IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 April 2024, accepted 26 May 2024, date of publication 3 June 2024, date of current version 14 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3408712

RESEARCH ARTICLE

Gaussian Processes Based Data Augmentation and Expected Signature for Time Series Classification

FRANCESCO TRIGGIANO[1] **AND MARCO ROMITO**[2]
[1]Scuola Normale Superiore, 56126 Pisa, Italy
[2]Department of Mathematics, University of Pisa, 56127 Pisa, Italy

Corresponding author: Marco Romito (marco.romito@unipi.it)

The work of Marco Romito was support in part by the European Commission, the NextGeneration EU Programme through the Project PNRR-M4C2-Investimento 1.3, Partenariato Esteso PE00000013-Future Artificial Intelligence Research (FAIR)-Spoke 1 Human-Centered AI; in part by MUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) Bando 2022 through the Project Noise in Fluid Dynamics and Related Models under Grant 20222YRYSP; in part by the University of Pisa through the Project Analysis and Probability in Science (APRISE) under Grant PRA_2022_85; and in part by the Department of Mathematics, University of Pisa, CUP, through MUR Excellence Department Project Awarded under Grant I57G22000700001.

ABSTRACT Time series classification tasks play a crucial role in extracting relevant information from data equipped with a temporal structure. In various scientific domains, such as biology or finance, this kind of data comes from complex and hardly predictable phenomena. Therefore, classification algorithms for time series should be able to deal with the uncertainty contained in data and capture the relevant statistical properties of the underlying phenomenon. The main object of interest of this work is the development of a model for time series that tackles the classification task by interpreting time series as realisations of stochastic processes, the natural mathematical description of chaotic behaviour. The focus thus is on time series that can be thought as signals of some nature, and that convey some kind of statistical information. We propose a data-driven feature extraction model for time series built upon a Gaussian process based data augmentation and on the expected signature. The signature is a fundamental object that describes paths, much alike Fourier or wavelet expansion, but in a non-linear fashion. Likewise, the expected signature provides a statistical description of the law of stochastic processes. One of the main features is that an optimal feature extraction is learnt through the supervised task that uses the model. The model can be adapted to more complicated supervised tasks, as it integrates seamlessly in a neural network architecture and is fully compatible with back-propagation, and it can be easily accommodated to perform regressive tasks. The effectiveness of the model is demonstrated with numerical experiments on some benchmark time series.

INDEX TERMS Expected signature, Gaussian process regression model, stochastic data augmentation, time series classification.

I. INTRODUCTION

Whenever we are dealing with structured data, such as images or time series, we need to deploy some feature extraction mechanisms in order to solve both supervised and unsupervised tasks. There are various well-known time series features extraction approaches, such as *Catch22* [1] or *Tsfresh* [2], which are widely used in general problems. In this work, we are interested in analyzing problems where the supervised task should be able to capture the values as well as the statistical features of input data. We have in mind problems coming from hardly predictable phenomena such as classification of ECG traces, or of the motion of single cells, see for instance [3], [4]. To this end, we focus on a collection of mathematical tools that come from the analysis of irregular signals, the so-called rough paths. In particular, we will use

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

the *signature* of a path as the main device for detecting the most significant features, in terms of the classification task, of time series.

Signature has arisen in the context of rough path theory [5], [6], [7], a theory initially developed to analyse irregular signals and to construct solutions of differential equations driven by such irregular signals. The signature has shown to be a powerful tool to capture the peculiarities of path-like data. In particular, it is able to characterize any path up to adding the time component [8] (see also Proposition 5) and a universal approximation theorem holds [9] (see Theorem 4). As the signature is able to give a non-parametric description of paths, the *expected signature* [10], [11] is the suitable object to extract features from distributions on paths: the expected signature is a well-suited transform whenever a time series is thought as a trajectory of a stochastic process thanks to its capability of identifying the law of various random processes [12].

A. ORIGINAL CONTRIBUTION

We propose a new time series classification model that uses expected signature as feature extraction procedure. Our method goes beyond an architecture of two models placed in series, namely the feature extraction through the evaluation of the expected signature, followed by the classification task based on these features. In our architecture the two models interact and our algorithm learns an optimal evaluation of the expected signature through the feedback of the classification task. The proposed new feature extraction approach can be compared to the convolutional layers when working on images [13]. Indeed, in both cases the features extraction procedure is learned by the model itself based on the prediction task at stake. So, it should be deployed to solve supervised tasks, such as classification. The feature extraction phase combines two main ideas. The first is a stochastic data augmentation based on a generalized Gaussian process regression model. The second idea is to capture the relevant features of paths by means of the expected signature, computed over the ensemble obtained in the phase of data augmentation.

In conclusion, in the present work,

1) we show that the expected signature is an effective tool for supervised tasks, for instance classification, involving time series, when one wants to capture the statistical features of the series and exploit them for better accuracy;
2) we develop a *data augmentation/Gaussian Process regression/computation of expected signature* module that is fully compatible with back-propagation, and thus can be seamlessly integrated into any neural network architecture;
3) we find that signature normalization, which is a crucial step to ensure that signature fully captures the statistical properties of paths, turns out to be crucial to ensure computational stability in the evaluation and use of signature;
4) we show that the proposed method effectively increases the performance of signature-based models.

B. RELATED WORK

Gaussian processes are a fundamental tool for non-parametric models, which has found extensive application in the general field of machine learning and in particular for time series [14]. Besides Gaussian processes, several strategies for data augmentation of time series have been developed [15], [16], mainly designed to address the problem of limited dataset sizes. Our approach differs in that we aim to create a large statistical ensemble, rather than a larger dataset.

The *signature* of a time series has recently emerged in the machine learning community as a universal non-parametric descriptor of a stream of time ordered data [17], [18]. The signature transform has been used as features extraction mechanism in neural network-like models [19] and has been integrated in kernel-based models [20], [21]. Moreover, signature-based models have been applied in various scientific domains, such as anomaly detection [22] and handwritten text recognition [23].

In [24] a Gaussian laws based augmentation is combined with the signature transform. Their model differs from ours in two fundamental aspects. First, they augment any starting time series using a classical GP model where the structure of the mean and of the covariance functions are defined a-priori. This a-priori choice imposes particular properties on the Gaussian processes used, such as stationarity. Instead, our model learns these quantities completely on its own without any prescribed constraint on the laws we sample from. More details on this are given in Section. II-B1. Second, they use the signature transform instead of the expected signature because in their augmentation phase randomness is ruled out by passing only the posterior mean and/or variance. We are able to exploit the expected signature since we combine it with a stochastic augmentation based on the Gaussian process (GP) regression model.

The expected signature has so far found more limited use in machine learning in general and in time series classification problems in particular. In [25] the expected signature has been applied to solve distribution regression on sequential data problems, namely the task of learning a function that takes as input a group of time series and produces a single scalar target. Therefore, they make use of the expected signature to identify relevant features for any sample, a group of various time series. In contrast in our model the expected signature is used to characterize the ensemble generated through the augmentation step starting from a single time series. So, our usage of the expected signature is not directly related to any specific task, but it can be deployed for various supervised problems.

C. STRUCTURE OF THE PAPER

The paper is organised as follows. In Section II we describe how to use the expected signature feature extraction module

in a simple classification task. The module is flexible and can be used in more complex classification tasks, as well as regression problems. We discuss some experimental results in Section III. The algorithm is analysed both on synthetic and real world datasets, a description of the datasets we have used is given in Section III-A. Finally, in Appendix we outline the theoretical background and prove some additional results.

II. MODEL ARCHITECTURE

In this section we propose an extended description of the easiest classification model that can be built using our new feature extraction approach and point out various possible architecture variations.

A. PRELIMINARIES

A time series that describes a phenomenon extended in time and that includes the influence of random components can be thought as a set of values sampled from a trajectory of a stochastic process. We preliminary introduce a series of ideas and notions that aim at modeling the description of time-extended phenomena with random components and that will help in illustrating our method.

1) SIGNATURE OF PATHS AND PROCESSES

We start with a short introduction of the signature of a path and of the expected signature of a stochastic process. Technical details are given in Appendix. The interest is in paths, that are continuous functions, that in general have poor properties of regularity, and so can be thought of as functions with strong degrees of oscillations. The signature is a universal object that describes the intrinsic nonlinear nature of the path and the response of systems governed by the path. The signature of a path $(X_t)_{t\in[0,T]}$ is composed by the set of all iterated integrals,

$$S(X)^n = \int \cdots \int_{0<s_1<s_2<\cdots<s_n<T} dX_{s_1} \otimes \cdots \otimes dX_{s_n}, \quad (1)$$

with the convention that $S(X)^0 = 1$. Here $\otimes$ is the tensor product. The expected signature of a stochastic process, which is nothing else but the family of expectations of all iterated integrals of the process, seen as a path, is able to characterize, in a large number of interesting cases, the law of the random signature, and in turns the statistical properties of the process. More precisely, the expected signature characterizes the law of the process only if properly normalized. A (tensor) normalization λ is simply a function that for a given signature $S = (1, s^1, s^2, \dots)$ returns the new element $(1, \lambda(S)s^1, \lambda(S)^2 s^2, \dots)$, where $\lambda(S)$ is a suitable real value. Tensor normalization is fully illustrated in Appendix B.

2) GAUSSIAN PROCESS REGRESSION MODEL

We briefly introduce the idea of the Gaussian Process regression model. A full description can be found in [14]. First, recall that a Gaussian process is a family of random variables $(X_t)_{t\in[0,T]}$ such that all finite dimensional time-marginals have a joint Gaussian distribution. The law of a Gaussian process is completely determined by its mean and covariance functions, namely $m(t) = \mathbb{E}[X_t]$ and $R(s,t) = \mathrm{Cov}(X_s, X_t)$.

Given a time series $x = (x_{t_i})_{i=1}^N$, we look for a function or a set of functions that might have possibly generated the known data and that can be used for interpolating the series at unknown time instants. Let $m(t)$ and $R(s,t)$ be a mean and a covariance function, the corresponding Gaussian process induces a prior distribution over the set of functions. Roughly speaking, this choice reduces the functions that we are taking into account.

By binding together the data and the prior distribution we obtain a set of possible interpolating functions and their likelihood of having generated the time series. Indeed, the possible values assumed by an interpolating function at unknown time instants $\bar{s} = (s_j)_{j=1}^M$ and the likelihood of each possible set of values for that time instants are described by the conditional law, that is the Gaussian law with mean and covariance given respectively by

$$\begin{aligned} &m(\bar{s}) + R(\bar{s},\bar{t})R(\bar{t},\bar{t})^{-1}(x - m(\bar{t})), \\ &R(\bar{s},\bar{s}) - R(\bar{s},\bar{t})R(\bar{t},\bar{t})^{-1}R(\bar{t},\bar{s}), \end{aligned} \quad (2)$$

where $m(\bar{s}) = (m(s_1), .., m(s_M))$ and

$$R(\bar{s},\bar{t}) = \begin{bmatrix} R(s_1,t_1) & \dots & R(s_1,t_N) \\ \vdots & \ddots & \vdots \\ R(s_M,t_1) & \dots & R(s_M,t_N) \end{bmatrix}. \quad (3)$$

Typically the structure of the mean and covariance functions is chosen a-priori, for instance the square exponential covariance function $R(s,t) = \sigma \exp(-\frac{1}{2l^2}(t-s)^2)$, and it remains only to estimate the parameters, which are σ and l in the squared exponential covariance function case.

B. THE MODEL ARCHITECTURE

The architecture of the simplest model deploying the module implementing the optimal evaluation of the expected signature subject to the accuracy of the classification task is made of two main parts.

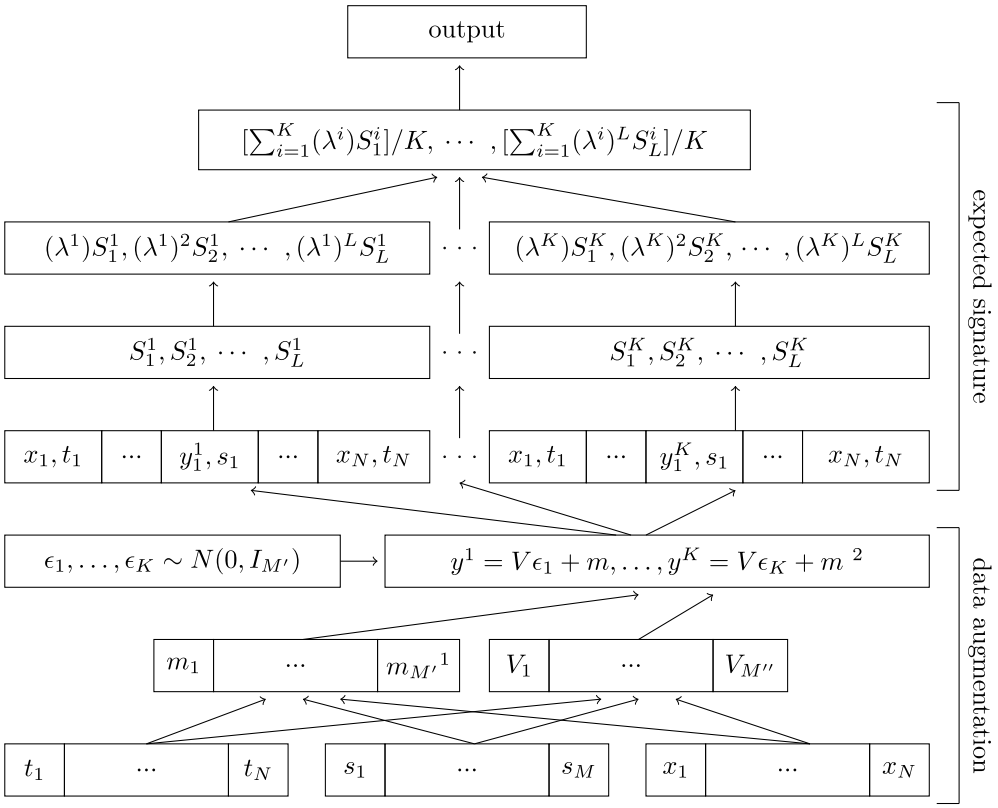
The first element is constituted of three layers and produces a sample of K time series evaluated in a set of new times. The new series are sampled through a generalized Gaussian Process regression model, whose mean and covariance functions are not defined a-priori but they are parameters of the architecture.

The second element is made of four layers, takes as input the sample of K new time series and evaluates the (normalised) expected signature.

A graphical representation of the architecture is shown in Fig. 1. In the following we describe the elements of the model in full detail.

1) DATA AUGMENTATION

We turn to the description of the first element, whose role essentially is to perform data augmentation and sampling of



[a]If $(x_i)_{i=1}^N$ is a d-dimensional time series, then $M' = Md$ and $M'' = \frac{dM(dM+1)}{2}$

[b]Here,

$$m = (m_1, .., .m_{M'}),\ V = \begin{bmatrix} V_1 & 0 & \dots & 0 \\ V_2 & V_3 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ V_{M''-M'+1} & V_{M''-M'+2} & \dots & V_{M''} \end{bmatrix}.$$

FIGURE 1. Graphical representation of the model

K new time series which are coherent with the initial input. This element is constituted by three layers.

The input layer receives the time series values and the corresponding time instants, $(x_i, t_i)_{i=1}^N$, and the sequence $(s_i)_{i=1}^M$ of new time instants. In principle, the set of new time instants can be arbitrarily chosen. Nevertheless, we propose two different choices in Section II-C.

The first hidden layer receives a vector m and a lower triangular matrix V, the output of a linear transformation. They should ideally represent respectively the mean and the square root of the covariance of the conditional law of $X_{s_1}, \ldots, X_{s_M}$ given $X_{t_1} = x_1, \ldots, X_{t_N} = x_N$, that is $(X_{s_1}, \ldots, X_{s_M})$ *conditional* to $(X_{t_1} = x_1, \ldots, X_{t_N} = x_N)$ has the multivariate Gaussian distribution $N(m, VV^T)$ with mean m and covariance matrix VV^T. The third layer, finally, samples K vectors $(y^j)_{j=1}^K$ from $N(m, VV^T)$. A full description of how the sampling is performed can be found in Section II-C. Any vector y^j has M components and it represents the values assumed at the new time instants $(s_i)_{i=1}^M$ by a trajectory that might have generated the starting time series $(x_i)_{i=1}^N$. Then, any vector y^j is combined with the original input by following the temporal order, ending up with a larger and richer time series.

We emphasize that the outcome of this part provides an ensemble of time series that on the one hand are richer than the original input, on the other hand are coherent with the starting time series.

In other words, the first part of the model deploys a data augmentation scheme that makes use of the sampling procedure outlined above to obtain more information about the original time series.

This scheme is strongly connected to the GP regression model described above since they both exploit a Gaussian process based interpolation procedure. We stress again that the main difference between them is how the mean and the covariance are tuned. In a classical GP regression model, the mean and covariance structure are specified a-priori. In our approach the mean and the covariance are fully learnt by the

model. This different approach is both a necessity and an improvement. Indeed, we cannot make use of hand-designed mean and covariance functions since we would need one for each time series. At the same time, this different tuning procedure is a strength of our model because we are not introducing any constraint on the Gaussian law we are sampling from.

2) EVALUATION OF THE EXPECTED SIGNATURE

The second part of the model is responsible for the extraction of the relevant features from the K enriched time series, and is made of four layers. These layers estimate the expected signature based on the ensemble provided by the first part of the model. Clearly, the estimate becomes more and more reliable as long as the size K of the sample increases. A quantitative version of this statement is given by Proposition 14.

The first layer of this phase receives the K enlarged time series and applies a dimensional augmentation by adding the time component. The following layer computes the signature of each time series up to a truncation level L. Then, the normalization procedure that allows the expected signature to characterize the law of stochastic processes is applied to each signature. The expected signature is estimated by averaging component-wise. Finally, the expected signature estimate is used by a softmax layer in order to classify the starting time series.

At this stage we can appreciate that the normalization procedure, which is a requirement to ensure that the expected signature would characterize the law at the theoretical level, turns out to be a crucial step also from a computational point of view. Indeed, we will see that a loose normalization can make the training unstable, see Section III-B1 for experimental evidence.

In addition, the normalization procedure can also be interpreted as a time series preprocessing technique. Indeed, $(\lambda S_1, \lambda^2 S_2, \dots, \lambda^L S_L)$ is both the normalized signature of a given time series $z = (z_{t_i})_i$ and the signature of the rescaled time series λz. We point out that in the machine learning literature one can find several time series normalization methods. Here, they would not be equally effective, since they are not able to preserve the fine theoretical properties of the expected signature. See Remark 16 for further details.

A technical novelty of our work is that the normalization constant λ is found using only the truncated signature. In Corollary 10 we show that the value λ we use is a proper approximation of the theoretical value λ_T. In particular, we prove that λ converges to λ_T, as the truncation threshold of the signature diverges to infinity, and we find an estimate of the convergence rate.

C. TRAINING PROCEDURE AND ARCHITECTURE MODIFICATIONS

One of the main features of the model is that it can be trained by using any classical gradient based optimization scheme (e.g. SGD) since back-propagation can be performed. Indeed, the sampling layer does not interfere with the gradient computation because we are exploiting a well-known Gaussian laws property (if $X \sim N(0, I)$, then $Y = VX + m \sim N(m, VV^T)$) in order to take samples of $N(m, VV^T)$ by just sampling from a standard Gaussian distribution.[1]

The signature layer and normalization procedure are both differentiable thanks to formula (18) and the gradient computed in Corollary 11.

The usage of back-propagation suggests that the proposed model can be easily introduced in more complex and deeper architectures. The easiest possible modification of our model architecture can be obtained by increasing the number of layers in the prediction phase.

There are other possible changes that can be easily implemented. For example, we can introduce any different signature computation algorithm, such as the log-Signature transform [18], or any time series transformation. Indeed, we have been using the time augmentation because it has a relevant role in various theoretical results (Proposition 3 and Theorem 4), but it can be replaced by various transformations. An extended list of possible and useful time series transformations can be found in [27].

Another possible modification can be obtained by introducing a limitation on the square root V of the covariance function of the conditional law, in order to reduce the computational burden. For instance a reasonable modification is to set to zero some sub-diagonals, that is, if $V = (v_{i,j})_{i,j=1}^{M'}$, to set $v_{i,j} = 0$ for all (i, j) such that $i < j$ and $i < \alpha$. The parameter α can be interpreted as a control on the correlation time-scale. Indeed, with this choice, X_{s_l} and X_{s_m} are correlated if $|l - m| < \alpha$.

Lastly, we indicate two possible strategies to select the new time instants $(s_i)_{i=1}^{M}$. The first one considers the middle points of the sub-intervals $[t_i, t_{i+1}]$ for $i = 1 \dots N - 1$ as new time instants. A second choice takes time instants smaller than t_1 or/and bigger than t_N together with the middle points. Even if these two possibilities look quite similar, they produce a substantial difference: all the time series generated using the first strategy have some components of their signature that are shared by all the other time series. For example, they all have the same components of the first level of the signature since these components depend only on the first and last value of each time series and they all have as first and last values the corresponding values of the original time series. Instead, the second strategy makes all the components of the signature affected by the sampling procedure.

III. EXPERIMENTAL RESULTS

In this section we perform some experiments on real and synthetic datasets in order to analyze the effect of the new hyperparameters and to assess the inference capability of the model described in Section II-B.

[1] In the machine learning community this trick is also known as 'the reparameterization trick' [26].

A. IMPLEMENTATION DETAILS

All models have been trained using the SGD optimizer as implemented by `Pytorch` [28]. Signature computations were done using the package `signatory` [29]. The hyperparameters have been tuned using a grid search strategy and cross-validation as validation procedure. Weighted accuracy has been chosen as validation metric due to the strong unbalance of some datasets. The code implementing the model and the generation of the synthetic datasets used is available on a dedicated *GitHub* page [30].

1) DATASETS

The datasets used are of two different types. We have created three synthetic datasets sampling time series with length equal to 100 from trajectories of the following stochastic processes over the time interval [0, 1] (see for instance [31] for further details on the definition of these mathematical objects),

- standard Brownian motion;
- fractional Brownian motion, sampled using the package `fbm`;
- geometric Brownian motion, namely the solution of

$$dX_t = \mu X_t\, dt + \sigma X_t\, dB_t, \tag{4}$$

  which has an explicit solution given by the formula

$$X_t = X_0 \exp\Big((\mu - \frac{\sigma^2}{2})t + \sigma B_t\Big); \tag{5}$$

- Ornstein-Uhlenbeck process, namely the solution of

$$dX_t = \alpha(\gamma - X_t)\, dt + \beta\, dB_t, \tag{6}$$

  with the explicit solution given by the formula

$$X_t = X_0 e^{-\alpha t} + \gamma(1 - e^{-\alpha t}) + \int_0^t \beta e^{\alpha(s-t)}\, dB_s; \tag{7}$$

- the solution of a *stochastic differential equation* with non-linear coefficients,

$$dX_t = \big(\sqrt{1 + X_t^2} + \tfrac{1}{2}X_t\big)\, dt + \sqrt{1 + X_t^2}\, dB_t, \tag{8}$$

  with the explicit solution given by

$$X_t = \sinh\big(\log(\sqrt{1 + X_0^2} + X_0) + t + B_t\big); \tag{9}$$

- white noise perturbations of the following smooth function

$$f(t) = 6\sin^3(4\pi t)\cos^2(4\pi t). \tag{10}$$

In particular, the first two synthetic problems, called **FBM** and **OU**, aim at discriminating two different fractional Brownian motions and two different Ornstein-Uhlenbeck processes, respectively. Instead, the third dataset, called **Bidim**, is composed by bidimensional time series obtained from all the stochastic processes listed above, where first and second components of any time series are trajectories of the same random process.

The second group of datasets has been collected from the *Time series classification* website [32]. We have chosen datasets with not too many observations, in order to reduce the computational burden, while keeping reliable results. At the same time, we have tried to use datasets coming from different topics (Ecg, Sensor, . . .). In particular, we have used the following datasets: *ECG200* (electrical activity recorded during one heartbeat. Here the classes are normal heartbeat and Myocardial Infarction), *Epilepsy* (tri-axial accelerometer data of healthy participants performing one of four class activities), *PowerCons* (household electric power consumption in warm/cold season), *FacesUCR* (rotationally aligned facial outlines of 14 grad students), *Ham* (spectrographs of French or Spanish dry-cured hams). A full description of these datasets can be found on the *Time series classification* website [32].

2) BENCHMARK MODELS

We have used a series of benchmark models in order to compare deterministic augmentation schemes with the stochastic scheme proposed here. The benchmark models used are the following ones:

- `NoAug model`: no augmentation scheme,
- `FFT model`: fast Fourier transform,
- `CS model`: cubic spline interpolation,
- `GP model`: Gaussian Process regression model.

In particular, `NoAug` model receives as input the signature of each time series and applies the normalization procedure discussed in Appendix B and a linear layer with a softmax. Instead, `FFT model`, `CS model` and `GP model` apply the `NoAug model` after the preprocessing phase of the time series. Indeed, any time series is augmented using Fast Fourier transform, cubic spline interpolation or classical GP regression model, respectively. In particular, in the `GP model` each time series is augmented using the posterior mean obtained by a GP regression model assuming that the mean is a constant function and that the covariance is a squared exponential function.

B. RESULTS

In this section we firstly analyze how the new hyperparameters introduced by the proposed data augmentation scheme can affect model performance and stability. Then, we compare the performance of our model with well-known models in the literature and the benchmark models we have introduced in Section III-A2. The comparison with these last models will indicate that our stochastic augmentation module can strongly improve signature-based models.

We preliminary point out that since our model is intrinsically stochastic, we have estimated its performance by running it multiple times with the full test set, and by averaging the obtained accuracy and weighted accuracy. The variance of the output was estimated in the following way:

- The trained model receives 50 times each time series of the test set producing as output 50 vectors of length equal to the number of possible labels, where each vector is a probability distribution over the set $\{1, \dots, D\}$, and where D is the number of possible labels.
- Every time series in the test set is associated with a $D \times D$ covariance matrix obtained from the 50 corresponding vectors;
- The empirical density of the 2-norm of the covariance matrices is computed.

1) HYPERPARAMETERS TUNING AND NORMALIZATION

Our model raises the problem of analysing the effect of a new set of hyperparameters on the performance of the supervised task. As in all signature based models, the level L of the signature truncation is a hyperparameter. Likewise, the data augmentation phase introduces the number M of new time instants that interpolate the original time series. Our model requires two new hyperparameters: the sample size K required for the statistical estimates of the expected signature, and the shape parameter C for the tensor normalization of the signature, whose role is explained below. In this section we focus in detail on K and C, and show that they should be properly chosen in order to achieve competitive results.

We first consider the shape parameter C. We recall that the introduction of a normalization procedure allows the expected signature to characterize the law of the corresponding stochastic process, see Appendix B for further details. In particular, the normalization takes the signature $S = (1, s^1, s^2, \cdots)$ and produces the vector $\lambda S = (1, \lambda(S)s^1, \lambda(S)^2 s^2, \cdots)$, with $\lambda(S)$ that is chosen as the only scalar such that $|\lambda S|$ is equal to $\psi(|S|)$. The function ψ should satisfy various theoretical properties, as stated in Proposition 8. A possible ψ function is given by

$$\psi(\sqrt{x}) = \begin{cases} x, & \text{if } x \leq C, \\ C + C^2(C^{-1} - x^{-1}), & \text{otherwise.} \end{cases} \tag{11}$$

We wish to emphasize that the actual value of C, and thus the tensor normalization, does not play a significant role in the characterization of the law of a stochastic process by means of the normalized expected signature (Theorem 12). In other words, any normalization would fit. On the other hand, our results shown below prove that the shape parameter C plays a relevant role from an experimental point of view. Indeed, it actually determines if the deployed normalization is too strict or too loose and in turn, if the model may underperform or show instabilities. Both these cases should be avoided when training the proposed model.

Fig. 2 shows that when C is close to 1, that is in the case the normalization is very rigid, the model can strongly underperform. In contrast, Fig. 3 shows that a loose normalization can make the training process quite unstable. In particular, the appearance of instabilities even when working with the

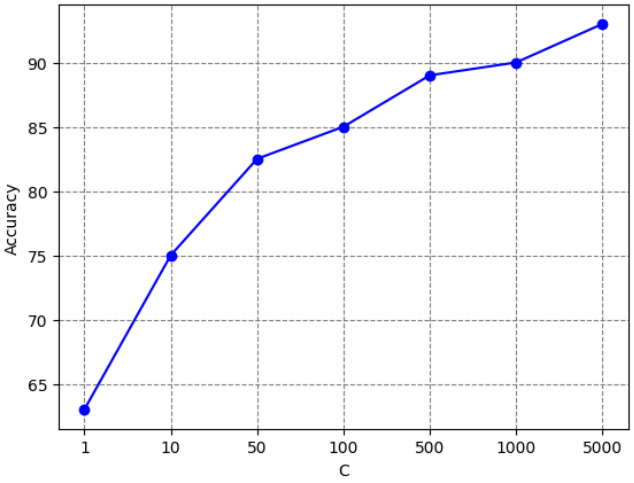


FIGURE 2. Accuracy on *Bidim* dataset depending on the shape parameter C.

TABLE 1. Accuracy results on real datasets. Accuracy and variance averaged over the test set are reported for our model.

Dataset\Model	NoAug	FFT	CS	GP	Model	HC2	1NN-DTW
ECG200	67.5	67	67	70	81 (1e-4)	**87**	77
PowerCons	90	90.5	90	92.7	**99.9 (1e-6)**	98.3	92.2
Epilepsy	68.4	70.4	68	72.5	83 (4e-4)	**100**	97.8
Ham	62.6	58	60	65.7	**74.2 (2e-4)**	72.3	46.6
FacesUCR	63	63	63	53	51 (1e-5)	**96.4**	90.4

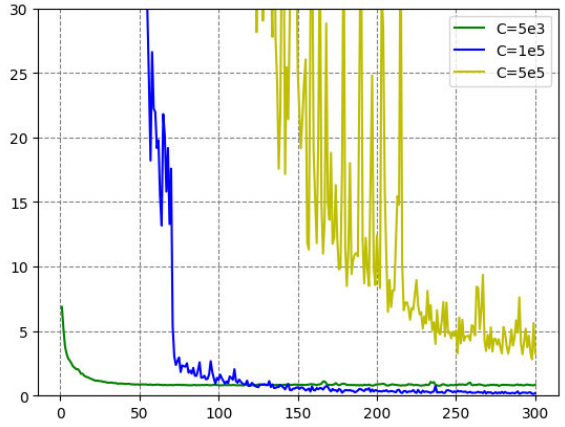
NoAug model defined in Section III-A2, the easiest model that can be built using the signature as feature extraction mechanism, suggests that the normalization procedure should be deployed and tuned whenever the signature transform is used.

We turn to the analysis of the number K of generated augmented time series. Proposition 14 shows that the empirical mean of the K signatures obtained by the K enlarged time series is a good approximation of the expected signature as long as K is sufficiently large. Fig. 4 empirically shows that by increasing K, that is by getting a better and better approximation of the expected signature, the output of the model becomes more and more stable with respect to the sampling procedure. Indeed, if the model is fed multiple times with the same input, then the variance in the output is small for K big enough.

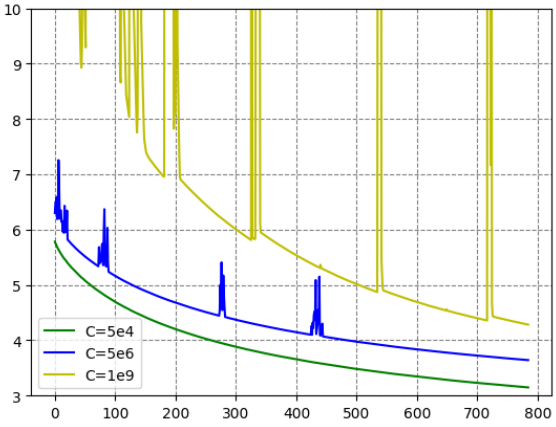
At last, we highlight that a large value of K can slow down the training phase. Hence, we suggest looking for the smallest K up to a reasonably low variance in the output.

C. MODEL PERFORMANCE

In this section we compare the results of our model with some other models on the dataset described in Section III-A1. In particular, we compare the proposed model with two well-known models, 1NN with DTW [33] and Hive-Cote 2 [34], which are considered state-of-the-art models for time series classification problems, see for instance [35], and

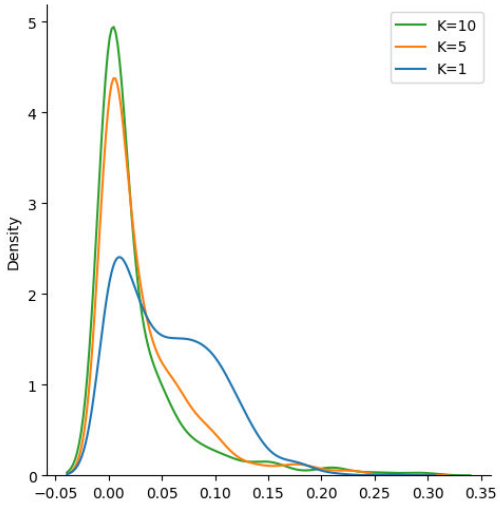


(a) Learning curve of the proposed model.

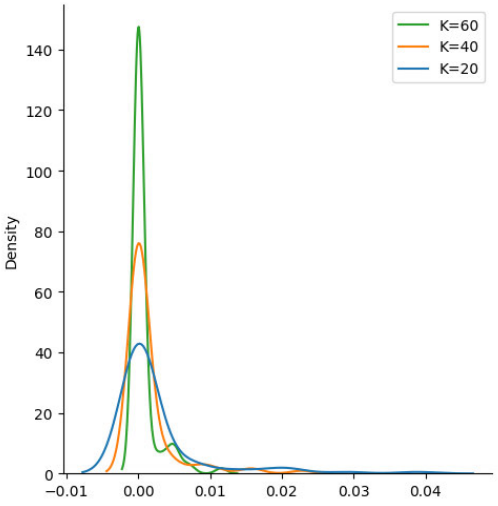


(b) Learning curve of the NoAug model.

FIGURE 3. Results obtained working on *Bidim* dataset.

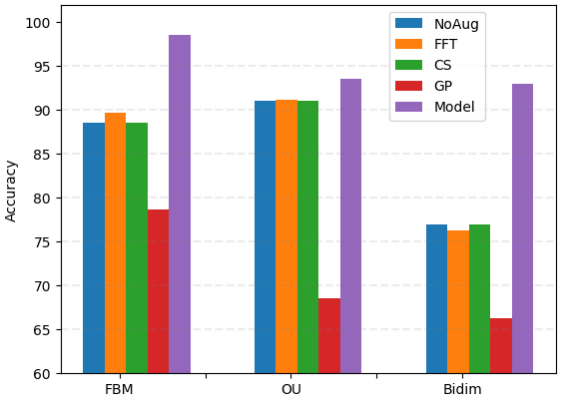


(a) Results obtained working on *Bidim* dataset.

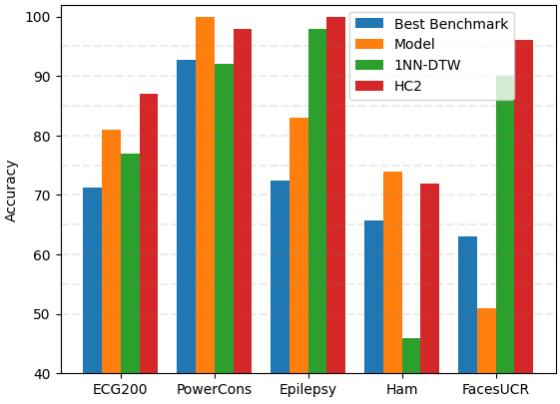


(b) Results obtained working on *ECG200* dataset.

FIGURE 4. Variance in the output.



(a) Performance on synthetic data.



(b) Performance on real data.

FIGURE 5. Performance values in synthetic and real data.

with the benchmark models introduced in Section III-A2, in order to show the effectiveness of the stochastic augmentation. The results are summarised in Table 1 and Fig. 5.

The comparison with the benchmark models indicates that the stochastic augmentation can strongly increase the inference capability with respect to the model that uses the signature instead of the expected signature, i.e. the `NoAug` model.

In addition, these results suggest that classical deterministic interpolation schemes are not as effective as the Gaussian processes based augmentation introduced here. For the sake of completeness we have shown the results of `HC2` and `1NN-DTW`. Clearly, we cannot state that the proposed model is statistically comparable to `HC2`, but the results indicate that the introduced ideas can strongly increase the performance of signature based models.

IV. DISCUSSION AND CONCLUSION

In this study, we proposed a new time series classification model that can be integrated into any neural network architecture and can be easily modified in order to tackle any supervised task. The main idea is the construction of a novel feature extraction scheme based on the combination of a stochastic data augmentation and the expected signature transform. This module is interesting because it has strong theoretical foundations and it allows to increase the performance of signature-based model.

This work can stimulate various future research both from a theoretical and a computational point of view. In particular, it would be extremely useful to look for a differentiable formula for the expected signature of Gaussian processes. A similar result can reduce the computational burden of our model since it will allow to avoid the sampling procedure. We highlight that a first result in this direction can be found in [36]. It could be also interesting to try to understand if a different family of stochastic processes can be deployed in the stochastic augmentation scheme. Such a result will make our model even more flexible and effective. At last, we report that it can be worth introducing our feature extraction module in more complex signature-based architectures, such as the one proposed in [19], in order to improve the performances and obtain additional statistical comparisons with state-of-the-art models for time series classification task.

APPENDIX
THEORETICAL BACKGROUND AND SOME RESULTS

The first part of this appendix contains the definition and property of the signature. The second part introduces the expected signature.

A. SIGNATURE

Almost all the results contained in this section can be found in [5] and in [7]. Firstly, we need to introduce the tensor space and the truncated tensor space.

Definition 1: The tensor product space of $\mathbb{R}^d$ is

$$T(\mathbb{R}^d) = \{(a^n)_{n\in\mathbb{N}} \mid a^n \in (\mathbb{R}^d)^{\otimes n}\}, \tag{12}$$

the set of the formal series of tensors of $\mathbb{R}^d$. The truncated tensor product space at degree N is the set

$$T(\mathbb{R}^d)_L = \{(a^n)_{n\le L} \mid a^n \in (\mathbb{R}^d)^{\otimes n}\}. \tag{13}$$

They are algebras w.r.t the component-wise addition, the component-wise multiplication by scalars and the tensor product

$$a \otimes b = \left(\sum_{i=0}^{n} a^{n-i} \otimes b^i\right)_{n\in\mathbb{N}}. \tag{14}$$

Any tensor component a^n can be represented by its components with respect to the canonical basis of $(\mathbb{R}^d)^{\otimes n}$. In other words a^n can be identified by a set of scalar values $(a^{i_1,\dots,i_n})_{i_1,\dots,i_n=1,\dots,d}$.

Moreover, the set $T_1(\mathbb{R}^d)$ defined by

$$\left\{a \in T(\mathbb{R}^d) \,\middle|\, a_0 = 1, \sum_{n=0}^{\infty} \sum_{i_1,\dots,i_n=1}^{d} |a^{i_1,\dots,i_n}|^2 < \infty\right\}, \tag{15}$$

is actually a Banach space.

Definition 2: Consider $X : [0,T] \to \mathbb{R}^d$ a continuous function with bounded variation. The signature of the path X is the sequence of iterated integrals $S(X)_{0,T} = (1, S(X)^1, S(X)^2, \dots, S(X)^n, \dots)$, where for every $n \ge 1$, $S(X)^n \in (\mathbb{R}^d)^{\otimes n}$ is defined through its components with respect to the canonical basis by

$$S(X)^n = \left(\int_0^T \cdots \int_0^{u_3} \int_0^{u_2} dX_{u_1}^{i_1} dX_{u_2}^{i_2} \dots dX_{u_n}^{i_n}\right), \tag{16}$$

with indices $i_1, \dots, i_n$ running over the set $\{1 \dots, d\}$.

The signature has various properties that make it well-suited for dealing with path-like data.

Proposition 3: Let $X, Y : [0,T] \to \mathbb{R}^d$ be BV and continuous functions and $\phi : [0,T] \to [0,T]$ be a C^1, increasing and surjective function. Then:

1) $S(X)_{s,t} = S(X)_{s,u} \otimes S(X)_{u,t}$ *for all* $s < u < t$;
2) $S(X) = S(X \circ \phi)$;
3) $|S(X)^n| \le \frac{1}{n!}|X|^n_{1,[0,T]}$, *for all* $n \in \mathbb{N}$;
4) $S(\overline{X}) = S(\overline{Y})$ *if and only if* $X = Y$ *(see [8]).*

Here, $S(X)_{s,t}$ stands for the signature of X restricted to the interval $[s,t]$, $|X|_{1,[0,T]}$ is the total variation of X on the interval $[0.T]$, and $\overline{X}_t = (X_t, t)$.

The third property in Proposition 3 shows that by truncating the signature we do not lose a huge amount of information. The fourth property in Proposition 3 on the other hand shows that, up to adding the time component, the signature uniquely identifies the corresponding path.

Moreover, it holds a universal approximation theorem.

Theorem 4: [(I.P. Arribas, [9])] Let $F : K \to \mathbb{R}$ be a continuous function defined over a compact set K, composed by continuous and BV functions from $[0,T]$ to $\mathbb{R}^d$. Then, for any $\epsilon > 0$, there exists a linear map L such that for all $X \in K$,

$$|F(X) - L(S(\overline{X}))| \le \epsilon. \tag{17}$$

At last, we indicate how to compute the signature of a time series.

Definition 5: Let $x = (x_{t_i})_{i=0}^N$ be a time series. Its signature is given by the signature of a linear interpolation of x.

A-priori the definition depends on the choice of the linear interpolation (that is the speed at which one traverses the gap between the x_i), but Proposition 3 ensures that the definition of signature for a stream of data is well-defined and independent from the choice of the linear interpolation. Moreover, Proposition 3 allows to easily compute the signature of a time series. Indeed,

$$S(x) = \exp(x_1 - x_0) \otimes \cdots \otimes \exp(x_N - x_{N-1}), \tag{18}$$

where we recall that $\exp(a) = \sum_{n=0}^{\infty} \frac{1}{n!} a^{\otimes n}$.

B. EXPECTED SIGNATURE

All the results without proof can be found in [12].

Definition 6: Consider a stochastic process $(X_t)_{t\in[0,T]}$ such that almost every trajectory is continuous with bounded variation. The sequence $(\mathbb{E}[S(X)^{i_1,\dots,i_n}])_{i_1,\dots,i_n=1,\dots,d}$ is called the *expected signature* of X.

The expected signature is able to identify the law of the corresponding stochastic process only if it is properly normalized.

Definition 7: A continuous and injective map $\lambda : T_1(\mathbb{R}^d) \to T_1(\mathbb{R}^d)$ is called a *tensor normalization* if there is $\lambda : T_1(\mathbb{R}^d) \to (0, \infty)$ such that:

- $\lambda(t) = \delta_{\lambda(t)}(t) := (1, \lambda(t)t^1, \lambda(t)^2 t^2, \dots)$ for all $t \in T_1(\mathbb{R}^d)$,
- $|\lambda(t)| \le R$ for all $t \in T_1(\mathbb{R}^d)$.

Let us show how such a tensor normalization can be built.

Proposition 8: Let $\psi : [1, \infty) \to [1, \infty)$ be a bounded, injective and K-Lipschitz function such that $\psi(1) = 1$ and $\sup_{x\ge1} \frac{\psi(x)}{x^2} \le 1$. Given $t \in T_1(\mathbb{R}^d)$, let $\lambda(t)$ be the only non-negative value such that $|\delta_{\lambda(t)}(t)|^2 = \psi(|t|)$. Then, the map $\lambda(t) = \delta_{\lambda(t)}(t)$ is a tensor normalization and there exists a constant $c > 0$ such that for all $s, t \in T_1(\mathbb{R}^d)$,

$$|\lambda(s) - \lambda(t)| \le c \min(\sqrt{|t-s|}, |t-s|). \tag{19}$$

Example 9: The function

$$\psi(\sqrt{x}) = \begin{cases} x & \text{if} x \le C, \\ C + \dfrac{C^{1+a}}{a}(C^{-a} - x^{-a}) & \text{otherwise,} \end{cases} \tag{20}$$

$a > 0$ and $C \ge 1$ meets the assumptions of the previous proposition.

Corollary 10: Let ψ be a function as in the previous proposition, $M \in \mathbb{N}^*$, $t \in T_1(\mathbb{R}^d)$ and $t_L = (1, t^1, \dots, t^L, 0, \dots)$. Then,

$$\lambda_L := \lambda(t_L) \to \lambda(t),$$

as $L \to \infty$.

Moreover, suppose that $t = S(X)$ for some continuous function with bounded variation and consider $r = \min\{j \in \mathbb{N} : t^j \neq 0\}$, then for all $L \ge r$,

$$|\lambda_L - \lambda| \le C \min\left(\sqrt[4]{\sum_{j=L+1}^{\infty} \frac{1}{j!}|X|^j_{1,[0,T]}}, \sqrt{\sum_{j=L+1}^{\infty} \frac{1}{j!}|X|^j_{1,[0,T]}}\right)^{\frac{1}{r}}. \tag{21}$$

Proof: If $r = 0$, then $t = (1, 0, \dots, 0, \dots) = t_L$ and the result is trivial. Suppose that $r \neq 0$ and consider $L \ge r$, then

$$\begin{aligned} |\lambda_L^r - \lambda^r|^2 &= \frac{1}{|t^r|^2}|\lambda_L^r t^r - \lambda^r t^r|^2 \\ &\le \frac{1}{|t^r|^2}\Big(\sum_{j=r}^{L} |\lambda_L^j t^j - \lambda^j t^j|^2 + \sum_{j=L+1}^{\infty} |\lambda^j t^j|^2\Big) \\ &= \frac{1}{|t^r|^2}|\lambda(t_L) - \lambda(t)|^2. \end{aligned}$$

Therefore, the convergence follows from the continuity of λ. The inequality Eq. (21) follows from Proposition 3 and the inequality Eq. (19). □

Since the normalization procedure is introduced in the proposed model, we need to be able to compute the gradient of $\lambda(t)$.

Corollary 11: Let ψ be a C^1 function that satisfies the assumptions of Proposition 10, and consider the corresponding $\lambda(t)$ function.

Given $s \in T_1(\mathbb{R}^d)_L$ such that $s \neq (1, 0, \dots, 0)$, there exists an open neighbourhood U of s in $T_1(\mathbb{R}^d)_L$ such that $\lambda|_U$ is a C^1 function and

$$\nabla\lambda(s) = \left(\frac{s_j^{i_1,\dots,i_j}(\lambda(s)^{2j} - \frac{1}{2|s|}\frac{d}{dx}\psi(|s|))}{\sum_{k=1}^{L} k\lambda(s)^{2k-1}\sum_{i_1,\dots,i_k=1}^{d}|s_k^{i_1,\dots,i_k}|^2}\right), \tag{22}$$

with indices $j = 1, \dots, L$ and $i_1, \dots, i_j = 1, \dots, d$.

Proof: Consider the function $F : (0, \infty) \times T_1(\mathbb{R}^d)_L \to \mathbb{R}$ defined by $F(\lambda, t) = |\delta_\lambda(t)|^2 - \psi(|t|)$. Its derivatives are given by the following formulas:

$$\begin{aligned} \partial_\lambda F(\lambda, t) &= \sum_{k=1}^{L} 2k\lambda^{2k-1} \sum_{i_1,\dots,i_k=1}^{d} |t_k^{i_1,\dots,i_k}|^2, \\ \partial_{t_j^{i_1,\dots,i_j}} F(\lambda, t) &= 2t_j^{i_1,\dots,i_j}\Big(\lambda^{2j} - \frac{1}{2|t|}\frac{d}{dx}\psi(|t|)\Big). \end{aligned}$$

Hence, the result follows directly from the implicit function theorem. □

We can finally state the main property of the expected signature.

Theorem 12: Consider a tensor normalization λ and let μ and ν be the laws of $(X_t)_{t\in[0,T]}$ and $(Y_t)_{t\in[0,T]}$, stochastic processes with continuous and BV trajectories. Then, $\mu = \nu$ if and only if $\mathbb{E}[\lambda(S(\overline{X}))] = \mathbb{E}[\lambda(S(\overline{Y}))]$

Since we estimate the expected normalized signature by averaging the normalized signature of K trajectories, we report a concentration inequality.

Lemma 13: (Hoeffding's inequality, [37]) Let $Y_1, \dots, Y_n$ independent random variables such that Y_i takes values in $[a_i, b_i]$ almost surely for all $i \leq n$. Then for every $\sigma > 0$,

$$\mathbb{P}\Big[\sum_{i=1}^{n}(Y_i - E[Y_i]) \geq \sigma\Big] \leq \exp\Big(-\frac{2\sigma^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\Big). \quad (23)$$

Proposition 14: Let λ be a tensor normalization and $(X_t)_{t\in[0,T]}$ a stochastic process with continuous and BV trajectories. Consider $Y_1, \dots, Y_K$ iid random variables with values in $T_1(\mathbb{R}^d)$ such that any Y_i has the same law of the random variable $\lambda(S(X_t))$. Then for all $\sigma > 0$,

$$\mathbb{P}\Big[|\sum_{i=1}^{K}\frac{Y_i - \mathbb{E}[Y_i]}{K}| \geq \sigma\Big] \leq \exp\Big(-\frac{2\sigma^2 K}{(2R)^2}\Big). \quad (24)$$

Proof: We have

$$\mathbb{P}\Big[|\sum_{i=1}^{K}\frac{Y_i - E[Y_i]}{K}| \geq \sigma\Big] \leq \mathbb{P}\Big[\sum_{i=1}^{K}\frac{|Y_i - \mathbb{E}[Y_i]|}{K} \geq \sigma\Big] \leq \exp\Big(-\frac{2\sigma^2 K^2}{\sum_{i=1}^{K}(2R)^2}\Big),$$

where the last inequality is due to Hoeffding's inequality applied to the random variables $\{\frac{1}{K}(|Y_i - \mathbb{E}[Y_i]|)\}_{i=1,\dots,K}$. Indeed, any $\frac{1}{K}(|Y_i - \mathbb{E}[Y_i]|)$ takes values in $[0, \frac{2R}{K}]$ since λ is a tensor normalization. □

At last, we report a concrete example where the normalization is crucial.

Example 15: Consider two $\mathbb{R}^2$-valued stochastic processes

$$(X_t)_{t\in[0,1]} = (tN_1, tN_2)_{t\in[0,1]},$$
$$(Y_t)_{t\in[0,1]} = (tM_1, tM_2)_{t\in[0,1]}, \quad (25)$$

where $N = (N_1, N_2)$ and $M = (M_1, M_2)$ have, respectively, density

$$p(n_1, n_2) = \prod_{i=1}^{2}\frac{1}{n_i\sqrt{2\pi}}\exp(-\tfrac{1}{2}\log^2(n_i)),$$
$$q(m_1, m_2) = p(m_1, m_2)\prod_{i=1}^{2}(1 + \sin(2\pi\log(m_i))). \quad (26)$$

These two stochastic processes have the same expected signature. The proof is an elementary consequence of the following equalities:

- $S(X)^m = \frac{1}{m!}((X(1) - X(0))^{\otimes m})$, for all $m \in \mathbb{N}$;
- $S(X)^m = \frac{1}{m!}((Y(1) - Y(0))^{\otimes m})$, for all $m \in \mathbb{N}$;
- $\mathbb{E}[(X(1) - X(0))^{\otimes m}] = \mathbb{E}[(Y(1) - Y(0))^{\otimes m}]$, for all $m \in \mathbb{N}$.

Remark 16: We have already highlighted that the normalized signature of a path is, actually, the signature of the path multiplied by a constant. Indeed, $(\lambda S_1, \lambda^2 S_2, \dots, \lambda^L S_L)$ is both the normalized signature of a given path $(Z_t)_t$ and the signature of the rescaled path $(\lambda Z_t)_t$. So, the normalization procedure can be thought as a path preprocessing mechanism. We point out that normalization procedures that are well-known in the machine learning field, such as z-normalization or min-max normalization, are not able to produce a result such as Theorem 12. This can be easily shown by applying them to the previous example.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Center for High Performance Computing (CHPC) at Scuola Normale Superiore (SNS) for providing the computational resources used in this work.

REFERENCES

[1] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: CAnonical time-series CHaracteristics: Selected through highly comparative time-series analysis," *Data Mining Knowl. Discovery*, vol. 33, no. 6, pp. 1821–1852, Nov. 2019.

[2] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series FeatuRe extraction on basis of scalable hypothesis tests (tsfresh—A Python package)," *Neurocomputing*, vol. 307, pp. 72–77, Sep. 2018.

[3] M. Darrin, A. Samudre, M. Sahun, S. Atwell, C. Badens, A. Charrier, E. Helfer, A. Viallat, V. Cohen-Addad, and S. Giffard-Roisin, "Classification of red cell dynamics with convolutional and recurrent neural networks: A sickle cell disease case study," *Sci. Rep.*, vol. 13, no. 1, Jan. 2023, Art. no. 745, doi: 10.1038/s41598-023-27718-w.

[4] A. R. Vega, S. A. Freeman, S. Grinstein, and K. Jaqaman, "Multistep track segmentation and motion classification for transient mobility analysis," *Biophysical J.*, vol. 114, no. 5, pp. 1018–1025, Mar. 2018, doi: 10.1016/j.bpj.2018.01.012.

[5] T. J. Lyons, M. Caruana, and Lévy, *Differential Equations Driven by Rough Paths* (Lecture Notes in Mathematics), vol. 1908. Berlin, Germany: Springer, 2007.

[6] P. K. Friz and N. B. Victoir, *Multidimensional Stochastic Processes as Rough Paths* (Cambridge Studies in Advanced Mathematics), vol. 120. Cambridge, U.K.: Cambridge Univ. Press, 2010, doi: 10.1017/CBO9780511845079.

[7] P. K. Friz and M. Hairer, *A Course on Rough Paths* (Universitext). Cham, Switzerland: Springer, 2014, doi: 10.1007/978-3-319-08332-2.

[8] B. Hambly and T. Lyons, "Uniqueness for the signature of a path of bounded variation and the reduced path group," *Ann. Math.*, vol. 171, no. 1, pp. 109–167, Mar. 2010.

[9] I. Perez Arribas, "Derivatives pricing using signature payoffs," 2018, *arXiv:1809.09466*.

[10] H. Ni, "The expected signature of a stochastic process," Ph.D. dissertation, Dept. Math., Oxford Univ., Oxford, U.K., 2012. [Online]. Available: https://ora.ox.ac.uk/objects/uuid:e0b9e045-4c09-4cb7-ace9-46c4984f16f6

[11] I. Chevyrev and T. Lyons, "Characteristic functions of measures on geometric rough paths," *Ann. Probab.*, vol. 44, no. 6, pp. 4049–4082, Nov. 2016, doi: 10.1214/15-aop1068.

[12] I. Chevyrev and H. Oberhauser, "Signature moments to characterize laws of stochastic processes," *J. Mach. Learn. Res.*, vol. 23, no. 42, p. 176, 2022, doi: 10.5486/pmd.1976.23.1-2.24.

[13] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[15] B. K. Iwana and S. Uchida, "An empirical survey of data augmentation for time series classification with neural networks," *PLoS ONE*, vol. 16, no. 7, Jul. 2021, Art. no. e0254841.

[16] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu, "Time series data augmentation for deep learning: A survey," in *Proc. 30th Int. Joint Conf. Artif. Intell. (IJCAI)*, Z.-H. Zhou, Ed., 2021, pp. 4653–4660.

[17] T. Lyons and A. D. McLeod, "Signature methods in machine learning," 2022, *arXiv:2206.14674*.

[18] I. Chevyrev and A. Kormilitzin, "A primer on the signature method in machine learning," 2016, *arXiv:1603.03788*.

[19] P. Kidger, I. P. Arribas, C. Salvi, and T. J. Lyons, "Deep signature transforms," in *Proc. Neural Inf. Process. Syst.*, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/hash/d2cdf047a6674cef251d56544a3cf029-Abstract.html

[20] F. J. Király and H. Oberhauser, "Kernels for sequentially ordered data," *J. Mach. Learn. Res.*, vol. 20, no. 31, pp. 1–45, 2019. [Online]. Available: http://jmlr.org/papers/v20/16-314.html

[21] C. Tóth and H. Oberhauser, "Bayesian learning from sequential data using Gaussian processes with signature covariances," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 9548–9560.

[22] E. Akyildirim, M. Gambara, J. Teichmann, and S. Zhou, "Applications of signature methods to market anomaly detection," 2022, *arXiv:2201.02441*.

[23] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten Chinese text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1903–1917, Aug. 2018.

[24] M. Moor, M. Horn, C. Bock, K. Borgwardt, and B. Rieck, "Path imputation strategies for signature models of irregular time series," in *Proc. ICML Workshop Art Learn. Missing Values*, 2020. [Online]. Available: https://openreview.net/forum?id=P0DL7M6T57o

[25] M. Lemercier, C. Salvi, T. Damoulas, E. V. Bonilla, and T. J. Lyons, "Distribution regression for sequential data," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 3754–3762.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[27] J. Morrill, A. Fermanian, P. Kidger, and T. Lyons, "A generalised signature method for multivariate time series feature extraction," 2020, *arXiv:2006.00873*.

[28] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[29] P. Kidger and T. Lyons, "Signatory: Differentiable computations of the signature and logsignature transforms, on both CPU and GPU," in *Proc. ICLR Poster*, 2021. [Online]. Available: https://iclr.cc/virtual/2021/poster/2566 and https://openreview.net/forum?id=lqU2cs3Zca

[30] F. Triggiano. *A New Signature Model*. Accessed: Oct 13, 2023. [Online]. Available: https://github.com/frtrigg5/A-new-signature-model

[31] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion* (Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]), vol. 293, 3rd ed. Berlin, Germany: Springer-Verlag, 1999, doi: 10.1007/978-3-662-06400-9.

[32] *Time Series Classification Dataset Collection*. Accessed: Oct. 1, 2023. [Online]. Available: http://timeseriesclassification.com/

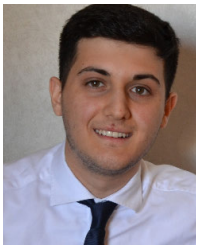
[33] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 358–386, Mar. 2005.

[34] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. Bagnall, "HIVE-COTE 2.0: A new meta ensemble for time series classification," *Mach. Learn.*, vol. 110, nos. 11–12, pp. 3211–3243, Dec. 2021.

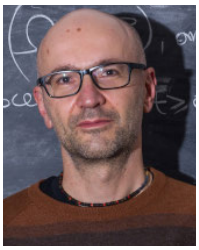
[35] M. Middlehurst, P. Schäfer, and A. Bagnall, "Bake off redux: A review and experimental evaluation of recent time series classification algorithms," *Data Mining Knowl. Discovery*, Apr. 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10618-024-01022-1

[36] T. Cass and E. Ferrucci, "On the Wiener chaos expansion of the signature of a Gaussian process," in *Probability Theory and Related Fields*, 2024, pp. 1–39. [Online]. Available: https://link.springer.com/article/10.1007/s00440-023-01255-z

[37] S. Boucheron, G. Lugosi, and O. Bousquet, "Concentration inequalities," in *Advanced Lectures on Machine Learning* (Lecture Notes in Computer Science), O. Bousquet, U. von Luxburg, and G. ätsch, Eds. Berlin, Germany: Springer, 2004, doi: 10.1007/978-3-540-28650-9_9.



FRANCESCO TRIGGIANO received the master's degree in mathematics from the University of Pisa, in 2022. He is currently pursuing the Ph.D. degree in computational methods and mathematical models for sciences and finance course with Scuola Normale Superiore, Pisa.



MARCO ROMITO received the Ph.D. degree in mathematics from the University of Pisa, in 2001. He joined the Department of Mathematics, University of Florence, as a Lecturer in mathematical analysis. He is currently a Full Professor in probability and mathematical statistics with the University of Pisa. He has been a Visiting Research Member with MSRI Berkeley, HIM Bonn, Newton Institute Cambridge, and Bernoulli Centre Lausanne, and an Invited Professor with ENS Bretagne and the Center for Mathematical Sciences, Wuhan. His research interests include stochastic analysis, stochastic PDEs, and the mathematical theory of turbulence. His research interests include the mathematics of machine learning and artificial intelligence.

Open Access funding provided by 'Università di Pisa' within the CRUI CARE Agreement