Distribution Regression for Sequential Data

Maud Lemercier[1], Cristopher Salvi[2], Theodoros Damoulas[1], Edwin V. Bonilla[3], and Terry Lyons[2]

[1]University of Warwick & Alan Turing Institute
{maud.lemercier, t.damoulas}@warwick.ac.uk
[2]University of Oxford & Alan Turing Institute
{salvi, lyons}@maths.ox.ac.uk
[3]CSIRO's Data61, edwin.bonilla@data61.csiro.au

Abstract

Distribution regression refers to the supervised learning problem where labels are only available for groups of inputs instead of individual inputs. In this paper, we develop a rigorous mathematical framework for distribution regression where inputs are complex data streams. Leveraging properties of the *expected signature* and a recent *signature kernel trick* for sequential data from stochastic analysis, we introduce two new learning techniques, one feature-based and the other kernel-based. Each is suited to a different data regime in terms of the number of data streams and the dimensionality of the individual streams. We provide theoretical results on the universality of both approaches and demonstrate empirically their robustness to irregularly sampled multivariate time-series, achieving state-of-the-art performance on both synthetic and real-world examples from thermodynamics, mathematical finance and agricultural science.

1 INTRODUCTION

Distribution regression (DR) on sequential data describes the task of learning a function from a group of data streams to a single scalar target. For instance, in thermodynamics (Fig. 1) one may be interested in determining the temperature of a gas from the set of trajectories described by its particles (Hill, 1986; Reichl, 1999; Schrödinger, 1989). Similarly in quantitative finance practitioners may wish to estimate mean-reversion parameters from observed market dynamics (Papavasiliou et al., 2011; Gatheral et al., 2018; Balvers et al., 2000). Another example arises in agricultural science where the challenge consists in predicting the overall end-of-year crop yield from high-resolution climatic data across a field (Panda et al., 2010; Dahikar and Rode, 2014; You et al., 2017).

DR techniques (Póczos et al., 2013; Oliva et al., 2014; Szabó et al., 2016) have been successfully applied to handle situations where the inputs in each group are vectors in $\mathbb{R}^d$. Recently, there has been an increased interest in extending these techniques to non-standard inputs such as images (Law et al., 2018b) or persistence diagrams (Kusano et al., 2016). However DR for sequential data, such as multivariate time-series, has been largely ignored. The main challenges in this direction are the non-exchangeability of the points in a sequence, which naturally come with an order, and the fact that in many real world scenarios the points in a sequence are irregularly distributed across time.

In this paper we propose a framework for DR that addresses precisely the setting where the inputs within each group are complex data streams, mathematically thought of as *Lipschitz continuous paths* (Sec. 2). We formulate two distinct approaches, one feature-based and the other kernel-based, both relying on a recent tool from stochastic analysis known as the *expected signature* (Chevyrev and Oberhauser, 2018; Chevyrev et al., 2016; Lyons et al., 2015; Ni, 2012). Firstly, we construct a new set of features that are universal in the sense that any continuous function on distributions on paths can be uniformly well-approximated by a linear combination of these features (Sec. 3.1). Secondly, we introduce a universal kernel on distributions on paths given by the composition of the expected signature and a Gaussian kernel (Sec. 3.2), which can be

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

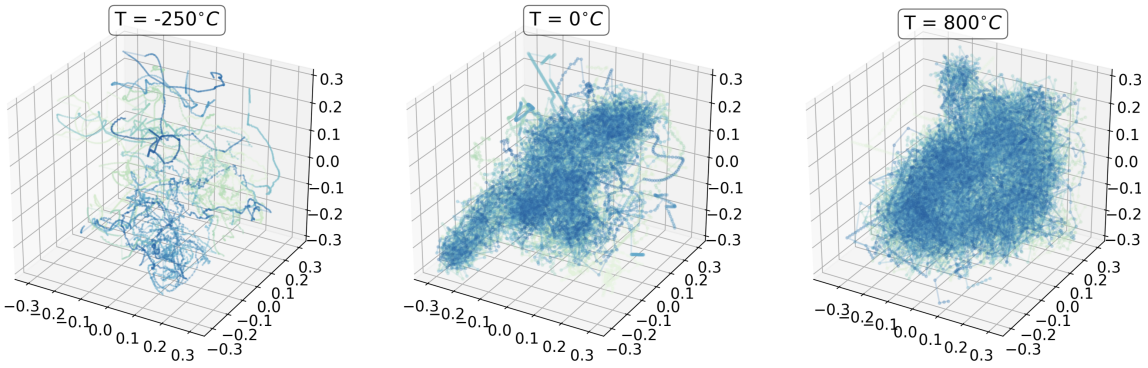

Figure 1: Simulation of the trajectories traced by 20 particles of an ideal gas in a 3-d box under different thermodynamic conditions. Higher temperatures equate to a higher internal energy in the system which increases the number of collisions resulting in different large-scale dynamics of the gas.

evaluated with a kernel trick. The former method is more suitable to datasets containing a large number of low dimensional streams, whilst the latter is better for datasets with a low number of high dimensional streams. We demonstrate the versatility of our methods to handle interacting trajectories like the ones in Fig. 1. We show how these two methods can be used to provide practical DR algorithms for time-series, which are robust to irregular sampling and achieve state-of-the-art performance on synthetic and real-world DR examples (Sec. 5).

1.1 Problem definition

Consider M input-output pairs $\{(\{\mathbf{x}^{i,p}\}_{p=1}^{N_i}, y^i)\}_{i=1}^{M}$, where each pair is given by a scalar target $y^i \in \mathbb{R}$ and a group of N_i d-dimensional time-series of the form

$$\mathbf{x}^{i,p} = \{(t_1, \mathbf{x}_1^{i,p}), \dots, (t_{\ell_{i,p}}, \mathbf{x}_{\ell_{i,p}}^{i,p})\}, \tag{1}$$

of possibly unequal lengths $\ell_{i,p} \in \mathbb{N}$, with time-stamps $t_1 < \ldots < t_{\ell_{i,p}}$ and values $\mathbf{x}_k^{i,p} \in \mathbb{R}^d$. Every d-dimensional time-series $\mathbf{x}^{i,p}$ can be naturally embedded into a Lipschitz-continuous path

$$x^{i,p} : [t_1, t_{l_{i,p}}] \to \mathbb{R}^d, \tag{2}$$

by piecewise linear interpolation with knots at $t_1, \dots, t_{l_{i,p}}$ such that $x_{t_k}^{i,p} = \mathbf{x}_k^{i,p}$. After having formally introduced a set of probability measures on this class of paths, we will summarize the information on each set $\{x^{i,p}\}_{p=1}^{N_i}$ by the *empirical measure* $\delta^i = \frac{1}{N_i}\sum_{p=1}^{N_i} \delta_{x^{i,p}}$ where $\delta_{x^{i,p}}$ is the *Dirac measure* centred at the path $x^{i,p}$. The supervised learning problem we propose to solve consists in learning an unknown function $F : \delta^i \mapsto y^i$.

2 THEORETICAL BACKGROUND

We begin by formally introducing the class of paths and the set of probability measures we are considering.

2.1 Paths and probability measures on paths

Let $0 \leq a < T$ and $I = [a, T]$ be a closed time interval. Let E be a Banach space of dimension $d \in \mathbb{N}$ (possibly infinite) with norm $\|\cdot\|_E$. For applications we will take $E := \mathbb{R}^d$. We denote by $\mathcal{C}(I, E)$ the Banach space (Friz and Victoir, 2010) of Lipschitz-continuous functions $x : I \to E$ equipped with the norm

$$\|x\|_{Lip} = \|x_a\| + \sup_{s,t \in I} \frac{\|x_t - x_s\|}{|t - s|}. \tag{3}$$

We will refer to any element $x \in \mathcal{C}(I, E)$ as an E-valued *path*.[1] Given a compact subset of paths $\mathcal{X} \subset \mathcal{C}(I, E)$, with respect to the topology induced by $\|\cdot\|_{Lip}$, we denote by $\mathcal{P}(\mathcal{X})$ the set of (Borel) *probability measures* on $\mathcal{X}$.

The signature has been shown to be an ideal feature map for paths (Lyons, 2014). Analogously, the expected signature is an appropriate feature map for probability measures on paths. Both feature maps take values in the same feature space. In the next section we introduce the necessary mathematical background to describe the structure of this space.

2.2 A canonical Hilbert space of tensors $\mathcal{T}(E)$

In what follows $\oplus$ and $\otimes$ will denote the direct sum and the tensor product of vector spaces respectively.

[1] For technical reasons, we remove from $\mathcal{C}(I, E)$ a subset of pathological paths called *tree-like* (Sec. 2.3 Fermanian, 2019; Hambly and Lyons, 2010). This removal has no theoretical or practical impact on what follows.

For example, $(\mathbb{R}^d)^{\otimes 2} = \mathbb{R}^d \otimes \mathbb{R}^d$ is the space of $d \times d$ matrices and $(\mathbb{R}^d)^{\otimes 3}$ is the space of $d \times d \times d$ tensors. By convention $E^{\otimes 0} = \mathbb{R}$. The following vector space will play a central role in this paper

$$\mathcal{T}(E) = \bigoplus_{k=0}^{\infty} E^{\otimes k} = \mathbb{R} \oplus E \oplus E^{\otimes 2} \oplus \dots \tag{4}$$

If $\{e_1, \dots, e_d\}$ is a basis of E, the elements $\{e_{i_1} \otimes \dots \otimes e_{i_k}\}_{(i_1,\dots,i_k)\in\{1,\dots,d\}^k}$ form a basis of $E^{\otimes k}$. For any $A \in \mathcal{T}(E)$ we denote by $A_k \in E^{\otimes k}$ the k-tensor component of A and by $A^{(i_1,\dots,i_k)} \in \mathbb{R}$ its $(i_1 \dots i_k)^{th}$ coefficient. If E is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_E$, then there exists a canonical inner product $\langle \cdot, \cdot \rangle_{E^{\otimes k}}$ on each $E^{\otimes k}$ which extends by linearity to an inner product

$$\langle A, B \rangle_{\mathcal{T}(E)} = \sum_{k \geq 0} \langle A_k, B_k \rangle_{E^{\otimes k}} \tag{5}$$

on $\mathcal{T}(E)$ that thus becomes also a Hilbert space (Chevyrev and Oberhauser, 2018, Sec. 3).

2.3 The Signature of a path

The *signature* (Chen, 1957; Lyons, 1998, 2014) turns the complex structure of a path x into a simpler vectorial representation given by an infinite sequence of iterated integrals. In this paper, the iterated integrals are defined in the classical *Riemann-Stieltjes* sense.

Definition 2.1. *The signature $S : \mathcal{C}(I, E) \to \mathcal{T}(E)$ is the map defined elementwise in the following way: the 0^{th} coefficient is always $S(x)^{(0)} = 1$, whilst all the others are defined as*

$$S(x)^{(i_1 \dots i_k)} = \int_{a<u_1<\dots<u_k<T} \dots \int dx_{u_1}^{(i_1)} \dots dx_{u_k}^{(i_k)} \in \mathbb{R}, \tag{6}$$

where $t \mapsto x_t^{(i)}$ denotes the i^{th} path-coordinate of x.

It is well known that any continuous function on a compact subset of $\mathbb{R}^d$ can be uniformly well approximated by polynomials (Conway, 2019, Thm. 8.1). In full analogy, the collection of iterated integrals defined by the signature provides a basis for continuous functions on compact sets of paths as stated in the following result (Fermanian, 2019, Prop. 3.).

Theorem 2.1. *Let $\mathcal{X} \subset \mathcal{C}(I, E)$ be a compact set of paths and consider a continuous function $f : \mathcal{X} \to \mathbb{R}$. Then for any $\epsilon > 0$ there exists a truncation level $n \geq 0$ such that for any path $x \in \mathcal{X}$*

$$\left| f(x) - \sum_{k=0}^{n} \sum_{J \in \{1,\dots,d\}^k} \alpha_J S(x)^J \right| < \epsilon, \tag{7}$$

where $\alpha_J \in \mathbb{R}$ are scalar coefficients.

2.4 Truncating the Signature

In view of numerical applications (Bonnier et al., 2019; Graham, 2013; Arribas et al., 2018; Moore et al., 2019; Kalsi et al., 2020), the signature of a path $S(x)$ might need to be truncated at a certain level $n \in \mathbb{N}$ yielding the approximation in $\mathcal{T}^{\leq n}(E) := \mathbb{R} \oplus E^{\otimes 1} \oplus \dots E^{\otimes n}$,

$$S^{\leq n}(x) = (1, S(x)_1, \dots, S(x)_n) \in \mathcal{T}^{\leq n}(E). \tag{8}$$

This approximation is given by the collection of the first $(d^{n+1} - 1)/(d - 1)$ iterated integrals in equation (6). Nonetheless, the resulting approximation is reasonable thanks to Lyons et al. (2007, Proposition 2.2) which states that the absolute value of all neglected terms decays factorially as $|S(x)^{(i_1,\dots,i_n)}| = \mathcal{O}(\frac{1}{n!})$. This factorial decay ensures that when the signature of a path x is truncated, only a negligible amount of information about x is lost (Bonnier et al., 2019, Sec. 1.3).

2.5 Robustness to irregular sampling

The invariance of the signature to a special class of transformations on the time-domain of a path (Friz and Victoir, 2010, Proposition 7.10) called time reparametrizations, such as shifting $t \mapsto t + b$ and acceleration $t \mapsto t^b$ $(b \geq 0)$, partially explains its effectiveness to deal with irregularly sampled datastreams (Bonnier et al., 2019; Chevyrev and Kormilitzin, 2016). In effect, the iterated integrals in equation (6) disregard the time parametrization of a path x, but focus on describing its shape. To retain the information carried by time it suffices to augment the state space of x by adding time t as an extra dimension yielding $t \mapsto \hat{x}_t = (t, x_t^{(1)}, \dots, x_t^{(d)})$. This augmentation becomes particularly useful in the case of univariate time-series where the action of the signature becomes somewhat trivial as there are no interesting dependencies to capture between the different path-coordinates (Chevyrev and Kormilitzin, 2016, Example 5).

3 METHODS

The *distribution regression* (DR) setting for sequential data we have set up so far consists of M groups of input-output pairs of the form

$$\left\{ \left(\{x^{i,p} \in \mathcal{C}(I, E)\}_{p=1}^{N_i}, y^i \in \mathbb{R} \right) \right\}_{i=1}^{M}, \tag{9}$$

such that the finite set of paths $\mathcal{X} = \bigcup_{i=1}^{M} \{x^{i,p}\}_{p=1}^{N_i}$ is a compact subset of $\mathcal{C}(I, E)$. As mentioned in Sec. 1.1, we can summarize the information carried by the collection of paths $\{x^{i,p}\}_{p=1}^{N_i}$ in group i by considering the empirical measure $\delta^i = \frac{1}{N_i} \sum_{p=1}^{N_i} \delta_{x^{i,p}} \in \mathcal{P}(\mathcal{X})$, where

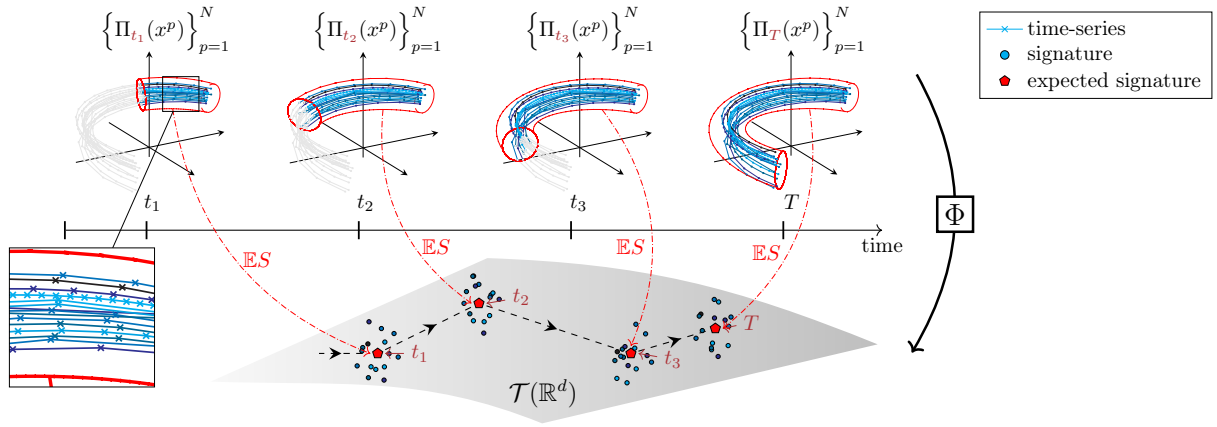

Figure 2: Schematic overview of the action of *pathwise expected signature* Φ on a group of time-series $\{x^p\}_{p=1}^N$. (top) Representation of the information about the group of time-series available from start up to time t_k. (bottom) At each time t_k this information gets embedded into a single point in $\mathcal{T}(\mathbb{R}^d)$.

$\delta_{x^{i,p}}$ is the Dirac measure centred at $x^{i,p}$. This way the input-output pairs in (9) can be represented as follows

$$\left\{\left(\delta^i \in \mathcal{P}(\mathcal{X}), y^i \in \mathbb{R}\right)\right\}_{i=1}^M. \tag{10}$$

The sequence of moments $(\mathbb{E}[Z^{\otimes m}])_{m\geq 0}$ is classically known to characterize the law $\mu_Z = \mathbb{P}\circ Z^{-1}$ of any finite-dimensional random variable Z (provided the sequence does not grow too fast). It turns out that in the infinite dimensional case of laws of paths-valued random variables (or equivalently of probability measures on paths) an analogous result holds (Chevyrev and Oberhauser, 2018). It says that one can fully characterise a probability measure on paths (provided it has compact support) by replacing monomials of a vector by iterated integrals of a path (i.e. signatures). At the core of this result is a recent tool from stochastic analysis that we introduce next.

Definition 3.1. *The expected signature is the map* $\mathbb{E}S : \mathcal{P}(\mathcal{X}) \to \mathcal{T}(E)$ *defined elementwise*

$$\mathbb{E}S(\mu)^{(i_1,\dots,i_k)} = \int_{x\in\mathcal{X}} S(x)^{(i_1,\dots,i_k)}\mu(dx), \tag{11}$$

for any $k \geq 0$ *and any* $(i_1,\dots,i_k) \in \{1,\dots,d\}^k$.

We will rely on the following important theorem in order to prove the universality of the proposed techniques for DR on sequential data presented in the next two sections.

Theorem 3.1. *The expected signature map is injective and weakly continuous.*

Proof. The injectivity has been proved in Chevyrev and Oberhauser (2018, Thm. 5.3). We prove the weak continuity in Appendix A.1. □

3.1 A feature-based approach (SES)

As stated in Thm. 2.1, linear combinations of path-iterated-integrals are universal approximators for continuous functions f on compact sets of paths. In this section we prove the analogous density result for continuous functions F on probability measures on paths. We do so by reformulating the problem of DR on paths as a linear regression on the iterated integrals of an object that we will refer to as the *pathwise expected signature*. We start with the definition of this term followed by the density result. Ultimately, we show that our DR algorithm materializes as extracting signatures on signatures. For any $t \in I = [a, T]$ consider the projection

$$\Pi_t : \mathcal{C}(I, E) \to \mathcal{C}([a, t], E) \tag{12}$$

that maps any path x to its restriction to the subinterval $[a, t] \subset I$, such that $\Pi_t(x) = x_{|_{[a,t]}}$ (see Fig. 2).

Definition 3.2. *The pathwise expected signature is the function* $\Phi : \mathcal{P}(\mathcal{X}) \to \mathcal{C}(I, \mathcal{T}(E))$ *that to a probability measure* $\mu \in \mathcal{P}(\mathcal{X})$ *associates the path* $\Phi(\mu) : I \to \mathcal{T}(E)$ *defined as*

$$\Phi(\mu) : t \mapsto \mathbb{E}_{x\sim\mu}\left[S\left(\Pi_t(x)\right)\right]. \tag{13}$$

The action of Φ is illustrated on Fig. 2, and its implementation is outlined in Alg. 1.[2] In line 6 of the algorithm we use an algebraic property for fast computation of the signature, known as Chen's relation (see Appendix B.2). The next theorem states that any weakly continuous function on $\mathcal{P}(\mathcal{X})$ can be uniformly well approximated by a linear combination of terms in the signature of the pathwise expected signature.

[2] Equivalently $\Phi(\mu) = \mathbb{E}S(\Pi_t \# \mu)$ where $\Pi_t \# \mu$ is the push-forward measure of μ by the measurable map Π_t.

Algorithm 1 Pathwise Expected Signature (PES)

Input: N streams $\{\mathbf{x}^p\}_{p=1}^N$ each of length ℓ
Create array Φ to store the PES
Create array S to store the signatures
Initialize $S[p] \leftarrow 1$ for $p \in \{1, \dots, N\}$
for each time-step $k \in \{2, \dots, \ell\}$ **do**
    // Compute the signature via Chen's relation
    $S[p] \leftarrow S[p] \otimes \exp(\mathbf{x}_k^p - \mathbf{x}_{k-1}^p)$ for $p \in \{1, \dots, N\}$
    $\Phi[k] \leftarrow \text{avg}(S)$
Output: The pathwise expected signature Φ

Algorithm 2 DR on sequential data with SES

Input: $\{(\{\mathbf{x}^{i,p}\}_{p=1}^{N_i}, y^i)\}_{i=1}^M$
Create array A to store M signatures of the PES.
for each group $i \in \{1, ..., M\}$ **do**
    $\Phi = \text{PES}(\{\mathbf{x}^{i,p}\}_{p=1}^{N_i})$ *// Using Alg. 1*
    for each time-step $k \in \{2, \dots, \ell_i\}$ **do**
        // Compute the signature of the PES
        $A[:, i] \leftarrow A[:, i] \otimes \exp(\Phi_k - \Phi_{k-1})$
$(\alpha_0, \dots, \alpha_c) \leftarrow \text{LinearRegression}(A, (y^i)_{i=1}^M)$
Output: Regression coefficients $(\alpha_0, \dots, \alpha_c)$

Theorem 3.2. *Let $\mathcal{X} \subset \mathcal{C}(I, E)$ be a compact set of paths and consider a weakly continuous function $F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$. Then for any $\epsilon > 0$ there exists a truncation level $m \geq 0$ such that for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$*

$$\left| F(\mu) - \sum_{k=0}^{m} \sum_{J \in \{1,\dots,d\}^k} \alpha_J S(\Phi(\mu))^J \right| < \epsilon, \qquad (14)$$

where $\alpha_J \in \mathbb{R}$ are scalar coefficients.

Proof. $\mathcal{P}(\mathcal{X})$ is compact (see proof of Thm. 3.3) and the image of a compact set by a continuous function is compact. Therefore, the image $K = \Phi(\mathcal{P}(\mathcal{X}))$ is a compact subset of $\mathcal{C}(I, \mathcal{T}(E))$. Consider a weakly continuous function $F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$. Given that Φ is injective (see Appendix A.2), Φ is a bijection when restricted to its image K. Hence, there exists a continuous function $f : K \to \mathbb{R}$ (w.r.t $\|\cdot\|_{Lip}$) such that $F = f \circ \Phi$. By Thm. 2.1 we know that for any $\epsilon > 0$, there exists a linear functional $\mathcal{L} : \mathcal{T}(E) \to \mathbb{R}$ such that $\|f - \mathcal{L} \circ S\|_\infty < \epsilon$. Thus $\|F \circ \Phi^{-1} - \mathcal{L} \circ S\|_\infty < \epsilon$, implying $\|F - \mathcal{L} \circ S \circ \Phi\|_\infty < \epsilon$. The approximation error decays factorially with the truncation level m. □

The practical consequence of this theorem is that the complex task of learning a highly non-linear regression function $F : P(\mathcal{X}) \to \mathbb{R}$ can be reformulated as a linear regression on the signature (truncated at level m) of the pathwise expected signature (truncated at level n). The resulting SES algorithm is outlined in Alg.2, and has time complexity $\mathcal{O}(M\ell d^n(N + d^m))$, where M is the total number of groups, ℓ is the largest length across all time-series, d is the state space dimension, N is the maximum number of input time series in a single group. The factorial decay mentioned in Sec. 2.4 also applies to the terms of the (pathwise) expected signature hence low truncation levels $n, m \in \{2, 3\}$ will usually be sufficient in practice to achieve good predictive performances.

3.2 A kernel-based approach (KES)

The SES algorithm is well suited to datasets containing a possibly large number $M \times N$ of relatively low dimensional paths. If instead the input paths are high dimensional, it would be prohibitive to deploy SES since the number of terms in the signature increases exponentially in the dimension d of the path. To address this, in this section we construct a new kernel function $k : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ combining the expected signature with a Gaussian kernel and prove its universality to approximate weakly continuous function on probability measures on paths. The resulting kernel-based algorithm (KES) for DR on sequential data is well-adapted to the opposite data regime to the one above, i.e. when the dataset consists of few number $M \times N$ of high dimensional paths.

Theorem 3.3. *Let $\mathcal{X} \subset \mathcal{C}(I, E)$ be a compact set of paths and $\sigma > 0$. The kernel $k : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ defined by*

$$k(\mu, \nu) = \exp\Big(- \sigma^2 \left\| \mathbb{E}S(\mu) - \mathbb{E}S(\nu) \right\|_{\mathcal{T}(E)}^2 \Big), \qquad (15)$$

is universal, i.e. the associated RKHS is dense in the space of continuous functions from $\mathcal{P}(\mathcal{X})$ to $\mathbb{R}$.

Proof. By Christmann and Steinwart (2010, Thm. 2.2) if K is a compact metric space and H is a separable Hilbert space such that there exists a continuous and injective map $\rho : K \to H$, then for $\sigma > 0$ the Gaussian-type kernel $k_\sigma : K \times K \to \mathbb{R}$ is a universal kernel, where $k_\sigma(z, z') = \exp\Big(- \sigma^2 \left\| \rho(z) - \rho(z') \right\|_H^2 \Big)$. With the metric induced by $\|\cdot\|_{Lip}$, $\mathcal{X}$ is a compact metric space. Hence the set $\mathcal{P}(\mathcal{X})$ is weakly-compact (Walkden, 2014, Thm. 10.2). Given that $(\mathcal{X}, d_{\mathcal{X}})$—where $d_{\mathcal{X}}$ is the topology induced by $\|\cdot\|_{Lip}$—is a compact metric space, the topology describing weak convergence of (Borel) probability measures can be metrized (e.g. by the Prohorov metric $d_{P(\mathcal{X})}$). Therefore $(P(\mathcal{X}), d_{P(\mathcal{X})})$ is also a compact metric space. By Thm. 3.1, the *expected signature* $\mathbb{E}S : \mathcal{P}(\mathcal{X}) \to \mathcal{T}(E)$ is injective and weakly continuous. Furthermore $\mathcal{T}(E)$ is a Hilbert space with a countable basis, hence it is

separable. Setting $K = \mathcal{P}(\mathcal{X}), H = \mathcal{T}(E)$ and $\rho = \mathbb{E}S$ concludes the proof. □

We note that Thm. 3.3 holds more generally for any Taylor-type kernel of the form

$$k(\mu,\nu) = \sum_{n=0}^{\infty} a_n ||\mathbb{E}S(\mu) - \mathbb{E}S(\nu)||^{2n}, \quad a_n > 0 \quad (16)$$

including the Gaussian-type kernel in eq. (15).

3.3 Evaluating the universal kernel k

When the input measures are two empirical measures $\delta^1 = \frac{1}{N_1}\sum_{p=1}^{N_1} \delta_{x^{1,p}}$ and $\delta^2 = \frac{1}{N_2}\sum_{q=1}^{N_2} \delta_{x^{2,q}}$, the evaluation of the kernel k in Equation (15) requires the ability to compute the tensor norm on $\mathcal{T}(E)$

$$\left\|\mathbb{E}S(\delta^1) - \mathbb{E}S(\delta^2)\right\|^2 = E_{11} + E_{22} - 2E_{12},$$
$$E_{ij} = \frac{1}{N_i N_j}\sum_{p,q=1}^{N_i,N_j} \left\langle S(x^{i,p}), S(x^{j,q})\right\rangle, \ i,j \in \{1,2\} \quad (17)$$

where all the inner products are in $\mathcal{T}(E)$. Each of these inner products defines another recent object from stochastic analysis called the signature kernel k_{sig} (Király and Oberhauser, 2019). Recently, Cass et al. (2020) have shown that k_{sig} is actually the solution of a surprisingly simple partial differential equation (PDE). This result provides us with a "kernel trick" for computing the inner products in Equation (17) by a simple call to any numerical PDE solver of choice.

Theorem 3.4. *(Cass et al., 2020, Thm. 2.2) The signature kernel defined as*

$$k_{sig}(x,y) := \langle S(x), S(y)\rangle_{\mathcal{T}(E)} \quad (18)$$

is the solution $u : [a,T]\times[a,T] \to \mathbb{R}$ at $(s,t) = (T,T)$ of the following linear hyperbolic PDE

$$\frac{\partial^2 u}{\partial s \partial t} = (\dot{x}_s^T \dot{y}_t)u \qquad u(a,\cdot) = 1, \ u(\cdot,a) = 1. \quad (19)$$

In light of Thm. 3.3, DR on paths with KES can be performed via any kernel method (Drucker et al., 1997; Quiñonero-Candela and Rasmussen, 2005) available within popular libraries (Pedregosa et al., 2011; De G. Matthews et al., 2017; Gardner et al., 2018) using the Gram matrix computed via Alg. 3 and leveraging the aforementioned kernel trick. When using a finite difference scheme (referred to as PDESolve in Alg. 3) to approximate the solution of the PDE, the resulting time complexity of KES is $\mathcal{O}(M^3 + M^2N^2\ell^2 d)$.

Algorithm 3 Gram matrix for KES

Input: $\{x^{i,p}\}_{p=1}^{N_i}$, $i = 1,\dots,M$ and $\sigma > 0$.
Initialize 0-array $G \in \mathbb{R}^{M\times M}$
for each pair of groups (i,j) such that $i \le j$ **do**
  Initialize 0-array $K_{ij} \in \mathbb{R}^{N_i \times N_j}$
  Similarly initialize $K_{ii}, K_{jj} \in \mathbb{R}^{N_i\times N_i}, \mathbb{R}^{N_j\times N_j}$
  for p, p' in group i and q, q' in group j **do**
    $K_{ii}[p,p'] \leftarrow \text{PDESolve}(x^{i,p}, x^{i,p'})$
    $K_{jj}[q,q'] \leftarrow \text{PDESolve}(x^{j,q}, x^{j,q'})$
    $K_{ij}[p,q] \leftarrow \text{PDESolve}(x^{i,p}, x^{j,q})$
  $G[i,j] \leftarrow \text{avg}(K_{ii}) + \text{avg}(K_{jj}) - 2\times\text{avg}(K_{ij})$
  $G[j,i] \leftarrow G[i,j]$
$G \leftarrow \exp(-\sigma^2 G)$ // *elementwise exp*
Output: The gram matrix G.

Remark In the case where the observed paths are assumed to be i.i.d. samples $\{x^p\}_{p=1}^N \sim \mu$ from the law of an underlying random process one would expect the bigger the sample size N, the better the approximation of μ, and therefore of its expected signature $\mathbb{E}S(\mu)$. Indeed, for an arbitrary multi-index $\tau = (i_1,\dots,i_k)$, the *Central Limit Theorem* yields the convergence (in distribution)

$$\sqrt{N}\Big(\mathbb{E}_{x\sim\mu}[S^\tau(x)] - \frac{1}{N}\sum_{p=1}^{N} S^\tau(x^p)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma_\tau^2)$$

as the variance $\sigma_\tau^2 = \mathbb{E}_{x\sim\mu}[S^\tau(x)^2] - (\mathbb{E}_{x\sim\mu}[S^\tau(x)])^2$ is always finite; in effect for any path x, the product $S^\tau(x)S^\tau(x)$ can always be expressed as a finite sum of higher-order terms of $S(x)$ (Chevyrev and Kormilitzin, 2016, Thm. 1). However, we note that Monte Carlo sampling is only one way of estimating the expected signature. There are stochastic processes such as Brownian motion, for which the expected signature can be computed by solving a PDE (Ni, 2012).

4 RELATED WORK

Recently, there has been an increased interest in extending regression algorithms to the case where inputs are sets of numerical arrays (Hamelijnck et al., 2019; Law et al., 2018a; Musicant et al., 2007; Wagstaff et al., 2008; Skianis et al., 2020). Here we highlight the previous work most closely related to our approach.

Deep learning techniques DeepSets (Zaheer et al., 2017) are examples of neural networks designed to process each item of a set individually, aggregate the outputs by means of well-designed operations (similar to pooling functions) and feed the aggregated output to a second neural network to carry out the regression. However, these models depend on a large

number of parameters and results may largely vary with the choice of architecture and activation functions (Wagstaff et al., 2019).

Kernel-based techniques In the setting of DR, elements of a set are viewed as samples from an underlying probability distribution (Szabó et al., 2016; Law et al., 2018b; Muandet et al., 2012; Flaxman, 2015; Smola et al., 2007). This framework can be intuitively summarized as a two-step procedure. Firstly, a probability measure μ is mapped to a point in an RKHS $\mathcal{H}_1$ by means of a *kernel mean embedding* $\Phi : \mu \to \int_{x\in\mathcal{X}} k_1(\cdot, x)\mu(dx)$, where $k_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the associated reproducing kernel. Secondly, the regression is finalized by approximating a function $F : \mathcal{H}_1 \to \mathbb{R}$ via a minimization of the form $F \approx \arg\min_{g\in\mathcal{H}_2} \sum_{i=1}^{M} \mathcal{L}(y^i, g \circ \Phi(\mu^i))$, where $\mathcal{L}$ is a loss function, resulting in a procedure involving a second kernel $k_2 : \mathcal{H}_1 \times \mathcal{H}_1 \to \mathbb{R}$. In Sec. 5 we denote by DR-k_1 the models produced by choosing k_2 to be a Gaussian-type kernel. Despite the theoretical guarantees of these methods (Szabó et al., 2016), the feature map $k_1(\cdot, x)$ acting on the support $\mathcal{X}$ is rarely provided explicitly, especially in the setting of non-standard input spaces $\mathcal{X} \not\subset \mathbb{R}^d$, requiring manual adaptations to make the data compatible with standard kernels.

The signature method The signature method consists in using the terms of the signature as features to solve supervised learning problems on time-series, with successful applications for detection of bipolar disorder (Arribas et al., 2018) and human action recognition (Yang et al., 2017) to name a few. The signature features have been used to construct neural network layers (Bonnier et al., 2019; Graham, 2013) in deep architectures. To the best of our knowledge, we are the first to use signatures in the context of DR.

5 EXPERIMENTS

We benchmark our feature-based (SES) and kernel-based (KES) methods against DeepSets and the existing kernel-based DR techniques discussed in Sec. 4 on various simulated and real-world examples from physics, mathematical finance and agricultural science. With these examples, we show the ability of our methods to handle challenging situations where only a few number of labelled groups of multivariate time-series are available. We consider the kernel-based techniques DR-k_1 with $k_1 \in \{\text{RBF}, \text{Matern32}, \text{GA}\}$, where GA refers to the Global Alignment kernel for time-series from Cuturi et al. (2007). Unlike our methods, DeepSets, DR-RBF and DR-Matern32 are all for static arrays on $\mathbb{R}^d$. This is why, we also construct the DR-GA method, which can be seen as a simplification of KES, where some smaller terms are deleted in the signature (see Király and Oberhauser (2019, Sec 5.)).

For KES and DR-k_1 we perform Kernel Ridge Regression, whilst for SES we use Lasso Regression. All models are run 5 times and we report the mean and standard deviation of the predictive mean squared error (MSE). Other metrics are reported in Appendix C. The hyperparameters of KES, SES and DR-k_1 are selected by cross-validation via a grid search on the training set of each run. Additional details about hyperparameters search and model architecture can be found in Appendix B. The code to reproduce the experiments is available at `https://github.com/maudl3116/Distribution_Regression_Streams`.

5.1 A defective electronic device

We start with a toy example to show the robustness of our methods to irregularly sampled time-series. For this, we propose to infer the phase φ of an electronic circuit from multiple recordings of its voltage $v^\varphi(t) = \sin(\omega t)$ and current $i^\varphi(t) = \sin(\omega t - \varphi)$. The data consists of M simulated circuits with phases $\{\varphi^i\}_{i=1}^M$ selected uniformly at random from $[\pi/8, \pi/2]$. Each circuit is attached to N measuring devices recording the two sine waves over 20 periods at a frequency 25 points per period. We then randomly subsample the data at rates ranging from 0% to 75% independently for each defective device. As shown in Fig. 3, the predictive performances of DR-RBF drastically deteriorate when the subsampling rate increases, whilst results for KES and SES remain roughly unchanged.

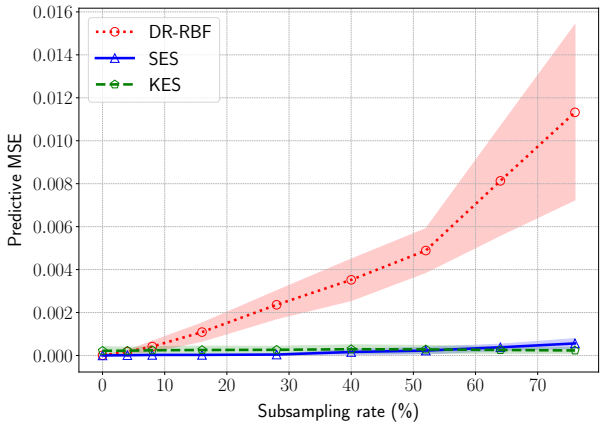

Figure 3: Predictive MSE at various subsampling rates for $M = 50$ circuits and $N = 15$ devices. The shaded area indicates the standard deviation.

Table 1: Ideal gas dataset. Radii of all particles composing the simulated gases: $r_1 = 3.5 \cdot 10^{-1}(V/N)^3$ (few collisions) and $r_2 = 6.5 \cdot 10^{-1}(V/N)^3$ (many collisions).

Model	Predictive MSE [$\times 10^{-2}$]	
	r_1	$r_2 > r_1$
DeepSets	8.69 (3.74)	5.61 (0.91)
DR-RBF	3.08 (0.39)	4.36 (0.64)
DR-Matern32	3.54 (0.48)	4.12 (0.39)
DR-GA	2.85 (0.43)	3.69 (0.36)
KES	**1.31** (0.34)	**0.08** (0.02)
SES	**1.26** (0.23)	**0.09** (0.03)

5.2 Inferring the temperature of an ideal gas

The thermodynamic properties of an *ideal gas* of N particles inside a 3-d box of volume V (3 cm^3) can be described in terms of the temperature T (K), the pressure P (Pa) and the total energy U (J) via the two equations of state $PV = Nk_BT$ and $U = c_V Nk_BT$, where k_B is the *Boltzmann constant* (Adkins and Adkins, 1983), and c_V the heat capacity. The large-scale behaviour of the gas can be related to the trajectories of the individual particles (through their *momentum* = *mass* $\times$ *velocity*) by the equation $U = \frac{1}{2}\sum_{p=1}^{N} m_p|\overrightarrow{v_p}|^2$. The complexity of the large-scale dynamics of the gas depends on T (see Fig. 1) as well as on the radius of the particles. For a fixed T, the larger the radius the higher the chance of collision between the particles. We simulate $M = 50$ different gases of $N = 20$ particles each by randomly initialising all velocities and letting particles evolve at constant speed.[3] The task is to learn T (sampled uniformly at random from $[1, 1\,000]$) from the set of trajectories traced by the particles in the gas. In Table 1 we report the results of two experiments, one where particles have a small radius (few collisions) and another where they have a bigger radius (many collisions). The performance of DR-k_1 is comparable to the ones of KES and SES in the simpler setting. However, in the presence of a high number of collisions our models become more informative to retrieve the global temperature from local trajectories, whilst the performance of DR-k_1 drops with the increase in system-complexity. With a total number of $MN = 1\,000$ time-series of dimension $d = 7$ (after path augmentation discussed in Sec. 2.5 and in Appendix B), KES runs in 50 seconds, three times faster than SES on a 128 cores CPU.

[3]We assume (Chang, 2015) that the environment is frictionless, and that particles are not subject to other forces such as gravity. We make use of python code from `https://github.com/labay11/ideal-gas-simulation`.

5.3 Parameter estimation in a pricing model

Financial practitioners often model asset prices via an SDE of the form $dP_t = \mu_t dt + \sigma_t dW_t$, where μ_t is a drift term, W_t is a 1-d Brownian motion (BM) and σ_t is the volatility process (Arribas et al., 2020). This setting is often too simple to match the volatility observed in the market, especially since the advent of electronic trading (Gatheral et al., 2018). Instead, we model the (rough) volatility process as $\sigma_t = \exp\{P_t\}$ where $dP_t = -a(P_t - m)dt + \nu dW_t^H$ is a *fractional Ornstein-Uhlenbeck* (fOU) process, with $a, \nu, m \geq 0$. The fOU is driven by a *fractional Brownian Motion* (fBM) W_t^H of *Hurst exponent* $H \in (0, 1)$, governing the regularity of the trajectories (Decreusefond et al., 1999).[4] In line with the findings in Gatheral et al. (2018) we choose $H = 0.2$ and tackle the task of estimating the mean-reversion parameter a from simulated sample-paths of σ_t. We consider 50 mean-reversion values $\{a^i\}_{i=1}^{50}$ chosen uniformly at random from $[10^{-6}, 1]$. Each a^i is regressed on a collection of $N = 20, 50, 100$ (time-augmented) trajectories $\{\hat{\sigma}_t^{i,p}\}_{p=1}^{N}$ of length 200. As shown in Table 2, KES and SES systematically yield the best MSE among all compared models. Moreover, the performance of KES and SES progressively improves with the number of time series in each group, in accordance to the remark at the end of Sec. 3, whilst this pattern is not observed for DR-RBF, DR-Matern32, and DeepSets. Both KES and SES yield comparable performances. However, whilst the running time of SES remains stable (≈ 1 min) when MN increases from 1 000 to 5 000, the running time of KES increases from ≈ 1 min to 15 min (on 128 cores).

Table 2: Predictive MSE (standard deviation) on the rough volatility dataset. N is the number of rough volatility trajectories and $(M, d, \ell) = (50, 2, 200)$.

Model	Predictive MSE [$\times 10^{-3}$]		
	N=20	N=50	N=100
DeepSets	74.43 (47.57)	74.07 (49.15)	74.03 (47.12)
DR-RBF	52.25 (11.20)	58.71 (19.05)	44.30 (7.12)
DR-Matern32	48.62 (10.30)	54.91 (12.02)	32.99 (5.08)
DR-GA	3.17 (1.59)	2.45 (2.73)	0.70 (0.42)
KES	**1.41** (0.40)	**0.30** (0.07)	**0.16** (0.03)
SES	**1.49** (0.39)	**0.33** (0.12)	**0.21** (0.05)

[4]We note that sample-paths of fBM are not in $\mathcal{C}([0, T], \mathbb{R})$ but we can assume that the interpolations obtained from market high-frequency data provide a sufficiently refined approximation of the underlying process.

5.4 Crop yield prediction from GLDAS data

Finally, we evaluate our methods on a crop yield prediction task. The challenge consists in predicting the yield of wheat crops over a region from the longitudinal measurements of climatic variables recorded across different locations of the region. We use the publicly available Eurostat dataset containing the total annual regional yield of wheat crops in mainland France—divided in 22 administrative regions—from 2015 to 2017. The climatic measurements (temperature, soil humidity and precipitation) are extracted from the GLDAS database (Rodell et al., 2004), are recorded every 6 hours at a spatial resolution of $0.25° \times 0.25°$, and their number varies across regions.[5] We further subsample at random 50% of the measurements. SES and KES are the two methods which improve the most against the baseline which consists in predicting the average yield on the train set (Table 3).

Table 3: MSE and MAPE (mean absolute percentage error) on the Eurostat/GLDAS dataset

Model	MSE	MAPE
Baseline	2.38 (0.60)	23.31 (4.42)
DeepSets	2.67 (1.02)	22.88 (4.99)
DR-RBF	0.82 (0.22)	13.18 (2.52)
DR-Matern32	0.82 (0.23)	13.18 (2.53)
DR-GA	0.72 (0.19)	12.55 (1.74)
KES	0.65 (0.18)	12.34 (2.32)
SES	**0.62** (0.10)	**10.98** (1.12)

6 CONCLUSION

We have developed two novel techniques for *distribution regression on sequential data*, a task largely ignored in the previous literature. In the first technique, we introduce the pathwise expected signature and construct a universal feature map for probability measures on paths. In the second technique, we define a universal kernel based on the expected signature. We have shown the robustness of our proposed methodologies to irregularly sampled multivariate time-series, achieving state-of-the-art performances on various DR problems for sequential data. Future work will focus on developing algorithms to handle simultaneously a large number of groups of high-dimensional time-series.

[5] http://ec.europa.eu/eurostat/data/database

Acknowledgments

We deeply thank Dr Thomas Cass and Dr Ilya Chevyrev for the very helpful discussions. ML and CS were respectively supported by the EPSRC grants EP/L016710/1 and EP/R513295/1. TD acknowledges support from EPSRC (EP/T004134/1), UKRI Turing AI Fellowship (EP/V02678X/1), and the Lloyd's Register Foundation programme on Data Centric Engineering through the London Air Quality project. ML, CS and TL were supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Adkins, C. J. and Adkins, C. J. (1983). *Equilibrium thermodynamics*. Cambridge University Press.

Arribas, I. P., Goodwin, G. M., Geddes, J. R., Lyons, T., and Saunders, K. E. (2018). A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational psychiatry*, 8(1):1–7.

Arribas, I. P., Salvi, C., and Szpruch, L. (2020). Sigsdes model for quantitative finance. In *ACM International Conference on AI in Finance*.

Balvers, R., Wu, Y., and Gilliland, E. (2000). Mean reversion across national stock markets and parametric contrarian investment strategies. *The Journal of Finance*, 55(2):745–772.

Bonnier, P., Kidger, P., Perez Arribas, I., Salvi, C., and Lyons, T. J. (2019). Deep signature transforms. In *Advances in Neural Information Processing Systems*, pages 3099–3109.

Cass, T., Lyons, T., Salvi, C., and Yang, W. (2020). Computing the full signature kernel as the solution of a goursat problem. *arXiv preprint arXiv:2006.14794*.

Chang, J. (2015). Simulating an ideal gas to verify statistical mechanics. `http://stanford.edu/~jeffjar/files/simulating-ideal-gas.pdf`.

Chen, K. (1957). Integration of paths, geometric invariants and a generalized baker-hausdorff formula.

Chevyrev, I. and Kormilitzin, A. (2016). A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*.

Chevyrev, I., Lyons, T., et al. (2016). Characteristic functions of measures on geometric rough paths. *The Annals of Probability*, 44(6):4049–4082.

Chevyrev, I. and Oberhauser, H. (2018). Signature moments to characterize laws of stochastic processes. *arXiv preprint arXiv:1810.10971*.

Christmann, A. and Steinwart, I. (2010). Universal kernels on non-standard input spaces. In *Ad-*

vances in neural information processing systems, pages 406–414.

Conway, J. B. (2019). *A course in functional analysis*, volume 96. Springer.

Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. (2007). A kernel for time series based on global alignments. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–413. IEEE.

Dahikar, S. S. and Rode, S. V. (2014). Agricultural crop yield prediction using artificial neural network approach. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, 2(1):683–686.

De G. Matthews, A. G., Van Der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. (2017). Gpflow: A gaussian process library using tensorflow. *The Journal of Machine Learning Research*, 18(1):1299–1304.

Decreusefond, L. et al. (1999). Stochastic analysis of the fractional brownian motion. *Potential analysis*, 10(2):177–214.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.

Fermanian, A. (2019). Embedding and learning with signatures. *arXiv preprint arXiv:1911.13211*.

Flaxman, S. R. (2015). *Machine learning in space and time*. PhD thesis, Ph. D. thesis, Carnegie Mellon University.

Friz, P. K. and Victoir, N. B. (2010). *Multidimensional stochastic processes as rough paths: theory and applications*, volume 120. Cambridge University Press.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, pages 7576–7586.

Gatheral, J., Jaisson, T., and Rosenbaum, M. (2018). Volatility is rough. *Quantitative Finance*, 18(6):933–949.

Graham, B. (2013). Sparse arrays of signatures for online character recognition. *arXiv preprint arXiv:1308.0371*.

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

Hamelijnck, O., Damoulas, T., Wang, K., and Girolami, M. (2019). Multi-resolution multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 14025–14035.

Hill, T. L. (1986). *An introduction to statistical thermodynamics*. Courier Corporation.

Kalsi, J., Lyons, T., and Arribas, I. P. (2020). Optimal execution with rough path signatures. *SIAM Journal on Financial Mathematics*, 11(2):470–493.

Király, F. J. and Oberhauser, H. (2019). Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20.

Kusano, G., Hiraoka, Y., and Fukumizu, K. (2016). Persistence weighted gaussian kernel for topological data analysis. In *International Conference on Machine Learning*, pages 2004–2013.

Law, H. C., Sejdinovic, D., Cameron, E., Lucas, T., Flaxman, S., Battle, K., and Fukumizu, K. (2018a). Variational learning on aggregate outputs with gaussian processes. In *Advances in Neural Information Processing Systems*, pages 6081–6091.

Law, H. C. L., Sutherland, D., Sejdinovic, D., and Flaxman, S. (2018b). Bayesian approaches to distribution regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR.

Lyons, T. (2014). Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537*.

Lyons, T., Ni, H., et al. (2015). Expected signature of brownian motion up to the first exit time from a bounded domain. *The Annals of Probability*, 43(5):2729–2762.

Lyons, T. J. (1998). Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310.

Lyons, T. J., Caruana, M., and Lévy, T. (2007). *Differential equations driven by rough paths*. Springer.

Moore, P., Lyons, T., Gallacher, J., Initiative, A. D. N., et al. (2019). Using path signatures to predict a diagnosis of alzheimer's disease. *PloS one*, 14(9).

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2012). Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pages 10–18.

Musicant, D. R., Christensen, J. M., and Olson, J. F. (2007). Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 252–261. IEEE.

Ni, H. (2012). *The expected signature of a stochastic process*. PhD thesis, Oxford University, UK.

Oliva, J., Neiswanger, W., Póczos, B., Schneider, J., and Xing, E. (2014). Fast distribution to real regression. In *Artificial Intelligence and Statistics*, pages 706–714. PMLR.

Panda, S. S., Ames, D. P., and Panigrahi, S. (2010). Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing*, 2(3):673–696.

Papavasiliou, A., Ladroue, C., et al. (2011). Parameter estimation for rough differential equations. *The Annals of Statistics*, 39(4):2047–2073.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Póczos, B., Singh, A., Rinaldo, A., and Wasserman, L. (2013). Distribution-free distribution regression. In *Artificial Intelligence and Statistics*, pages 507–515. PMLR.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959.

Reichl, L. E. (1999). A modern course in statistical physics.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D. (2004). The global land data assimilation system. *Bulletin of the American Meteorological Society*, 85(3):381–394.

Schrödinger, E. (1989). *Statistical thermodynamics*. Courier Corporation.

Skianis, K., Nikolentzos, G., Limnios, S., and Vazirgiannis, M. (2020). Rep the set: Neural networks for learning set representations. In *International conference on artificial intelligence and statistics*, pages 1410–1420. PMLR.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016). Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311.

Wagstaff, E., Fuchs, F., Engelcke, M., Posner, I., and Osborne, M. A. (2019). On the limitations of representing functions on sets. In *International Conference on Machine Learning*, pages 6487–6494. PMLR.

Wagstaff, K. L., Lane, T., and Roper, A. (2008). Multiple-instance regression with structured data. In *2008 IEEE International Conference on Data Mining Workshops*, pages 291–300. IEEE.

Walkden, C. (2014). Ergodic theory. *Lecture Notes University of Manchester*.

Yang, W., Lyons, T., Ni, H., Schmid, C., Jin, L., and Chang, J. (2017). Leveraging the path signature for skeleton-based human action recognition. *arXiv preprint arXiv:1707.03993*.

You, J., Li, X., Low, M., Lobell, D., and Ermon, S. (2017). Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). Deep sets. In *Advances in neural information processing systems*, pages 3391–3401.