Supplementary: Distribution Regression for Sequential Data

A PROOFS

In this section, we prove that the expected signature $\mathbb{E}S$ is weakly continuous (Appendix A.1), and that the pathwise expected signature Φ is injective and weakly continuous (Appendix A.2).

Recall that in the main paper we consider a compact subset of paths $\mathcal{X} \subset \mathcal{C}(I, E)$, where I is a closed interval and E is a Banach space of dimension d (possibly infinite, but countable). We will denote by $\mathcal{P}(\mathcal{X})$ the set of Borel probability measures on $\mathcal{X}$ and by $S(\mathcal{X}) \subset \mathcal{T}(E)$ the image of $\mathcal{X}$ by the signature $S : \mathcal{C}(I, E) \to \mathcal{T}(E)$.

As shown in (Chevyrev and Oberhauser, 2018, Section 3), if E is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_E$, then for any $k \geq 1$ the following bilinear form defines an inner product on $E^{\otimes k}$

$$\left\langle e_{i_1} \otimes \ldots \otimes e_{i_k}, e_{j_1} \otimes \ldots \otimes e_{j_k} \right\rangle_{E^{\otimes k}} = \prod_{r=1}^{k} \delta_{i_r, j_r}, \qquad \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \tag{1}$$

which extends by linearity to an inner product $\langle A, B \rangle_{\mathcal{T}(E)} = \sum_{k \geq 0} \langle A_k, B_k \rangle_{E^{\otimes k}}$ on $\mathcal{T}(E)$ that thus becomes also a Hilbert space.

A.1 Weak continuity of the expected signature

Definition A.1. *A sequence of probability measures $\mu_n \in \mathcal{P}(\mathcal{X})$ converges weakly to μ if for every $f \in C_b(\mathcal{X}, \mathbb{R})$ we have $\int_{\mathcal{X}} f d\mu_n \to \int_{\mathcal{X}} f d\mu$ as $n \to \infty$, where $C_b(\mathcal{X}, \mathbb{R})$ is the space of real-valued continuous bounded functions on $\mathcal{X}$.*

Remark. *Since $\mathcal{X}$ is a compact metric space, we can drop the word "bounded" in Def. A.1.*

Definition A.2. *Given two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$, the Wasserstein-1 distance is defined as follows*

$$W_1(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \int_{x, y \in \mathcal{X}} \|x - y\|_{Lip} \, d\gamma(x, y) \tag{2}$$

where the infimum is taken over all possible couplings of μ and ν.

Lemma A.1. *(Chevyrev and Oberhauser, 2018, Theorem 5.3) The signature $S : \mathcal{C}(I, E) \to \mathcal{T}(E)$ is injective.*[1]

Lemma A.2. *(Chevyrev et al., 2016, Corollary 5.5) The signature $S : \mathcal{C}(I, E) \to \mathcal{T}(E)$ is continuous w.r.t. $\|\cdot\|_{Lip}$.*

Lemma A.3. *(Chevyrev and Oberhauser, 2018, Theorem 5.6) The expected signature $\mathbb{E}S : \mathcal{P}(\mathcal{X}) \to \mathcal{T}(E)$ is injective.*[2]

Theorem A.1. *The expected signature $\mathbb{E}S : \mathcal{P}(\mathcal{X}) \to \mathcal{T}(E)$ is weakly continuous.*

Proof. Consider a sequence $\{\mu_n\}_{n \in \mathbb{N}}$ of probability measures on $\mathcal{P}(\mathcal{X})$ converging weakly to a measure $\mu \in \mathcal{P}(\mathcal{X})$. By Lemma A.2 the signature $S : x \mapsto S(x)$ is continuous w.r.t. $\|\cdot\|_{Lip}$. Hence, by definition of weak-convergence (and because $\mathcal{X}$ is compact), for any $k > 0$ and any multi-index $(i_1, \ldots, i_k) \in \{1, \ldots, d\}^k$ it follows that $\int_{x \in \mathcal{X}} S(x)^{(i_1, \ldots, i_k)} \mu_n(dx) \to \int_{x \in \mathcal{X}} S(x)^{(i_1, \ldots, i_k)} \mu(dx)$. The factorial decay given by (Lyons et al., 2007, Proposition 2.2) yields $\int_{x \in \mathcal{X}} S(x) \mu_n(dx) \to \int_{x \in \mathcal{X}} S(x) \mu(dx)$ in the topology induced by $\langle \cdot, \cdot \rangle_{\mathcal{T}(E)}$. □

[1] Up to tree-like equivalence (see (Chevyrev and Oberhauser, 2018, appendix B) for a definition and detailed discussion).

[2] This result was firstly proved in **?** for probability measures supported on compact subsets of $\mathcal{C}(I, E)$, which is enough for this paper. It was also proved in a more abstract setting in Chevyrev et al. (2016). The authors of Chevyrev and Oberhauser (2018) introduce a normalization that is not needed in case of compact supports, as they mention in (Chevyrev and Oberhauser, 2018, (I) - page 2)

A.2 Injectivity and weak continuity of the pathwise expected signature

Theorem A.2. *(Lyons et al., 2007, Theorem 3.7) Let $x \in \mathcal{C}(I, E)$ and recall the definition of the projection $\Pi_t : x \mapsto x_{|_{[a,t]}}$. Then, the $\mathcal{T}(E)$-valued path defined by*

$$S_{path}(x) : t \mapsto S \circ \Pi_t(x) \tag{3}$$

is Lipschitz continuous. Furthermore the map $x \mapsto S_{path}(x)$ is continuous w.r.t. $\|\cdot\|_{Lip}$.

Theorem A.3. *The pathwise expected signature $\Phi : \mathcal{P}(\mathcal{X}) \to \mathcal{C}(I, \mathcal{T}(E))$ is injective.*[3]

Proof. Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ be two probability measures. If $\Phi(\mu) = \Phi(\nu)$, then for any $t \in I$, $\mathbb{E}_{x\sim\mu}[S \circ \Pi_t(x)] = \mathbb{E}_{y\sim\nu}[S \circ \Pi_t(y)]$. In particular, for $t = T$, $\mathbb{E}S(\mu) = \mathbb{E}_{x\sim\mu}[S(x)] = \mathbb{E}_{y\sim\nu}[S(y)] = \mathbb{E}S(\nu)$. The result follows from the injectivity of the expected signature $\mathbb{E}S$ (Lemma A.3). □

Theorem A.4. *The pathwise expected signature $\Phi : \mathcal{P}(\mathcal{X}) \to \mathcal{C}(I, \mathcal{T}(E))$ is weakly continuous.*

Proof. Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}(\mathcal{X})$ converging weakly to $\mu \in \mathcal{P}(\mathcal{X})$. As S_{path} is continuous (Thm. A.2), it follows, by the *continuous mapping theorem*, that $S_{path}\#\mu_n \to S_{path}\#\mu$ weakly, where $S_{path}\#\mu$ is the pushforward measure of μ by S_{path}. Given that S_{path} is continuous and $\mathcal{X}$ is compact, it follows that the image $S_{path}(\mathcal{X})$ is a compact subset of the Banach space $\mathcal{C}(I, \mathcal{T}(E))$. By (**?**, Theorem 6.8) weak convergence of probability measures on compact supports is equivalent to convergence in *Wasserstein-1 distance*. By *Jensen's inequality* $\|\mathbb{E}[S_{path}\#\mu_n] - \mathbb{E}[S_{path}\#\mu]\|_{Lip} \leq \mathbb{E}[\|S_{path}\#\mu_n - S_{path}\#\mu\|_{Lip}]$. Taking the infimum over all couplings $\gamma \in \Pi(S_{path}\#\mu_n, S_{path}\#\mu)$ on the right-hand-side of the previous equation we obtain $\|\mathbb{E}[S_{path}\#\mu_n] - \mathbb{E}[S_{path}\#\mu]\|_{Lip} \leq W_1(S_{path}\#\mu_n, S_{path}\#\mu) \to 0$, which yields the convergence $\mathbb{E}[S_{path}\#\mu_n] \to \mathbb{E}[S_{path}\#\mu]$ in $\|\cdot\|_{Lip}$ over $\mathcal{C}(I, \mathcal{T}(E))$. Noting that $\mathbb{E}[S_{path}\#\mu] = \Phi(\mu)$ concludes the proof. □

B EXPERIMENTAL DETAILS

In our experiments we benchmark KES and SES against DeepSets and DR-k_1 where $k_1 \in \{\text{RBF}, \text{Matern32}, \text{GA}\}$. Both KES and SES do not take into account the length of the input time-series. Apart from DR-GA, all other baselines are designed to operate on vectorial data. Therefore, in order to deploy them in the setting of DR on sequential data, manual pre-processing (such as padding) is required. In the next section we describe how we turn discrete time-series into continuous paths on which the signature operates.

B.1 Transforming discrete time-series into continuous paths

Consider a d-dimensional time-series of the form $\mathbf{x} = \{(t_1, \mathbf{x}_1), \ldots, (t_\ell, \mathbf{x}_\ell)\}$ with time-stamps $t_1 \leq \ldots \leq t_\ell$ and values $\mathbf{x}_k \in \mathbb{R}^d$, and the continuous path x obtained by linearly interpolating between the points $\mathbf{x}_1, \cdots, \mathbf{x}_\ell$. The signature (truncated at level n) of x can be computed explicitly with existing Python packages **???**, does not depend on the time-stamps $(t_1, \ldots, t_{\ell_{i,j}})$, and produces $(d^{n+1} - 1)/(d - 1)$ terms when $d > 1$. When $d = 1$ the signature is trivial since $S^{\leq n}(x) = (1, (x_{t_\ell} - x_{t_1}), \frac{1}{2}(x_{t_\ell} - x_{t_1})^2, \cdots, \frac{1}{n!}(x_{t_\ell} - x_{t_1})^n)$. As mentioned in Sec. 2.5 we can simply augment the paths with a monotonous coordinate, such that $\hat{x} : t \mapsto (t, x_t)$, where $t \in [a, T]$, effectively reintroducing a time parametrization. Another way to augment the state space of the data and obtain additional signature terms is the *lead-lag* transformation (see Def. B.1) which turns a 1-d data stream into a 2-d path. For example if the data stream is $\{1, 5, 3\}$ one obtains the 2-d continuous path $\hat{x} : t \mapsto (x_t^{(lead)}, x_t^{(lag)})$ where $x^{(lead)}$ and $x^{(lag)}$ are the linear interpolations of $\{1, 5, 5, 3, 3\}$ and $\{1, 1, 5, 5, 3\}$ respectively. A key property of the lead-lag transform is that the difference between $S(\hat{x})^{(1,2)}$ and $S(\hat{x})^{(2,1)}$ is the quadratic variation $QV(x) = \sum_{k=1}^{\ell-1}(x_{t_{k+1}} - x_{t_k})^2$ Chevyrev and Kormilitzin (2016). Hence, even when $d > 1$, it may be of interest to lead-lag transform the coordinates of the paths for which the quadratic variation is important for the task at hand.

[3] For any $\mu \in \mathcal{P}(\mathcal{X})$ the path $\Phi(\mu) \in \mathcal{C}(I, \mathcal{T}(E))$. Indeed $\Phi(\mu)$ is a continuous path because x, Π_t, S and Φ are all continuous and the composition of continuous functions is continuous. The Lipschitzianity comes from the fact that $\|\Phi(\mu)\|_{Lip} \leq \mu(\mathcal{X}) \sup_{x\in\mathcal{X}} \|S_{path}(x)\|_{Lip} < +\infty$ by Thm. A.2.

Definition B.1 (Lead-lag)**.** *Given a sequence of points* $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ *in* $\mathbb{R}^d$ *the lead-lag transform yields two new sequences* $\mathbf{x}^{(lead)}$ *and* $\mathbf{x}^{(lag)}$ *of length* $2\ell - 1$ *of the following form*

$$\mathbf{x}_p^{(lead)} = \begin{cases} \mathbf{x}_k & \text{if } p = 2k-1 \\ \mathbf{x}_k & \text{if } p = 2k-2. \end{cases} \qquad \mathbf{x}_p^{(lag)} = \begin{cases} \mathbf{x}_k & \text{if } p = 2k-1 \\ \mathbf{x}_k & \text{if } p = 2k. \end{cases}$$

In our experiments we add time and lead-lag all coordinates except for the first task which consists in inferring the phase of an electronic circuit (see Sec. 5.1 in the main paper).

B.2 Implementation details

The distribution regression methods (including DR-k_1, KES and SES) are implemented on top of the Scikit-learn library Pedregosa et al. (2011), whilst we use the existing codebase `https://github.com/manzilzaheer/DeepSets` for DeepSets.

B.2.1 KES

The KES algorithm relies on the signature kernel trick which is referred to as PDESolve in the main paper. In the algorithm below we outline the finite difference scheme we use for the experiments. In all the experiments presented in the main paper, the discretization level of the PDE solver is fixed to $n = 0$ such that the time complexity to approximate the solution of the PDE is $\mathcal{O}(d\ell^2)$ where ℓ is the length of the longest data stream.

Algorithm 1 PDESolve

```
Input:  two streams $\{\mathbf{x}_k\}_{k=1}^{\ell_\mathbf{x}}$, $\{\mathbf{y}_k\}_{k=1}^{\ell_\mathbf{y}}$ of dimension $d$ and discretization level $n$ (step size = $2^{-n}$)
Create array $U$ to store the solution of the PDE
Initialize $U[i,:] \leftarrow 1$ for $i \in \{1, 2, \ldots, 2^n * (\ell_x - 1) + 1\}$
Initialize $U[:,j] \leftarrow 1$ for $j \in \{1, 2, \ldots, 2^n * (\ell_y - 1) + 1\}$
for each $i \in \{1, 2, \ldots, 2^n * (\ell_\mathbf{x} - 1)\}$ do
    for each $j \in \{1, 2, \ldots, 2^n * (\ell_\mathbf{y} - 1)\}$ do
        $\Delta_\mathbf{x} = (\mathbf{x}_{\lceil i/(2^n) \rceil + 1} - \mathbf{x}_{\lceil i/(2^n) \rceil})/2^n$
        $\Delta_\mathbf{y} = (\mathbf{y}_{\lceil j/(2^n) \rceil + 1} - \mathbf{y}_{\lceil j/(2^n) \rceil})/2^n$
        $U[i+1, j+1] = U[i, j+1] + U[i+1, j] + (\Delta_\mathbf{x}^T \Delta_\mathbf{y} - 1.) * U[i, j]$
Output: The solution of the PDE at the final times $U[-1,-1]$
```

B.2.2 SES

The SES algorithm from the main paper relies on an algebraic property for fast computation of signatures, known as Chen's relation. Given a piecewise linear path $x = \Delta x_{t_2} \star \ldots \star \Delta x_{t_\ell}$ given by the concatenation $\star$ of individual increments $\mathbb{R}^d \ni \Delta x_{t_k} = x_{t_k} - x_{t_{k-1}}$, $k = 2, \ldots, \ell$, one has $S(x) = \exp(\Delta x_{t_2}) \otimes \ldots \otimes \exp(\Delta x_{t_\ell})$, where exp denotes the tensor exponential and $\otimes$ the tensor product. Using Chen's relation, computing the signature (truncated at level n) of a sequence of length ℓ has complexity $\mathcal{O}(\ell d^n)$.

B.2.3 Baselines

For the kernel-based baselines DR-k_1, we perform Kernel Ridge regression with the kernel defined by $k(\delta^i, \delta^j) = \exp\left(-\sigma^2 \left\|\rho(\delta^i) - \rho(\delta^j)\right\|_{\mathcal{H}_1}^2\right)$, where $\rho(\delta^i) = N_i^{-1} \sum_{p=1}^{N_i} k_1(\cdot, x^{i,p})$. For $k_1 \in \{\text{RBF}, \text{Matern32}\}$, if the time-series are multi-dimensional, the dimensions are stacked to form one large vector $x \in \mathbb{R}^{d\ell}$. See Table 1 for the expressions of the kernels k_1 used as baselines.

Table 1: Kernels k_1 for the kernel-based baselines. See **?** for the definition of $\text{dtw}_{1/\gamma}$ in the GA kernel.

RBF	$\exp(-\gamma^2 \left\|x - x'\right\|^2)$
Matern32	$(1 + \sqrt{3}\gamma^2 \left\|x - x'\right\|) \exp(-\sqrt{3}\gamma^2 \left\|x - x'\right\|)$
GA	$\exp(-\gamma\, \text{dtw}_{1/\gamma}(x, x'))$

For DeepSets, the two neural networks are feedforward neural networks with ELU activations. We train by minimizing the mean squared error.

B.3 Hyperparameter selection

All models are run 5 times. The hyperparameters of KES, SES and DR-k_1 are selected by cross-validation via a grid search on the training set (80% of the data selected at random) of each run. The range of values for each parameter is specified in Table 2.

Table 2: Range of values for each parameter of DR-k_1, KES and SES. We denote by α the regularization parameter in Kernel Ridge regression and Lasso regression. The kernels parameters γ and σ are expressed in terms of lengthscales ℓ_1 and ℓ_2 such that $\gamma^2 = 1/(2\ell_1^2)$ and $\sigma^2 = 1/(2\ell_2^2)$.

Model	ℓ_1	ℓ_2	α	n	m
DR-RBF	$\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$	$\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$	$\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$	N/A	N/A
DR-Matern32	$\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$	$\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$	$\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$	N/A	N/A
DR-GA	$\{7 \cdot 10^{1}, 7 \cdot 10^{2}\}$	$\{10^{-3}, 10^{-2}, \cdots, 10^{2}, 10^{3}\}$	$\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$	N/A	N/A
KES	N/A	$\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$	$\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$	N/A	N/A
SES	N/A	N/A	$\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}$	$\{2, 3\}$	$\{2\}$

C ADDITIONAL RESULTS

C.1 Additional performance metrics

We report the mean absolute percentage error (MAPE) as well as the computational time on two synthetic examples (the ideal gas and the rough volatility examples). As discussed in the main paper, these two datasets represent two data regimes: in one case (the rough volatility model) there is a high number of low dimensional time-series (see Fig. 1), whist in the other case (ideal gas), there is a relatively small number of time-series with a higher state-space dimension. Apart from DeepSets (which is run on a GPU), all other models are run on a 128 cores CPU.

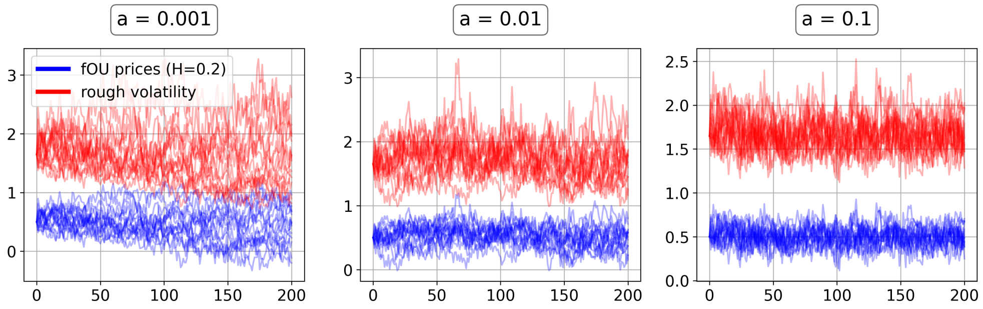

Figure 1: Visualization of fOU sample-paths and their corresponding volatility. Each panel corresponds to a different mean-reversion parameter $a \in \{0.001, 0.01, 0.1\}$.

Table 3: Ideal gas example

Model	Predictive MAPE		Time (s)
	r_1	$r_2 > r_1$	
DeepSets	82.50 (20.20)	53.49 (13.93)	31
DR-RBF	32.09 (5.78)	41.15 (11.81)	58
DR-Matern32	33.79 (5.16)	40.20 (12.45)	55
DR-GA	31.61 (5.60)	39.17 (13.87)	68
KES	16.57 (4.86)	4.20 (0.79)	49
SES	15.75 (2.65)	4.44 (1.36)	120

Table 4: Rough volatility example.

Model	Predictive MAPE			Time (min)		
	N=20	N=50	N=100	N=20	N=50	N=100
DeepSets	44.85 (17.80)	44.75 (17.93)	45.00 (18.21)	1.31	1.86	2.68
DR-RBF	43.86 (13.36)	45.54 (10.05)	41.00 (12.98)	0.71	1.38	7.50
DR-Matern32	40.97 (10.81)	43.59 (9.79)	35.35 (9.18)	0.73	1.00	7.80
DR-GA	11.94 (7.14)	9.54 (6.85)	5.51 (2.78)	0.68	2.60	9.80
KES	6.12 (1.00)	2.83 (0.49)	2.07 (0.42)	0.71	4.00	15.50
SES	6.67 (3.35)	3.58 (0.84)	2.14 (0.62)	0.60	0.65	0.78

C.2 Interpretability

When dealing with complex data-streams, cause-effect relations between the different path-coordinates might be an essential feature that one wishes to extract from the signal. Intrinsic in the definition of the signature is the concept of iterated integral of a path over an ordered set of time indices $a < u_1 < \ldots < u_k < T$. This ordering of the domain of integration, naturally captures causal dependencies between the coordinate-paths $x^{(i_1)}, \ldots, x^{(i_k)}$.

Taking this property into account, we revisit the crop yield prediction example (see Sec. 5.4 in the main paper, and Fig. 3) to show how the iterated integrals from the signature (of the pathwise expected signature) provide interpretable predictive features, in the context of *distribution regression* (DR) with SES. For this, we replace the climatic variables by two distinct multi-spectral reflectance signals: 1) near-infrared (nR) spectral band; 2) red (R) spectral band **?**. These two signals are recorded at a much lower temporal resolution than the climatic variables, and are typically used to assess the health-status of a plant or crop, classically summarized by the *normalized difference vegetation index* (NDVI) **?**. To carry out this experiment, we use a publicly available dataset **?** which contains multi-spectral time-series corresponding to geo-referenced French wheat fields from 2006 to 2017, and consider these field-level longitudinal observations to predict regional yields (still obtained from the Eurostat database).[4]Instead of relying on a predefined vegetation index signal, such as the aforementionned NDVI : $t \mapsto (x_t^{nR} - x_t^{R})/(x_t^{nR} + x_t^{R})$, we use the raw signals in the form of 2-dimensional paths $x : t \mapsto x_t = (x_t^{nR}, x_t^{R})$ to perform a Lasso DR with SES.

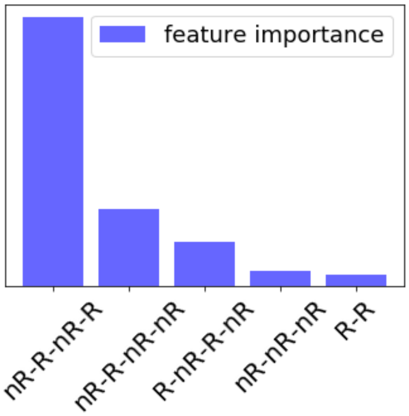

Figure 2: The 5 most predictive features provided by (Lasso) SES for the task of crop yield prediction.

Interpretation Chlorophyll strongly absorbs light at wavelengths around $0.67\mu m$ (red) and reflects strongly in green light, therefore our eyes perceive healthy vegetation as green. Healthy plants have a high reflectance in the near-infrared between 0.7 and $1.3\mu m$. This is primarily due to healthy internal structure of plant leaves **?**. Therefore, this absorption-reflection cycle can be seen as a good indicator of the health of crops. Intuitively, the healthier the crops, the higher the crop-yield will be at the end of the season. It is clear from Fig. 2 that the feature in the signature that gets selected by the Lasso penalization mechanism corresponds to a double red-infrared cycle, as described above. This simple example shows how the terms of the signature are not only good predictors, but also carry a natural interpretability that can help getting a better understanding of the underlying physical phenomena.

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

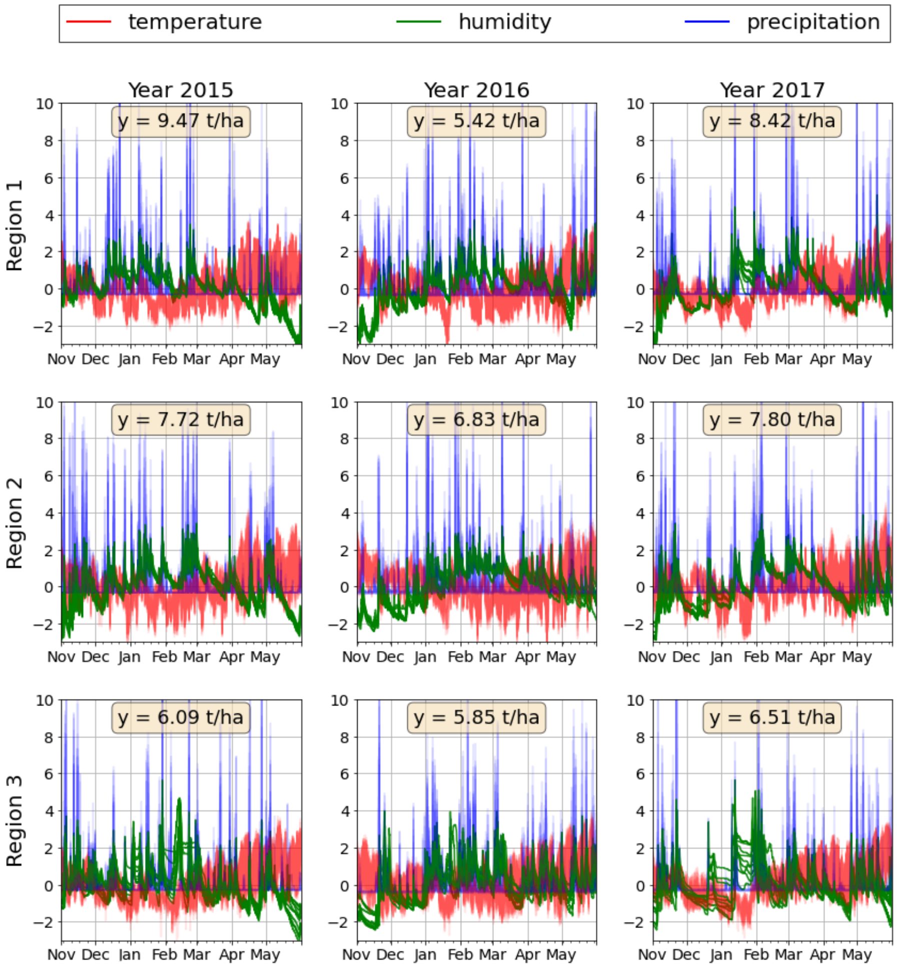

Figure 3: GLDAS/Eurostat dataset. Each panel shows the normalized time-series of temperature, humidity and precipitation, measured over 10 different locations across a region within a year.

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). Sig-sdes 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 *Advances 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.