Computing the full signature kernel as the solution of a Goursat problem

Thomas Cass Terry Lyons Cristopher Salvi Weixin Yang

DataSig - Imperial College, University College, University of Oxford, Alan Turing Institute

2nd July 2020

Representing functions from data

A basic problem in data science is to model real valued functions on sets from data. To transition from pairs $(x_i, y_i) \in X \times \mathbb{R}$ to $f \in C(X)$. It is a deep and substantial topic but still it is worth looking for a moment at the abstract foundations. One core step is to represent the data

- A feature map is
  - A bijective representation ϕ of a set X onto $K \subset E$ where E is a linear space.
  - A feature map is universal if ϕ can approximately linearise functions!

$$\overline{E^*|_K} = C(K)$$

  - For example K three points in 2 dimensions; product of linear functions is linear!
  - Theory of Choquet simplexes

Kernels

Kernels form a useful tool in machine learning, they offer concrete approaches to resolving the core problem, and also can be very useful in reducing the dimension and complexity of calculations

- What is a kernel?
  - A set X embedded into E and E^*

$$\begin{aligned} \phi &: X \hookrightarrow E \\ \psi &: X \hookrightarrow E^* \end{aligned}$$

    then one gets $K(x, x') := \langle \phi(x), \psi(x') \rangle$ and conversely.
  - Amari's statistical manifolds of probability measures all have this property

$$\begin{aligned} P_{\ll \mu} &\to L_p(\mu) \\ p d\mu &\to p^{1/p} \end{aligned}$$

  - Sometimes a Hilbert space (e.g. $p = 2$ in Amari)

Kernel based regression

Although there are many issues, one cannot deny the convenience of kernels. For every data point one gets a new function on points: $K(x, x') := \langle \phi(x), \psi(x') \rangle$.

- Kernels are useful for regression and machine learning
  - Some real observational data $(x_j, y_j)_{j=1...N}$ then solve this system

$$y_j = \sum_{k=1...N} \lambda_k K(x_j, x_k)$$

    to express
  - the observed function in terms of the kernel functions $\psi(x_i) \in E^*$.

$$F(\cdot) := \sum_{k=1...N} \lambda_k K(\cdot, x_k)$$

  - Crucially this calculation only depends on the N^2 numbers $K(x_j, x_k)$ and does not need the embedding - E can be infinite dimensional.

Inner products

It is not clear that kernels should induce inner products, but when they do, one can use Gaussian tricks. Symmetric kernels are an essential component in algorithms for pattern analysis (Bishop, 1995; Hastie et al., 2001; Scholköpf and Smola, 2002)

- suppose we have a symmetric and universal kernel
  - Consider a gaussian random variable X' on E^* with co-variance given by the inner product on E .

$$\mathbb{E}\left[X'(x_1)\,X'(x_2)\right] = < x_1, x_2 >, \qquad x_1, x_2 \in E$$

  - A measure on functions (images).
  - Can sample from the conditional distribution of X' given the evaluations $(x_j, y_j)_{j=1\ldots N}$.
  - Allows interpolation - get a random function defined everywhere and consistent with the monochrome data.

Tensor algebras

- Let E be d-dimensional Banach space with basis $\mathcal{E} = \{e_1, \dots, e_d\}$. Denote by

$$T(E) = \bigoplus_{k=0}^{\infty} E^{\otimes k}$$

and

$$T((E)) = \prod_{k=0}^{\infty} E^{\otimes k}$$

the spaces of formal polynomials, power series in the letters from $\mathcal{E}$.

  - $T(\,)$ and $T((\,))$ are functors from the category of vector spaces to that of algebras
  - moreover $T(E^*) \subset T((E))^*$.

From alphabets to Hilbert spaces

The basis $\mathcal{E} = \{e_1, \dots, e_d\} \subset E$ induces structure on $T((E))$.

- A basis for $E^{\otimes n}$ is the words of length n with letters drawn from the alphabet $\mathcal{E}$:

$$\{e_K = e_{k_1} \otimes \dots \otimes e_{k_n}\}_{K=(k_1,\dots,k_n)\in\{1,\dots,d\}^n}$$

- The choice of $\mathcal{E}$ induces an inner product on E, $E^{\otimes n}$, $T(E)$, $T((E))$.

$$\langle e_{i_1} \otimes \dots \otimes e_{i_n}, e_{j_1} \otimes \dots \otimes e_{j_n} \rangle = \delta_{i_1,j_1} \dots \delta_{i_n,j_n}, \qquad \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

- Making the canonical projection $\pi_n : T = (T^0, T^1, \dots, T^n, \dots) \to T^n \in E^{\otimes n}$ orthogonal.

- The inner product is
  - Defined for $A, B \in T((E))$ as

$$\langle A, B \rangle = \sum_{n=0}^{\infty} \langle \pi_n(A), \pi_n(B) \rangle$$

The signature as a universal feature set for unparameterised paths

- Suppose that x is a finite stream of information defined on $[s,t]$ with features in the vector space E
  - The signature is the solution of the universal differential equation driven by x

$$\begin{aligned} dS(x)_{s,u} &= S(x)_{s,u} \otimes dx_u \ , & S(x)_{s,s} &= 1 = (1,0,0,\ldots) \\ S(x)_{s,t} &= 1 + \int_s^t S(x)_{s,u} \otimes dx_u, & S(x)_{s,s} &= 1 = (1,0,0,\ldots) \end{aligned}$$

  - Informs about the stream $x|_{[s,t]}$ through the response $S(x)_{s,t}$ of the exponential nonlinear system. (*meaning without maths*).
  - If E is the formal span of a finite alphabet $A = \{a_1,\ldots,a_n\}$ then $S \in T((E))$ the space of infinite formal linear combinations of words with letters drawn from A. The solution lives in a vector space. S is a feature map!
  - S does not depend on the parameterisation of the path segment. It is a powerful nonlinear filter that removes sampling rate from data and faithfully preserves the order of events, the curve. Hambly, Lyons 2010

Signature Kernels

Signatures exist over any interval and (for rough paths) decay factorially.

- Suppose that x, $\mathbb{X}$ are rough paths/streams of finite 1 and p variation over $[s,t]$ and controlled by w_x, $w_{\mathbb{X}}$;Let $(s_1,s_2)\subset[s,t]$
  - then x, $\mathbb{X}$ have signatures:

$$(s_1,s_2)\mapsto S(x_{s_1,s_2})=(1,x^1_{s_1,s_2},\dots,\dots,\dots,x^m_{s_1,s_2},\dots)\in T((E))$$

$$(s_1,s_2)\mapsto S(\mathbb{X}_{s_1,s_2})=(1,\mathbb{X}^1_{s_1,s_2},\dots,\mathbb{X}^{\lfloor p\rfloor}_{s_1,s_2},\dots,\mathbb{X}^m_{s_1,s_2},\dots)\in T((E))$$

  - they are multiplicative functionals
  - and they have factorial decay (neoclassical inequality)

$$\begin{aligned}||x^k_{s_1,s_2}||_{E^{\otimes k}} &\leq \frac{\omega_x(s_1,s_2)}{k!},\quad \forall(s_1,s_2)\in\Delta_I\\ ||\mathbb{X}^k_{s_1,s_2}||_{E^{\otimes k}} &\leq \frac{\omega_{\mathbb{X}}(s_1,s_2)^{k/p}}{\beta_p(k/p)!},\quad \forall(s_1,s_2)\in\Delta_I\end{aligned}$$

- For any two unparameterised rough paths $\mathbb{X}$ and $\mathbb{Y}$ the signature kernel $K(\mathbb{X},\mathbb{Y}):=\langle S(\mathbb{X}_I),S(\mathbb{Y}_J)\rangle$ is always well defined.

Oberhauser and Kiraly

- In their paper *Kernels for Sequentially Ordered Data* (Franz J. Kiraly, Harald Oberhauser; JMLR 20(31):1-45, 2019) Franz and Harald observed that a kernel on the space carrying the data always implies a kernel on truncated signatures of sequences in the implied linear space, and importantly
  - they use dynamic programming and low-rank techniques to demonstrate that for the truncated signature kernel and bounded variation paths there were efficient algorithms to compute this truncated kernel.
  - Harald, with his student Csaba, *Bayesian Learning from Sequential Data using Gaussian Processes with Signature Covariances*. (Toth, Csaba, and Harald Oberhauser., ICML in press (2020)) explored the practical ramifications of this kernel in a range of practical contexts and demonstrate its value.

Our work - the Goursat PDE

It is easy to see that, unlike the truncated signature, the full signature kernel is universal, but calculating it directly would seem to involve infinite series of integrals with exponentially increasing numbers of terms.

- This talk is based on *Computing the full signature kernel as the solution of a Goursat problem (*Thomas Cass, Terry Lyons, Cristopher Salvi, Weixin Yang https://arxiv.org/abs/2006.14794)
  - We derive a PDE. evaluating the *full* signature kernel between two unparameterised paths
  - In the paper we further establish the PDE is well defined, has numerics and solutions for any rough streams. It is an interesting "S" PDE!

The basic statement

Theorem

Let $I = [u, u']$ and $J = [v, v']$ be two closed time intervals and let $x \in C^1(I, E)$ and $y \in C^1(J, E)$. Consider the bilinear form $k_{x,y} : I \times J \to \mathbb{R}$ defined as follows

$$k_{x,y} : (s, t) \mapsto \langle S(x)_s, S(y)_t \rangle$$

then $k_{x,y}$ is a solution of the following linear hyperbolic PDE

$$\frac{\partial^2 k_{x,y}}{\partial s \partial t} = \langle \dot{x}_s, \dot{y}_t \rangle k_{x,y}$$

with initial conditions $k_{x,y}(u, \cdot) = k_{x,y}(\cdot, v) = 1$ and where $\dot{x}_s = \frac{dx_p}{dp}|_{p=s}$ and $\dot{y}_t = \frac{dx_q}{dq}|_{q=t}$.

The derivation

Clearly, for any $t \in J$ we have $k_{x,y}(u,t) = \langle S(x)_{uu}, S(y)_{vt} \rangle = 1$ and for any $s \in I$, $k_{x,y}(s,v) = 1$

$$
\begin{aligned}
k_{x,y}(s,t) &= \langle S(x)_{us}, S(y)_{vt} \rangle \\
&= \langle 1 + \int_{p=u}^{s} S(x)_{up} \otimes dx_p,\ 1 + \int_{q=v}^{t} S(y)_{vq} \otimes dy_q \rangle \\
&= 1 + \langle \int_{p=u}^{s} S(x)_{up} \otimes \dot{x}_p dp,\ \int_{q=v}^{t} S(y)_{vq} \otimes \dot{y}_q dq \rangle \\
&= 1 + \int_{p=u}^{s} \int_{q=v}^{t} \langle S(x)_{up} \otimes \dot{x}_p,\ S(y)_{vq} \otimes \dot{y}_q \rangle dpdq \\
&= 1 + \int_{p=u}^{s} \int_{q=v}^{t} \langle S(x)_{up}, S(y)_{vq} \rangle \langle \dot{x}_p, \dot{y}_q \rangle dpdq \\
&= 1 + \int_{p=u}^{s} \int_{q=v}^{t} \langle S(x)_{up}, S(y)_{vq} \rangle \langle \dot{x}_p, \dot{y}_q \rangle dpdq \\
&= 1 + \int_{p=u}^{s} \int_{q=v}^{t} k_{x,y}(p,q) \langle \dot{x}_p, \dot{y}_q \rangle dpdq
\end{aligned}
$$

The Equation continued

By the fundamental theorem of calculus we can differentiate

$$1 + \int_{p=u}^{s} \int_{q=v}^{t} k_{x,y}(p,q) \langle \dot{x}_p, \dot{y}_q \rangle dpdq$$

firstly with respect to s

$$\frac{\partial k_{x,y}(s,t)}{\partial s} = \int_{q=v}^{t} k_{x,y}(s,q) \langle \dot{x}_s, \dot{y}_q \rangle dq$$

and then with respect to t to obtain the desired linear hyperbolic PDE

$$\frac{\partial^2 k_{x,y}(s,t)}{\partial s \partial t} = \langle \dot{x}_s, \dot{y}_t \rangle k_{x,y}(s,t)$$

Goursat

Theorem (Lees 1960 (Goursat 1916) Theorems 2 & 4)

Let $\sigma : I \to \mathbb{R}$ and $\tau : J \to \mathbb{R}$ be two absolutely continuous functions whose first derivatives are square integrable and such that $\sigma(u) = \tau(v)$. Let $C_1, C_2, C_3 : \mathcal{D} \to \mathbb{R}$ be a bounded and measurable over $\mathcal{D}$ and $C_4 : \mathcal{D} \to \mathbb{R}$ be square integrable. Then there exists a unique function $u : \mathcal{D} \to \mathbb{R}$ such that $u(s, v) = \sigma(s), u(u, t) = \tau(t)$ and (almost everywhere on $\mathcal{D}$)

$$\frac{\partial^2 u}{\partial s \partial t} = C_1(s,t)\frac{\partial u}{\partial s} + C_2(s,t)\frac{\partial u}{\partial t} + C_3(s,t)u + C_4(s,t)$$

If in addition $C_i \in C^{p-1}(\mathcal{D})$ $(i = 1, 2, 3, 4)$ and σ and τ are C^p, then the unique solution $u : \mathcal{D} \to \mathbb{R}$ of the Goursat problem is of class C^p.

Set $C_1 = C_2 = C_4 = 0$ and $C_3(s,t) = \langle \dot{x}_s, \dot{y}_t \rangle$ .If the two input paths x, y are C^p then their derivatives will be of class C^{p-1} and therefore the solution $k_{x,y}$ will be of class C^p. Finite difference approximation works.

The numerics

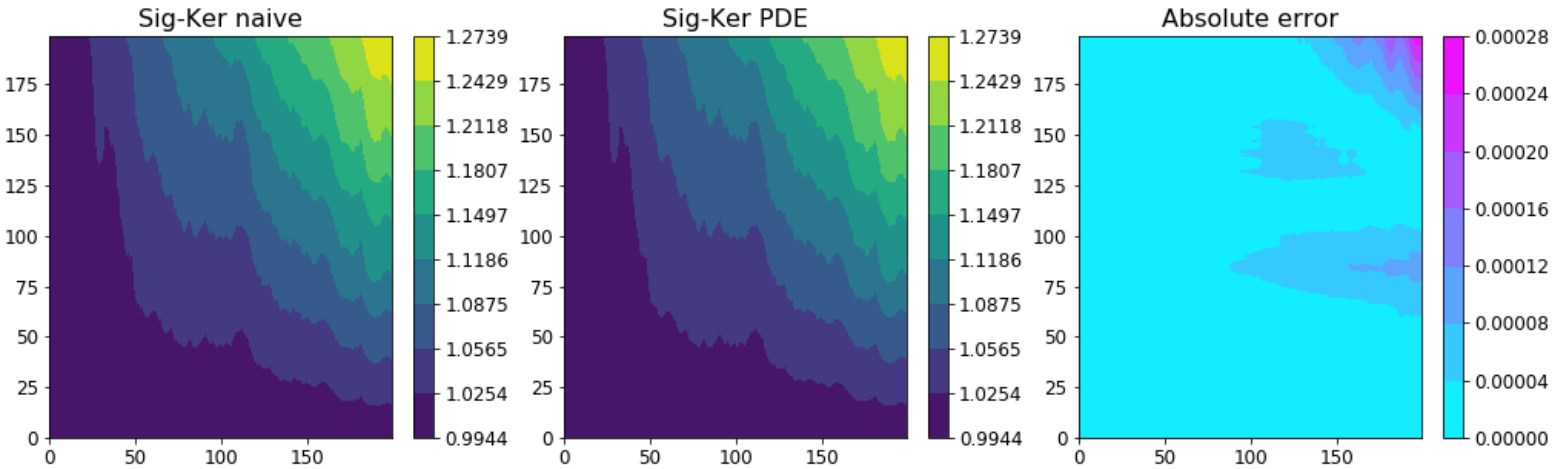

Figure: Example of error distribution of $k_{x,y}(s,t)$ on the whole grid $(s,t) \in \mathcal{D}$.

The analysis

The equation in the bounded variation case is already useful.

- A challenge is to make sense of
  - the equation when paths are rougher:

$$\partial^2 k_{x,y}(s,t) = \langle dx_s, dy_t \rangle k_{x,y}(s,t)$$

  - Since k makes sense it is reasonable that one can.
  - One successful approach (see the paper) is
    - to make a common parameterisation of x and y
    - use the extension theorem to add the cross integrals to make x and y jointly a rough path
    - solve the second order rough ode
    - show that the solution was independent of the choice of extension.

The end

Thank you!