Algebraic aspects of signatures

Nikolas Tapia based on joint work with J. Diehl & K. Ebrahimi-Fard

FG6

Introduction

Classification of time series is a very active field of research.

Most methods rely on extraction of *features*.

Signatures[a,b] provide features that are interesting for a number of applications.

Also useful for other tasks such as analysing control systems, pathwise solutions to Stochastic Differential Equations, among others.

[a] K.-T. Chen. „Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula". In: *Ann. of Math. (2)* 65 (1957), pp. 163–178.
[b] T. Lyons. „Differential equations driven by rough signals". In: *Revista Matemática Iberoamericana* 14 (1998), pp. 215–310.

Continuous-time signatures

Let $X\colon [0,1] \longrightarrow \mathbb{R}$ **continuous** path.

Definition

Given $p \geq 1$, define the p-variation of X over the interval $[s,t] \subseteq [0,1]$ by

$$\|X\|_{p;[s,t]} := \left(\sup_{\pi \in \mathcal{P}[s,t]} \sum_{[u,v]\in\pi} |X_v - X_u|^p \right)^{1/p}.$$

The space of all paths such that $\|X\|_{p;[s,t]} < \infty$ is denoted by $\mathrm{V}^p([s,t])$.

Can be generalized to functions $\Xi\colon [0,1]^2 \longrightarrow \mathbb{R}$ by replacing the increment $X_v - X_u$ by $\Xi_{u,v}$.

This generalization is an essential piece in T. Lyon's theory of *rough paths*.[a]

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

Young integration

Theorem (Young[a])

Let $X \in \mathrm{V}^p$, $Y \in \mathrm{V}^q$ with $\frac{1}{p} + \frac{1}{q} > 1$. The integral

$$\int_s^t Y_u \,\mathrm{d}X_u := \lim_{|\pi|\to 0} \sum_{i=0}^{n(\pi)} Y_{t_i}(X_{t_{i+1}} - X_{t_i})$$

is well defined and

$$\left| \int_s^t Y_u \,\mathrm{d}X_u - Y_s(X_t - X_u) \right| \leq C_{p,q} \|X\|_{p;[s,t]} \|Y\|_{q;[s,t]}.$$

In particular, the iterated integral $\int X \,\mathrm{d}X$ is well defined as long as $X \in \mathrm{V}^p$ for $1 \leq p < 2$.

More generally, if $X = (X^1, \ldots, X^d)$ takes values in $\mathbb{R}^d$ then the integrals $\int X^i \,\mathrm{d}X^j$ are defined.

[a] L. C. Young. „An inequality of the Hölder type, connected with Stieltjes integration“. In: *Acta Mathematica* 67.1 (1936), p. 251.

Signatures

Definition

The signature of the path $X : [0,1] \longrightarrow \mathbb{R}^d$ is the collection of iterated integrals

$$S(X)_{s,t} := 1 + \int_s^t \mathrm{d}X_u + \int_s^t \int_s^u \mathrm{d}X_v \otimes \mathrm{d}X_u + \cdots + \int \cdots \int_{s<u_1<\cdots<u_n<t} \mathrm{d}X_{u_1} \otimes \cdots \otimes \mathrm{d}X_{u_n} + \cdots$$

Theorem

The signature satisfies

1. *Chen's identity:* $S(X)_{s,u} \otimes S(X)_{u,t} = S(X)_{s,t}$.
2. *Reparametrization invariance:* $S(X \circ \varphi)_{s,t} = S(X)_{s,t}$.
3. *It is the unique solution to the fixed-point equation*

$$S(X)_{s,t} = 1 + \int_s^t S(X)_{s,u} \otimes \mathrm{d}X_u.$$

Signatures

Additionally, the *shuffle relations* are satisfied:

$$S(X)^I_{s,t} S(X)^J_{s,t} = \sum_{K \in \mathrm{Sh}(I,J)} S(X)^K_{s,t}.$$

This introduces some redundancy, e.g.

$$S(X)^{ji}_{s,t} = S(X)^i_{s,t} S(X)^j_{s,t} - S(X)^{ij}_{s,t}.$$

A way to compress the available information is to work with the so-called *log-signaure*

$$\Omega(X)_{s,t} := \log_{\otimes} S(X)_{s,t} \in \mathcal{L}(\mathbb{R}^d).$$

$\Omega(X)$ corresponds to a *pre-Lie Magnus expansion* w.r.t. the pre-Lie product

$$X \triangleright Y := \int_s^t \int_s^u [\mathrm{d}X_v, \mathrm{d}Y_u]$$

The map $I \mapsto S(X)^I_{s,t}$ defines a linear map from the *tensor algebra* to the reals. The shuffle relations then mean that this map is a *character*, i.e.

$$S(X)^I_{s,t} S(X)^J_{s,t} = S(X)^{I \sqcup\!\sqcup J}_{s,t}$$

where the shuffle product is recursively defined, for $I = (i_1, \dots, i_n)$, $J = (j_1, \dots, j_m)$, by

$$I \sqcup\!\sqcup J = (I' \sqcup\!\sqcup J)i_n + (I \sqcup\!\sqcup J')j_m$$

where $I' = (i_1, \dots, i_{n-1})$, $J' = (j_1, \dots, j_{m-1})$ and

$$S(X)^I_{s,t} = \int_{s<u_1<\cdots<u_n<t} \cdots \int \mathrm{d}X^{i_1}_{u_1} \cdots \mathrm{d}X^{i_n}_{u_n}.$$

Remark

We can think of $S(X)$ as a formal series

$$S(X)_{s,t} = \sum_I S(X)^I_{s,t} I.$$

Signatures

Why should we care?

1. Useful for the description of the solutions of *controlled systems*: if $\dot{Y}_t = \mathrm{V}(Y_t)\dot{X}_t$ then

$$Y_t - Y_s = \sum_I \mathrm{V}_I(Y_s) S(X)^I_{s,t} + R_{s,t}.$$

2. Captures *features* of X, useful for applications to Machine Learning, pattern recognition, time series analysis, etc.

In principle hard to compute. However, if X is piecewise linear then

$$S(X)_{s,t} = \exp_\otimes(v_1) \otimes \cdots \otimes \exp_\otimes(v_k)$$

and we can use the Baker–Campbell–Hausdorff formula.[a]

[a] J. Reizenstein and B. Graham. „The iisignature library: efficient calculation of iterated-integral signatures and log signatures“. In: (2018). arXiv: 1802.08252 [cs.DS].

However:

1. For a one-dimensional signal:

$$\int_{s<u_1<\cdots<u_n<t} \cdots \int dX_{u_1} \cdots dX_{u_n} = \frac{(X_t - X_s)^n}{n!}.$$

   This can be cured to some extent by introducing more dimensions[a] and other tricks[b].
2. In practice we are confronted with discrete data.

   This can also be avoided by *interpolation*.
3. A more severe problem is *tree-like equivalence*.[c]

We propose instead a new framework operating directly at the discrete level.

[a] T. Lyons and H. Oberhauser. „Sketching the order of events“. In: (2017). arXiv: 1708.09708 `[stat.ML]`.

[b] F. J. Király and H. Oberhauser. „Kernels for sequentially ordered data“. In: *Journal of Machine Learning Research* 20.31 (2019), pp. 1–45.

[c] B. Hambly and T. Lyons. „Uniqueness for the signature of a path of bounded variation and the reduced path group“. In: *Ann. Math.* 171.1 (Mar. 2010), pp. 109–167.

Discrete signatures

A composition of an integer n is a sequence $(i_1, \dots, i_k)$ with $i_1 + \cdots + i_k = n$.

Definition (Gessel[a])

Given a composition $I = (i_1, \dots, i_k)$ *define*

$$M_I(z) := \sum_{j_1 < j_2 < \ldots < j_k} z_{j_1}^{i_1} \cdots z_{j_k}^{i_k}.$$

For example

$$M_{(1)}(z) = \sum_j z_j, \quad M_{(1,1)} = \sum_{j_1 < j_2} z_{j_1} z_{j_2}, \quad M_{(2)}(z) = \sum_j z_j^2.$$

Note that

$$M_{(1)}(z)^2 = M_{(2)}(z) + 2M_{(1,1)}.$$

[a] I. M. Gessel. „Multipartite P-partitions and inner products of skew Schur functions“. In: *Combinatorics and algebra (Boulder, Colo., 1983)*. Vol. 34. Contemp. Math. Amer. Math. Soc., Providence, RI, 1984, pp. 289–317.

Discrete signatures

The map $\mathcal{I} \mapsto M_{\mathcal{I}}(z)$ defines a linear map from compositions to the reals. The product rule above can be expressed as

$$M_{\mathcal{I}}(z)M_{\mathcal{J}}(z) = M_{\mathcal{I}\star\mathcal{J}}(z)$$

where the *quasi-shuffle product*[a] $\star$ is recursively defined, for $\mathcal{I} = (i_1, \dots, i_n)$, $\mathcal{J} = (j_1, \dots, j_m)$, by

$$\mathcal{I} \star \mathcal{J} = (\mathcal{I}' \star \mathcal{J})i_n + (\mathcal{I} \star \mathcal{J}')j_m + c(\mathcal{I}, \mathcal{J})$$

where $\mathcal{I}'$ and $\mathcal{J}'$ are defined as before, and

$$c(\mathcal{I}, \mathcal{J}) := (i_1, \dots, i_{n-1}, j_1, \dots, j_{m-1}, i_n + j_m).$$

Definition

Given a discrete time series $x = (x_0, x_1, \dots, x_N)$, its discrete signature is

$$DS(x)_{n,m} = \sum_{\mathcal{I}} M_{\mathcal{I}}(\Delta_n^m x)\mathcal{I}$$

where $\Delta_n^m x = (x_{n+1} - x_n, \dots, x_m - x_{m-1})$.

[a] M. E. Hoffman. „Quasi-shuffle products“. In: *J. Algebraic Combin.* 11.1 (2000), pp. 49–68.

Discrete signatures

Why should we care?

1. Can be used to analyse solutions of *controlled recurrence equations* of the form

$$y_{k+1} = y_k + V(y_k)(x_{k+1} - x_k)$$

relevant e.g. for applications to Residual Neural Networks.[a,b,c]

2. Invariant under *time warping*, useful for applications to time series classification.[d]
3. No need to transform the data in any way. Even if x is one-dimensional we get more information, e.g.

$$DS(x)^{(2)}_{0,N} = \sum_{j=1}^{N} (x_j - x_{j-1})^2 \neq (x_N - x_0)^2.$$

4. No need for BCH formula.

[a] Current project with P. Friz (TU) and C. Bayer (WIAS)
[b] K. He et al. „Deep Residual Learning for Image Recognition“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
[c] E. Haber and L. Ruthotto. „Stable architectures for deep neural networks“. In: *Inverse Problems* 34.1 (2018), pp. 014004, 22.
[d] J. Diehl, K. Ebrahimi-Fard, and N. Tapia. „Time warping invariants of multidimensional time series“. In: (2019). arXiv: 1906.05823 [math.RA].

Discrete signatures

We can actually count the number of invariants.

Theorem (Diehl, Ebrahimi-Fard, T.; Novelli, Thibon[a])

The number of time-warping invariants of a d-dimensional time series has generating function

$$G(t) := \sum_{n=0}^{\infty} c_n(d) t^n = \frac{(1-t)^d}{2(1-t)^d - 1} = 1 + dt + \frac{d(3d+1)}{2} t^2 + \frac{d(13d^2+9d+2)}{6} t^3 + \cdots .$$

Compare with the corresponding generating function for the shuffle algebra:

$$H(t) = \frac{1}{1-dt} = 1 + dt + d^2 t^2 + d^3 t^3 + \cdots .$$

[a] J.-C. Novelli and J.-Y. Thibon. „Free quasi-symmetric functions and descent algebras for wreath products, and noncommutative multi-symmetric functions“. In: *Discrete Math.* 310.24 (2010), pp. 3584–3606.

Discrete signatures

Theorem (Diehl, Ebrahimi-Fard, T.)

Let x be a time series and define an infinite-dimensional path $X = (X^I)$ where for a composition I, X^I is the linear interpolation of the sequence

$$n \mapsto M_I(\Delta_0^n x).$$

Then

$$S(X)_{0,N}^{I} = DS(x)_{0,N}^{\Phi(I)}$$

where Φ is Hoffman's isomorphism[a].

In the one-dimensional case, the *elementary symmetric functions*

$$M_{(1,1,\dots,1)}(\Delta x) = \sum_{j_1 < \cdots < j_n} \Delta x_{j_1} \cdots \Delta x_{j_n}$$

arise as a left-point Riemann sum associated to a *piecewise constant* interpolation of x.

[a] M. E. Hoffman. „Quasi-shuffle products“. In: *J. Algebraic Combin.* 11.1 (2000), pp. 49–68.

Perspectives

A few possible extensions:

1. Multi-parameter data, e.g. one-dimensional time series depending on two parameters.
2. General functions on increments, e.g.

$$\sum_{j_1<j_2} f_{i_1}(\Delta x_{j_1}) f_{i_2}(\Delta x_{j_2}).$$

And some questions and ongoing projects:

1. Understanding $\log DS(x)$. Chow's theorem.
2. Numerical experiments and use in time warping.[a]
3. Robustness of Residual Neural Networks.
4. Learning dynamics of Stochastic Differential Equations.[b,c]

[a] B. J. Jain. „Making the dynamic time warping distance warping-invariant". In: *Pattern Recognition* 94 (2019), pp. 35–52.
[b] with C. Bayer and M. Eigel (WIAS)
[c] W. S. Gray, G. S. Venkatesh, and L. A. D. Espinosa. „Combining Learning and Model Based Control via Discrete-Time Chen-Fliess Series". In: (2019). arXiv: 1906.11084 [eess.SY].

Thanks!