Chapter 4
Learning with Signatures

Gérard Biau and Adeline Fermanian

Abstract Sequential and temporal data arise in many fields of research, such as quantitative finance, medicine, or computer vision. The present article is concerned with a novel approach for sequential learning, called the signature method and rooted in rough path theory. Its basic principle is to represent multidimensional paths, i.e., functions from $[0, 1]$ to $\mathbb{R}^d$, by a graded feature set of their iterated integrals, called the signature. This approach relies critically on an embedding principle, which consists in representing discretely sampled data as continuous paths. After a survey of basic principles of signatures, we investigate the influence of embeddings on prediction accuracy with an in-depth study of recent and challenging datasets. We show that a specific embedding, called lead-lag, is systematically better, whatever the dataset or algorithm used.

4.1 Introduction

Sequential or temporal data occur in many fields of research, due to an increase in storage capacity and to the rise of machine learning techniques. Sequential data are characterized by the fact that each sample consists of an ordered array of values. Although the ordering often corresponds to time, it is not always the case. For example, text documents or DNA sequences have an intrinsic ordering, and can, therefore, be considered as sequential. Besides, when time is involved, several values can be recorded simultaneously, giving rise to an ordered array of vectors, which is, in the field of time series, often referred to as multidimensional time series. To name only a few domains, market evolution is described by financial time series,

Gérard Biau (✉)
Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, 4 place Jussieu, 75005 Paris, France, e-mail: gerard.biau@sorbonne-universite.fr

Adeline Fermanian
Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, 4 place Jussieu, 75005 Paris, France, e-mail: adeline.fermanian@sorbonne-universite.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*, Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_4

and physiological variables (e.g., electrocardiograms, electroencephalograms...) are recorded simultaneously in medicine, yielding multidimensional time series. We can also mention smartphone and GPS sensors data, or character recognition problems, where data has both a spatial and temporal aspect. These high-dimensional datasets open up new theoretical and practical challenges, as both statistical models and algorithms need to be adapted to their sequential nature.

Our goal in the present article is to discuss a novel approach for sequential learning, called the signature method, and coming from rough path theory. Its main idea is to summarize temporal (or functional) inputs by the graded feature set of their iterated integrals, the signature. Note that, in rough path theory, functions are referred to as paths, to insist on their geometrical aspects. Indeed, the importance of iterated integrals had been noticed by geometers in the 60s, as presented in the seminal work of [2]. It has been rediscovered by [10] in the context of stochastic analysis and controlled differential equations, and is at the heart of rough path theory. This theory, of which [11] and [5] give a recent account, focuses on developing a new notion of paths to make sense of evolving irregular systems. In this context, it has been shown that the signature provides an accurate summary of a path and allows to obtain arbitrarily good linear approximations of continuous functions of paths. Therefore, assuming we want to learn an output $Y \in \mathbb{R}$, which depends on a random path $X : [0,1] \to \mathbb{R}^d$, rough path theory suggests that the signature is a relevant feature set to describe X.

As can be expected, the signature has recently received the attention of the machine learning community and has achieved a series of successful applications. To cite some of them, [14] have achieved state-of-the-art results for handwriting recognition with a recurrent neural network combined with signature features. [7] have used the same approach for character recognition, and [8] have coupled Lasso with signature features for financial data streams classification. [1] have investigated its use for the detection of bipolar disorders, and [15] for human action recognition. For a gentle introduction to the signature method in machine learning, we refer the reader to [3].

However, despite many promising empirical successes, a lot of questions remain open, both practical and theoretical. In particular, to compute the signature, it is necessary to embed discretely sampled data points into paths. While authors use different approaches, this embedding is only mentioned in some articles, and rarely discussed. Thus, our purpose in this paper is to take a step forward in understanding how signature features should be constructed for machine learning tasks, with a special focus on the embedding step.

Our document is organized as follows. First, in Section 4.2, we give a brief exposition of the signature definition and properties. Then, we compare the predictive performance of different embeddings in Section 4.3. We emphasize that the embedding is as a crucial step as the algorithm choice since it can drastically change accuracy results. Moreover, we point out that one embedding, called lead-lag, performs systematically better than others, and this consistently over different datasets and learning algorithms.

4.2 Signature Definition and First Properties

We introduce in this section the notion of signature and review some of its important properties. The reader is referred to [11] or [5] for a more involved mathematical treatment with proofs. Throughout the article, our basic objects are paths, that is functions from $[0,1]$ to $\mathbb{R}^d$, where $d \in \mathbb{N}^*$. The main assumption is that these paths are of bounded variation, i.e., they have finite length.

Definition 1 Let

$$\begin{aligned} X : [0,1] &\longrightarrow \mathbb{R}^d \\ t &\longmapsto (X_t^1, \dots, X_t^d). \end{aligned}$$

The total variation of X is defined by

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

where the supremum is taken over all finite partitions

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

of $[0,1]$, and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. The path X is said to be of bounded variation if its total variation is finite.

The assumption of bounded variation allows to define Riemann-Stieljes integrals along paths. From now on, we assume that the integral of a continuous path $Y : [0,1] \to \mathbb{R}^d$ against a path of bounded variation $X : [0,1] \to \mathbb{R}^d$ is well-defined on any $[s,t] \subset [0,1]$, and denoted by

$$\int_s^t Y_u dX_u = \begin{pmatrix} \int_s^t Y_u^1 dX_u^1 \\ \vdots \\ \int_s^t Y_u^d dX_u^d \end{pmatrix} \in \mathbb{R}^d,$$

where $X = (X^1, \dots, X^d)$ and $Y = (Y^1, \dots, Y^d)$. We are now in a position to define the signature.

Definition 2 Let $X : [0,1] \to \mathbb{R}^d$ be a path of bounded variation, $I = (i_1, \dots, i_k) \subset \{1, \dots, d\}^k$, $k \in \mathbb{N}^*$, be a multi-index of length k, and $[s,t] \subset [0,1]$ be an interval. The signature coefficient of X corresponding to the multi-index I on $[s,t]$ is defined by

$$S^I(X)_{[s,t]} = \int_{s \leq u_1 < \cdots < u_k \leq t} \cdots \int dX_{u_1}^{i_1} \dots dX_{u_k}^{i_k}. \tag{4.1}$$

$S^I(X)_{[s,t]}$ is then said to be a signature coefficient of order k.

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

$$S(X)_{[s,t]} = \left(1, S^{(1)}(X)_{[s,t]}, \dots, S^{(d)}(X)_{[s,t]}, S^{(1,1)}(X)_{[s,t]}, S^{(1,2)}(X)_{[s,t]}, \dots\right).$$

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

$$S^K(X)_{[s,t]} = \left(1, S^{(1)}(X)_{[s,t]}, S^{(2)}(X)_{[s,t]}, \dots, S^{\overbrace{(d,\dots,d)}^{K}}(X)_{[s,t]}\right).$$

For simplicity, when $[s,t] = [0,1]$, we omit the interval in the notations, and, e.g., write $S^K(X)$ instead of $S^K(X)_{[0,1]}$. We note that, for a path in $\mathbb{R}^d$, there are d^k coefficients of order k. The signature truncated at order K is therefore a vector of dimension

$$\sum_{k=0}^{K} d^k = \frac{d^{K+1}-1}{d-1} \quad \text{if } d \neq 1,$$

and $K+1$ if $d = 1$. Unless otherwise stated, we assume that $d \neq 1$, as this is in practice usually the case. Thus, the size of $S^K(X)$ increases exponentially with K, and polynomially with d. Finally, it should be noted that, due to the ordering in the integration domain in (4.1), signature coefficients are not symmetric. For example, $S^{(1,2)}(X)$ is a priori not equal to $S^{(2,1)}(X)$.

A crucial feature of the signature is that it encodes geometric properties of the path. Indeed, it is clear that coefficients of order 2 correspond to some areas outlined by the path, as shown in Figure 4.1. For higher orders of truncation, the signature contains information about the joint evolution of tuples of coordinates. Furthermore, the signature possesses several properties that make it a good statistical summary of paths, as shown in the next three propositions.

Proposition 1 *Let $X : [0,1] \to \mathbb{R}^d$ be a path of bounded variation, and $\psi : [0,1] \to [0,1]$ be a non-decreasing surjection. Then, if $\widetilde{X}_t = X_{\psi(t)}$ is the reparametrization of X under ψ,*

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

In other words, the signature of a path is the same up to any reasonable time change. There is, therefore, no information about the path travel time in signature coefficients, which may be a useful feature in some applications. Nevertheless, when relevant for the problem at hand, it is possible to include this information by adding the time parametrization as a coordinate of the path. A second important property is a condition ensuring uniqueness of signatures.

Proposition 2 *If X has at least one monotonous coordinate, then $S(X)$ determines X uniquely.*

It should be noticed that having a monotonous coordinate is a sufficient condition, but a necessary one can be found in the monograph by [9], together with a proof of the proposition. The principal significance of this result is that it provides a practical procedure to guarantee signature uniqueness: it is sufficient to add a monotonous coordinate to the path X. For example, the time embedding mentioned above will

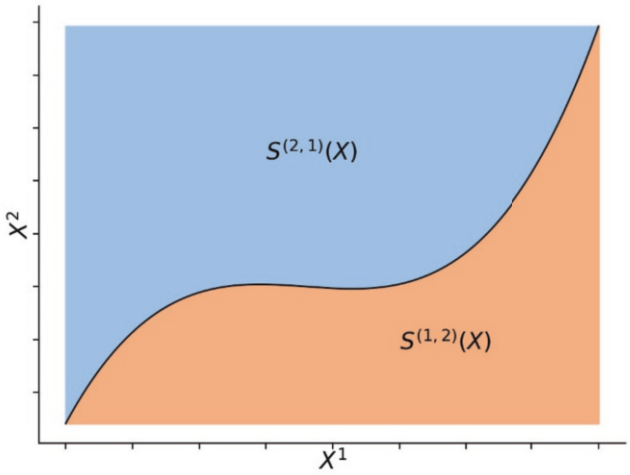

Fig. 4.1 Geometric interpretation of signature coefficients.

satisfy this condition. The next proposition reveals that the signature linearizes functions of X.

Proposition 3 *Let D be a compact subset of the space of bounded variation paths from $[0,1]$ to $\mathbb{R}^d$ that are not tree-like equivalent. Let $f : D \to \mathbb{R}$ be continuous. Then, for every $\epsilon > 0$, there exists $N \in \mathbb{N}^*$, $w \in \mathbb{R}^N$, such that, for any $X \in D$,*

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

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product on $\mathbb{R}^N$.

The notion of tree-like equivalence is closely related to the uniqueness of paths—the reader is referred to [9] for a definition. Proposition 3 is then a consequence of the Stone-Weierstrass theorem.

4.3 Embeddings

Now that we have presented the signature and its properties, we focus on its use in machine learning. In this context, we place ourselves in a statistical framework, and assume that our goal is to understand the relationship between a random input path $X : [0,1] \to \mathbb{R}^d$ and a random output $Y \in \mathbb{R}$. In a classical setting, we would be given a sample of independent and identically distributed (i.i.d.) observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, drawn from (X, Y). However, in applications, we only observe a realization X_i sampled at a discrete set of times $0 \leq t_1 < \cdots < t_{p_i} \leq 1$,

$p_i \in \mathbb{N}^*$. Therefore, we are given an i.i.d. sample $\{(\mathbf{x_1}, Y_1), \ldots, (\mathbf{x_n}, Y_n)\}$, where $\mathbf{x_i}$ takes the form of a matrix, i.e.,

$$\mathbf{x_i} = \begin{pmatrix} x_{i,1}^1 & \cdots & x_{i,p_i}^1 \\ \vdots & & \vdots \\ x_{i,1}^d & \cdots & x_{i,p_i}^d \end{pmatrix} \in \mathbb{R}^{d \times p_i}. \tag{4.2}$$

In this notation, $x_{i,j}^k$ denotes the kth coordinate of the ith sample observed at time t_j. If $d = 1$, we are in a classical setting of time series, where each observation is sampled in a finite number of points. However, d may here differ from 1, so we find ourselves in a more general situation where we want to learn from multidimensional time series. Moreover, it is worth noting the dependence of the number of sampled points p_i on i. In other words, each observation may have a different length. The signature dimension being independent of the number of sampled points, representing time series by their signature naturally handles inputs of various lengths, whereas traditional methods often require them to be normalized to a fixed length. To sum up, the signature method is appropriate for learning with discretely sampled multidimensional time series, possibly of different lengths.

To use signature features, one needs to embed the observations $\mathbf{x_i}$ into paths of bounded variation $X_i : [0, 1] \to \mathbb{R}^d$. Therefore, we need to choose an interpolation method, but, to ensure some properties such as signature uniqueness (see Proposition 2), we may also create new coordinates to the path and in this way increase the dimension of the embedding space. We refer the reader to [4] for a detailed description of the different embeddings that we use in the present article.

Our empirical study is based on three datasets of various nature. We present here the results on the Quick, Draw! dataset but similar results are obtained on two other datasets, described in [4]. The Quick, Draw! dataset has been made available by Google [6], and consists of drawing trajectories. It is made up of 50 million drawings, each drawing being a sequence of time-stamped pen stroke trajectories, divided into 340 categories. Some samples are shown in Figure 4.2.

We present in Figure 4.3 the results of our study on embedding performance, obtained with the following approach. Starting from the raw data, we first embed it into a continuous path, then compute its truncated signature, and use this vector as input for a learning algorithm. We want our findings to be independent of the data and the underlying statistical model so we use a range of different algorithms. The classification metric to assess prediction quality is the accuracy score. Then, to compare the quality of different embeddings, we plot the accuracy score against the log number of features, which yields one curve per embedding, where each point corresponds to a different truncation order. We then check whether one embedding curve is above the others, which would mean that, at equal input size, this embedding is better for learning.

A first striking fact is that some embeddings, namely the time and lead-lag, seem consistently better, whatever the algorithm used. It suggests that this performance is due to intrinsic theoretical properties of signatures and embeddings, not to domain-specific characteristics. The linear and rectilinear embeddings (red and pink curves),

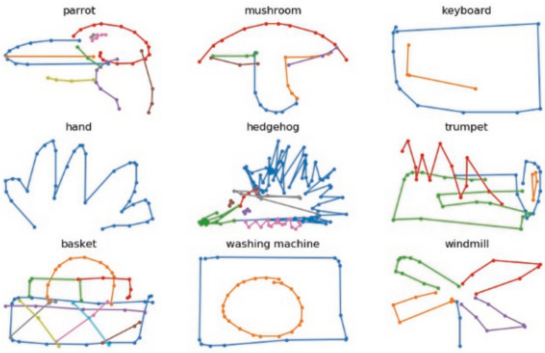

Fig. 4.2 9 samples from the Quick, Draw! dataset. Each color corresponds to a different pen stroke.

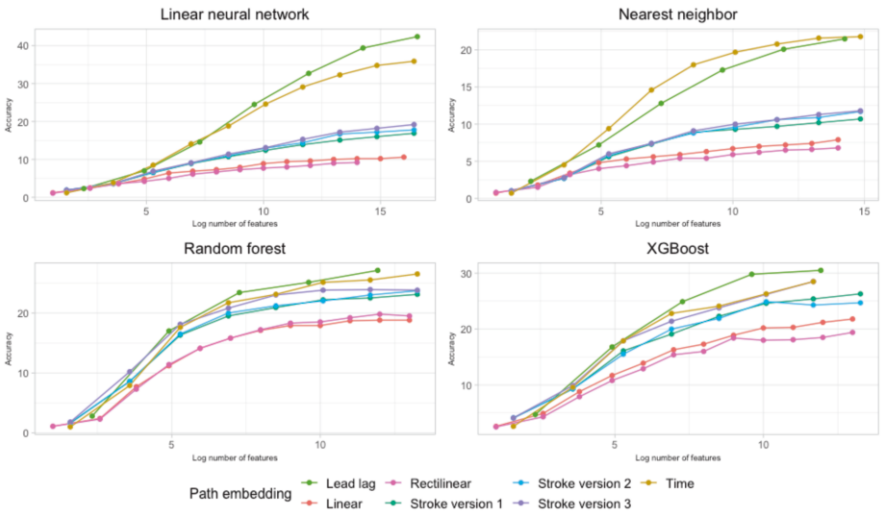

Fig. 4.3 Quick, Draw! dataset: prediction accuracy on the test set, for different algorithms and embeddings.

which are often used in the literature, appear to give the worst results. This bad performance can be explained by the fact that there is no guarantee that the signature characterizes paths when using the linear or rectilinear embeddings. Therefore, two different paths can have the same signature, without necessarily corresponding to the same class.

To conclude, the take-home message is that using the lead-lag embedding seems to be the best choice, regardless of the data and algorithm used. It does not cost anything computationally and can drastically improve prediction accuracy. Moreover,

the linear and stroke paths yield surprisingly poor results, despite their frequent use in the literature.

References

[1] Arribas, I.P., Goodwin, G.M., Geddes, J.R., Lyons, T., Saunders, K.E.: A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. Translational psychiatry **8**, 274 (2018)

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

[3] Chevyrev, I., Kormilitzin, A.: A primer on the signature method in machine learning. arXiv:1603.03788 (2016)

[4] Fermanian, A.: Embedding and learning with signatures. arXiv:1911.13211 (2019)

[5] Friz, P., Victoir, N.: Multidimensional Stochastic Processes as Rough Paths: Theory and Applications. Cambridge University Press, Cambridge (2010)

[6] Google Creative Lab: The Quick, Draw! Dataset. (2017) Available from https://github.com/googlecreativelab/quickdraw-dataset

[7] Graham, B.: Sparse arrays of signatures for online character recognition. arXiv:1308.0371 (2013)

[8] Gyurkó, L., Lyons, T., Kontkowski, M., Field, J.: Extracting information from the signature of a financial data stream. arXiv:1307.7244 (2014)

[9] Hambly, B., Lyons, T.: Uniqueness for the signature of a path of bounded variation and the reduced path group. Annals of Mathematics **171**, 109–167 (2010)

[10] Lyons, T.: Differential equations driven by rough signals. Revista Matemática Iberoamericana **14**, 215–310 (1998)

[11] Lyons, T., Caruana, M., Lévy, T.: Differential Equations Driven by Rough Paths. Springer, Berlin (2007)

[12] Malekzadeh, M., Clegg, R.G., Cavallaro, A., Haddadi, H.: Protecting sensory data against sensitive inferences. In: Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems, ACM (2018)

[13] Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM International Conference on Multimedia, 1041–1044. ACM (2014)

[14] Yang, W., Jin, L., Liu, M.: Deepwriterid: An end-to-end online text-independent writer identification system. IEEE Intelligent Systems **31**, 45–53 (2016)

[15] Yang, W., Lyons, T., Ni, H., Schmid, C., Jin, Li., Chang, J.: Leveraging the path signature for skeleton-based human action recognition. arXiv:1707.03993 (2017)