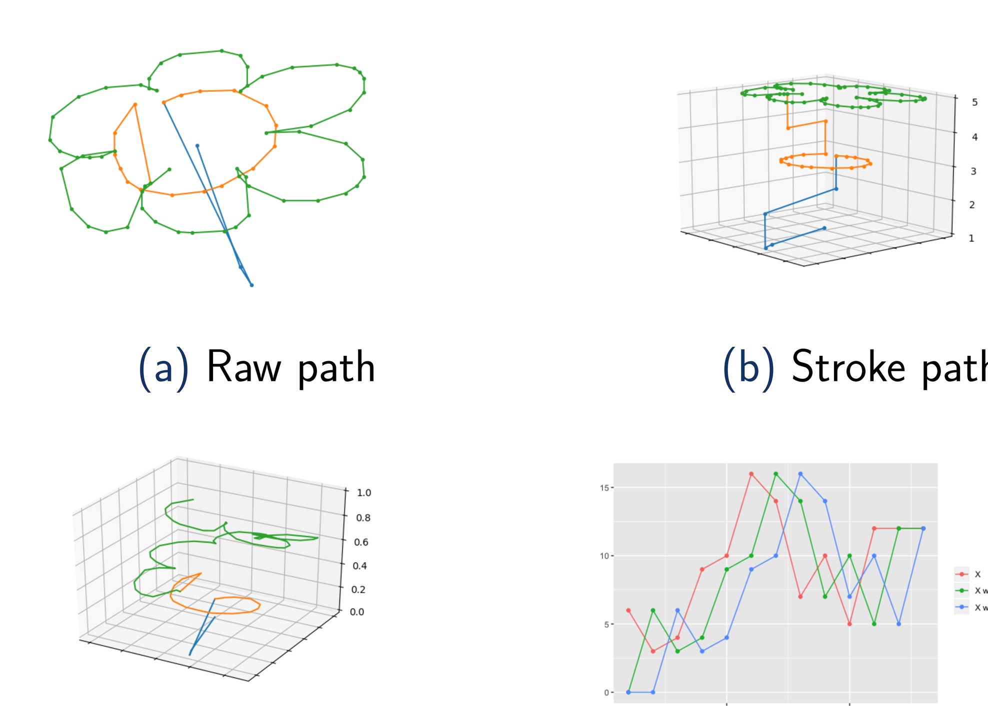


Figure: Quick, Draw! dataset [4].The task is to classify **340** different objects from 50 million samples.

Embedding results



(a) Raw path (b) Stroke path

(c) Time path (d) Lead-lag path

Figure: Different possible data embeddings.

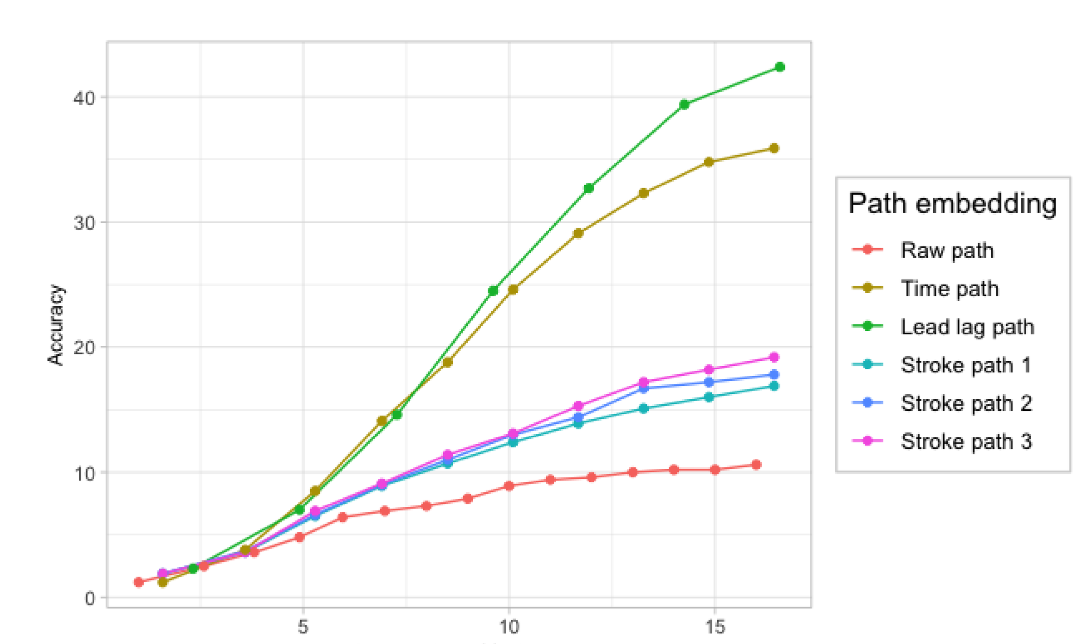


Figure: Prediction accuracy on the "Quick, Draw!" dataset with a small linear neural network with one hidden layer and different path embeddings.

References

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

[2] 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.

[3] 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.

[4] Quick, draw! dataset.
https://quickdraw.withgoogle.com/data.
Data made available by Google, Inc. under the Creative Commons Attribution 4.0 International license.