Generalised Interpretable Shapelets for Irregular Time Series

Terry Lyons[1,2]

[1] Mathematical Institute, University of Oxford
[2] The Alan Turing Institute, British Library
{tlyons}@maths.ox.ac.uk

Abstract

The shapelet transform is a form of feature extraction for time series, in which a time series is described by its similarity to each of a collection of 'shapelets'. However existing work has suffered from several limitations, such as fragility to noise, loss of interpretability, and a dependence on fully observed and regularly sampled data. In this work, we demonstrate how the shapelet transform may be improved and generalised in multiple ways: by using learnt pseudometrics as a measure of similarity between time series; by demonstrating that a regularisation penalty ensures interpretability; by allowing the length of each shapelet to be learnt differentiably (in contrast to its previously discrete formulation). Furthermore our *generalised shapelet transform* is applicable to the general case of irregularly sampled partially observed multivariate time series. We validate our method by demonstrating its empirical performance on several datasets.

1 Introduction

Shaplets are a form of feature extraction for time series. Given some fixed hyperparameter K, describing how many shapelets we are willing to consider, then each time series is represented by a vector of length K describing how similar it is to each of the k selected shapelets.

We begin by stating the classical definition of shapelets.

1.1 Classical shapelets

Given N regularly sampled multivariate time series, with D observed channels, where the i-th time series is of length T_n, then the n-th time series is a matrix

$$f^n = (f^n_t)_{t\in\{0,\dots,T_n-1\}} = (f^n_{t,d})_{t\in\{0,\dots,T_n-1\},d\in\{1,\dots,d\}}, \tag{1}$$

with each $f^n_{t,d} \in \mathbb{R}$. We assume without loss of generality that $0, \dots, T_n - 1$ are the times at which each sample is observed, so that the parameterisation t corresponds to the time of an observation.

Fix some hyperparameter $K \in \mathbb{N}$, which will describe the number of shapelets. Fix some $S \in \{0, \dots, \min_{i\in\{1,\dots,N\}} T_n - 1\}$, which will describe the length of each shapelet. We define the k-th shapelet as a matrix

$$w^k = (w^k_t)_{t\in\{0,\dots,S-1\}} = (w^k_{t,d})_{t\in\{0,\dots,S-1\},d\in\{1,\dots,d\}},$$

with each $w^k_{t,d} \in \mathbb{R}$.

Preprint. Under review.

Then the discrepancy between f^n and w^k is defined by:

$$\sigma_S(f^n, w^k) = \min_{s \in \{0, \ldots, T_n - S\}} \sum_{t=0}^{S-1} \|f^n_{s+t} - w^k_t\|_2^2, \tag{2}$$

where $\|\cdot\|_2$ describes the L^2 norm on $\mathbb{R}^D$. A small discrepancy implies that f^n and w^k are similar to one another. This corresponds to sweeping w^k over f^n, and finding the offset s at which w^k best matches f^n. The collection of $(\sigma_S(f^n, w^1), \ldots, \sigma_S(f^n, w^K)) \in \mathbb{R}^K$ is now a feature describing f^n. This may now be passed to some model to perform classification or regression.

In [**?**], a procedure is described for selecting shapelets as particular small intervals from particular training samples. However doing so is very expensive, requiring $\mathcal{O}(N^2 \cdot \max_n T_n^4)$ work. As such, [**?**] instead show that the discrepancy σ_S of equation (2) is differentiable with respect to $w^{k,\rho,S}$, and so the shapelets may be selected differentiably[1], as part of an end-to-end optimisation of the final loss function of the model that uses the shapelets as features.

The shapelet method is attractive for three reasons. First, it is invariant to the value of T_n, and as such provides a way to normalise variable-length time series. Second, it is interpretable, as use of a particular feature corresponds to the importance of the similarity to the shapelet w^k, which may for example describe some shape that is characteristic of a particular class; furthermore the value of s gives where the similarity occurs. Third, it typically demonstrates good performance [**?**].

1.2 Limitations

However, classical shapelet methods also suffer from a number of limitations.

1. The technique only applies to regularly spaced time series, due to the minimisation over s.
2. The choice of S is a hyperparameter; it is discrete, and choosing it is thus a relatively expensive optimisation procedure.
3. The technique is not robust to irrelevant channels (which will typically exist in many real world datasets, for example medical time series): equation (2) attempts to fit a $w^k_{t,d}$ even for uninformative channels d.
4. Selecting w^k is either expensive, following the procedure of [**?**], or loses interpretability, following the procedure of [**?**].
5. The formulation of equation (2) has essentially made several ad-hoc choices, for example in the choice of L^2 norm on $\mathbb{R}^D$. Indeed, there are many natural notions of discrepancy between time series [**?**, **?**, **?**, **?**] that do not fit this framework.

1.3 Contributions

We demonstrate how classical shapelets may be generalised in multiple ways, so as to address the collection of limitations just described.

We demonstrate how the discrepancy between a shapelet and a time series may be taken to be a learnt pseudometric; this makes our proposed method robust to noise in unrelated channels, and additionally introduces a great deal of flexibility into the method. Furthermore, we demonstrate how simple regularisation is enough to achieve shapelets that resemble characteristic features of the data, in order to achieve the desired interpretability.

Additionally, by treating the objects in continuous time rather than discrete time, we demonstrate how the length of each shapelet may be learnt, individually for each shapelet, in a differentiable manner. This continuous-time formulation also allows our *generalised shapelet transform* to extend to the general case of irregularly sampled partially observed multivariate time series.

Our code is available at TODO.

[1] Although they include a 'softmin' procedure which we believe to be unnecessary, as the minimum function is already almost everywhere differentiable.

2 Method

We move on to describing our method, which we present in a general form. In the next section we will discuss the specific choices made in our experiments.

2.1 Continuous-time objects

We interpret a time series as a discretised sample from an underlying process, observed only through the time series. Similarly, the shapelet previously constructed may be thought of as a discretisation of some underlying function. The first important step in our procedure is to construct continuous-time approximations to these underlying objects.

Continuous-time path interpolants Formally speaking, we assume that for $n \in \{1, \ldots, N\}$ indexing different time series, we observe a collection of time series

$$f^n = (f^n_{t_\tau})_{\tau \in \{1,\ldots,T_n\}},$$

where $t_\tau \in \mathbb{R}$ is the observation time of $f^n_{t_\tau} \in (\mathbb{R} \cup \{*\})^D$, where $*$ denotes the possibility of a missing observation. This description allows irregularly sampled partially observed time series to be treated on the same footing as regularly sampled and completely observed time series.

Next, interpolate to get a function $\iota(f^n)\colon [0, T_n - 1] \to \mathbb{R}^D$ such that $\iota(f^n)(t_\tau) = f^n_{t_\tau}$ for all $\tau \in \{0, \ldots, T_n - 1\}$ such that $f^n_{t_\tau}$ is observed. There are many possible choices for interpolations, for example splines [**?**], kernel methods [**?**], or Gaussian processes [**?**, **?**]. In our experiments, we use piecewise linear interpolation.

Continuous-time shapelets The shapelets themselves we are free to control, and so for $k \in \{1, \ldots, K\}$ indexing different shapelets, we take each $w^{k,\rho}\colon [0, 1] \to \mathbb{R}^D$ to be some learnt function depending on learnt parameters ρ. For example, this could be an interpolated sequence of learnt points, an expansion in some basis functions, or a neural network. In our experiments we use linear interpolation of a sequence of a learnt points.

Then for some learnt length $S_k > 0$, define $w^{k,\rho,S_k}\colon [0, S_k] \to \mathbb{R}^D$ by

$$w^{k,\rho,S_k}(t) = w^{k,\rho}\left(\frac{t}{S_k}\right).$$

Taking the length S_k to be continuous is a necessary prerequisite to training it differentiably. We will discuss the training procedure in a moment.

2.2 Generalised discrepancy

The core of the shapelet method is that the similarity or discrepancy between f^n and w^{k,ρ,S_k} is important. In general, we approach this by defining a *discrepancy function* between the two, which will typically be learnt, and which we require only to be a pseudometric, rather than a metric. By relaxing to allow pseudometricity, then the procedure becomes robust to noise in unrelated channels, as the learning procedure may learn to ignore extraneous dimensions.

We denote this discrepancy function by $\pi^A_{S_k}$; it depends upon the length S_k and a learnt parameter A, consumes two paths in $\mathbb{R}^D$, and returns a real number describing some notion of closeness between them. We are being deliberately vague about the domain of $\pi^A_{S_k}$, as it is a function space whose regularity will depend on ι.

Given some fixed $\pi^A_{S_k}$, then we define the discrepancy between f^n and w^{k,ρ,S_k} to be given by

$$\sigma^A_{S_k}(f^n, w^{k,\rho,S_k}) = \min_{s \in [0, T_n - S_k]} \pi^A_{S_k}(\iota(f^n)|_{[s,s+S_k]}(s + \cdot), w^{k,\rho,S_k}). \tag{3}$$

The collection of discrepancies $(\sigma^A_{S_k}(f^n, w^{1,\rho,S_k}), \ldots, \sigma^A_{S_k}(f^n, w^{K,\rho,S_k}))$ is now a feature describing f^n, and is invariant to the length T_n. Use of the particular feature $\sigma^A_{S_k}(f^n, w^{k,\rho,S_k})$ corresponds to the importance of the similarity between f^n and w^{k,ρ,S_k}. In this way, the choice of $\pi^A_{S_k}$ gives a great deal of flexibility: not only may it be selected for reasons of classification performance, but it may also be selected to aid interpretability, as we are about to see.

Existing shapelets fit into this framework A simple example, in analogy to the classical shapelet method of equation (2), is to take

$$\pi^A_{S_k}(f, w) = (\int_0^{S_k} \|f(t) - w(t)\|_2^2 \,\mathrm{d}t)^{\frac{1}{2}},$$

which in fact has no A dependence. If ι is taken to be a piecewise constant 'interpolation' then this will exactly correspond to (the square root of) the classical shapelet approach.

Learnt L^2 discrepancies The previous example may be generalised by taking our learnt parameter $A \in \mathbb{R}^{D\times D}$, and then letting

$$\pi^A_{S_k}(f, w) = (\int_0^{S_k} \|A(f(t) - w(t))\|_2^2 \,\mathrm{d}t)^{\frac{1}{2}}. \tag{4}$$

That is, allowing some learnt linear transformation before measuring the discrepancy. As we have allowed pseudometricity, then uninformative dimensions may be shrunk to zero. In our experiments we consider two possible formats for A: a general element of $\mathbb{R}^{D\times D}$, or restricted to be diagonal. Being a general element allows for the possibility of dimensions to interact with one another, but whether this is beneficial is likely to be problem-dependent.

More complicated discrepancies Moving on, we consider other more general choices of discrepancy, which may be motivated by the problem at hand. In particular we will discuss discrepancies based on the logsignature transform [**?**], and mel-frequency cepstrums (MFC) [**?**].

Our exposition on these two discrepancies will be deliberately brief, as the finer details on exactly when and how to use the logsignature and MFC transforms is not important to us here. The point is that our framework has the flexibility to consider general discrepancies motivated by other discplines, or which are known to extract information which is particular useful to the domain in question. An understanding of either logsignatures or mel-frequency cepstral coefficients will not be necessary to follow this paper.

Logsignature discrepancies The logsignature transform is a transform on paths, known to characterise its input whilst extracting statistics which describe how the path controls differential equations [**?**, **?**, **?**, **?**]. Let μ denote the Möbius function, and let

$$\beta_{D,R} = \sum_{r=1}^{R} \frac{1}{r} \sum_{\rho|r} \mu\left(\frac{r}{\rho}\right) D^\rho$$

which is Witt's formula [**?**]. Let

$$\mathrm{LogSig}^R\colon \left\{f\colon [0,T] \to \mathbb{R}^D \;\middle|\; T \in \mathbb{R}, f \text{ is of bounded variation}\right\} \to \mathbb{R}^{\beta_{D,R}}$$

be the depth-R logsignature transform. Let $A \in \mathbb{R}^{\beta_{D,R}\times\beta_{D,R}}$ be full or diagonal as before, and let $\|\cdot\|_p$ be the L^p norm on $\mathbb{R}^{\beta_{D,R}}$. Then we define the *p-logsignature discrepancy* between two functions to be

$$\pi^A_{S_k}(f, w) = \|A(\,\mathrm{LogSig}^R(f) - \,\mathrm{LogSig}^R(w))\|_p. \tag{5}$$

MFC discrepancies The computation of a MFC is a function-to-function map derived from the short-time Fourier transform, using additional processing to focus on information that is particularly relevant to human hearing. Letting MFC represent the computation of an MFC, then we compose this with the L^2 based discrepancy of equation (4) to produce

$$\pi^A_{S_k}(f, w) = (\int_0^{S_k} \|A(\mathrm{MFC}(f)(t) - \mathrm{MFC}(w)(t))\|_2^2 \,\mathrm{d}t)^{\frac{1}{2}}. \tag{6}$$

The generalised shapelet transform Whatever the choice of π^A_S, and in analogy to classical shapelet methods, we call the map

$$f \mapsto (\sigma^A_{S_1}(f, w^{1,\rho,S_1}), \ldots, \sigma^A_{S_K}(f, w^{K,\rho,S_K}))$$

the *generalised shapelet transform*.

2.3 Interpretable regularisation

Selecting shapelets by searching as in [**?**] is incredibly expensive. Selecting them as part of a differentiable optimisation procedure is much more attractive for its speed, and is faciliated by deep learning tools that typically optimise in the same way. However, it has been noted in [**?**] that this method sacrifices much of the interpretability, as the shapelets that are then selected need not look like any small extracts from the training data. In [**?**] they propose to solve this issue via adversarial regularisation.

We instead propose a much simpler method; train differentiably as before, and add on

$$\sum_{k=1}^{K} \min_{n \in \{1,\dots,N\}} \sigma_S^A(f^n, w^{k,\rho,s}) \tag{7}$$

as a regularisation term, so that minimising the discrepancy between f^n and $w^{k,\rho,S}$ is also important. Note the choice of minimisation over n, rather than a sum over n. A sum over n would ask that every shapelet should look like every training sample. Taking a minimum instead asks only that every shapelet should be similar to a single training sample.

2.4 Minimisation objective and training procedure

Overall, suppose we have some parametric (typically linear) model F^θ, some loss function $\mathcal{L}$, and some observed time series $f^1, \dots, f^N$ with targets $y_1, \dots, y_N$.

Then letting $\gamma > 0$ control the amount of regularisation, we propose to minimise

$$\frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(y_n, F^\theta(\sigma_{S_1}^A(f^n, w^{1,\rho,S_1}), \dots, \sigma_{S_K}^A(f^n, w^{K,\rho,S_K}))) + \gamma \sum_{k=1}^{K} \min_{n \in \{1,\dots,N\}} \sigma_{S_k}^A(f^n, w^{k,\rho,S_k}) \tag{8}$$

over model parameters θ, discrepancy parameters A, shapelet parameters ρ, and shapelet lengths S_k, via standard stochastic gradient descent based techniques.

Differentiability Some thought is necessary to verify that our constructions are in fact differentiable, and in particular differentiable with respect to S_k. Examining the definition of $\sigma_{S_k}^A$ in equation (3), there are also two operations that may seem to pose a problem, namely the restriction operator $\iota(f^n) \mapsto \iota(f^n)|_{[s,s+S_k]}$, and the minimum over a range $\min_{s \in [0, T_n - S_k]}$; neither of these are a standard part of an autodifferentiation framework such as PyTorch [**?**].

Practically speaking, however, it is straightforward to resolve all of these issues. The continuous-time paths $\iota(f^n)$ and continuous-time shapelets w^{k,ρ,S_k} must both be represented by some parameterisation of function space, and it is thus sufficient to restrict to considering differentiability with respect to this parameterisation. In our experiments we represent both $\iota(f^n)$ and w^{k,ρ,S_k} as a continuous piecewise linear function stored as a collection of knots.

In this context, the restriction operator is clearly differentiable, as a map from the unrestricted function, represented by one collection of knots, to the restricted function, represented by another collection of knots. Each knot is either kept (the identity function), thrown away (the zero function), or interpolated between to place a new knot at the boundary (a ratio of existing knots), and so collectively the map is differentiable.

For the minimum over a range, this may reasonably be approximated by a minimum over some collection of points $s \in \{0, \varepsilon, 2\varepsilon, \dots, T_n - S_k - \varepsilon, T_n - S_k\}$, for some $\varepsilon > 0$ small and dividing $T_n - S_k$. This is now a standard piece of an autodifferentiation package. The error of this approximation may be controlled by the modulus of continuity of $s \mapsto \pi_{S_k}^A(\iota(f^n)|_{[s,s+S_k]}(s + \cdot), w^{k,\rho,S_k})$, but in practice we found this to be unnecessary, and simply took ε equal to the smallest gap between observations.

Choice of F^θ Interpretability of the model will depend on an interpretable choice of F^θ. In our experiments we took a natural logarithm of every feature, and then used a linear model. The logarithm is because the discrepancies $\sigma_{S_k}^A(f^n, w^{k,\rho,S_k}) \in [0, \infty)$, so the logarithm maps this to $\mathbb{R}$. (As a technical point, a small number such as 10^{-5} may need to first be added to prevent $\log\colon 0 \mapsto -\infty$.)

This is now easily interpretable: a very negative coefficient of $\log \sigma^A_{S_k}(f^n, w^{k,\rho,S_k})$ corresponds to the importance of $\sigma^A_{S_k}(f^n, w^{k,\rho,S_k})$ being close to zero, and thus the importance of f^n and w^{k,ρ,S_k} being similar to each other.

3 Experiments

Our generalised shapelet transform, contrasted with the classical shapelet transform, has two extra degrees of freedom: the choice of interpolation scheme ι, and the choice of discrepancy function π^A_S. In our experiments, we consider π^A_S given by either of equations (4) or (5). Meanwhile, we take ι to be piecewise linear interpolation, because efficient algorithms for computing the logsignature transform only exist for piecewise linear paths [**?**].

TODO: we need to consider more disrepancy functions than this.

3.1 The UEA (Multivariate) Time-Series Archive

We begin by comparing the classification performance of the old method of shapelets to our new 'generalised' approach for which we consider three learnt metrics: a standard L2-metric, an learnt metric and a logsig-3-diagonal. We evaluate the methods on a subset of the UEA time-series archive [**?**]. This contains a wide range of multivariate time-series classification problems from various fields with significant differences in time-series length, number of classes, and amount of training data. The full collection contains 30 datasets, however due to algorithm run-time constraints we have considered a subset of these ensuring they still contain significant variation. The statistics of these dataset are given in Table **??** of the Appendix.

The results are given in Table 1. We see improved performance from the old shapelet method in ? of ? cases. Whats more, we see there are datasets for which the optimal accuracy is achieved with a learnt metric and the logsignature discrepancy. These results, whilst not extensive, show the feasibility of improving performance by discrepancy tailored to the structure of the data as opposed to a simple euclidean distance metric.

Table 1

	Discrepancy		
Dataset	logsig-3diagonal	L2-diagonal	old
AtrialFibrillation	0.333	0.333	0.467
BasicMotions	1.000	0.950	0.975
ERing	0.574	0.919	0.737
FingerMovements	0.450	0.560	0.490
JapaneseVowels	0.646	0.914	0.949
Libras	0.739	0.661	0.661
NATOPS	-	0.861	-
PenDigits	0.967	0.956	0.956
RacketSports	0.717	0.717	0.605
Wins	3	4	2

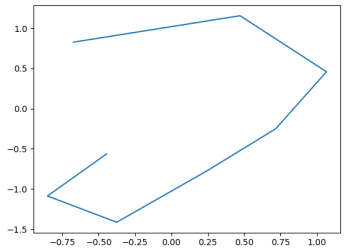
3.1.1 Algorithm Interpretability

TODO: It would be good to include more than 4 images here, we could include one of each number corresponding to the largest logreg value for that class and plot a row from 0-9 for each class. This would actually save space and hopefully demonstrate the interpretable aspect much better. It will also look like we havent just picked out two nice ones and two bad ones if we note they were the largest coefficient for each class.

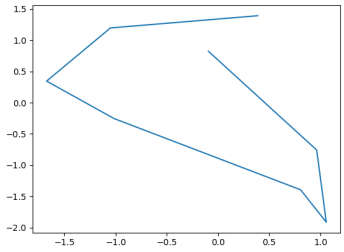
Here we explore the effect of the inclusion of the regularisation term from Equation [**?**] on shapelet interpretability. Recall that the term was chosen to ensure the resulting shapelets are 'close', in the sense of the chosen discrepancy, to some subsample of the training data giving a natural interpretability as representing those such subsamples of the data. To examine this we consider the

PenDigits dataset where participants were asked to write down a number from 0-9 and the goal is to classify the intended digit. In Figure 1 we plot two of the learnt shapelets for both the old algorithm (top) and the regularised algorithm (bottom). For the regularised algorithm one could immediately guess that the first shapelet is designed to distinguish the number two, as it looks like the top half of the number two, and the second clearly looks like a number 7. By looking at the coefficients of the logistic regression, one does indeed find that these are the main classes that these shapelets aim to discriminate. On the other hand it is much less clear what is represented by the old shapelets, it turns out these also aim to represent a two and a seven but it is less clear that this is the case.

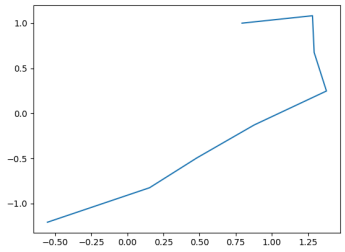
We chose PenDigits here because it is easy to understand visually what the shapelet represents. In general this is not the case, but provided we can make sense of isolated subintervals of the time-series, then the generated results will be interpretable in this context.



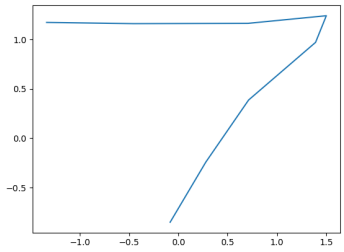
(a) Old 2



(b) Old 7



(c) New 2



(d) New 7

Figure 1: Learnt shapelets from the PenDigits dataset with the old algorithm (top) and the new algorithm (bottom).

3.1.2 Missing Data and Length Ablation

We now demonstrate both the ability of the proposed framework to handle partially observed data, as well as show the effectiveness of having learnable shapelet lengths.

3.2 Speech Commands

Finally we examine the performance on the speech commands dataset **??** which includes a selection of one-second audio files with each representing a single spoken word. The aim is to build a model to detect the word that has been spoken. We chose this dataset as it is significantly larger than any in the UEA archive so as to demonstrate that the method is not restricted to these (relatively) small datasets. We do reiterate that the computational cost of shapelet methods is high in comparison to other more traditional deep-learning approaches and so can make datasets such as this prohibitive, in particular, it is why we have left the logsig-3 discrepancy out of our analysis in this instance as it takes significantly longer to train compared with the L2 methods.

3.2.1 Interpretability of Speech Commands

Pray for me.

Table 2

Dataset	**Discrepancy** L2-diagonal-False	L2-diagonal-True	old
BasicMotions3	0.360 ± 0.175	0.320 ± 0.157	0.520 ± 0.262
BasicMotions30	0.250 ± 0.000	0.280 ± 0.067	0.360 ± 0.213
BasicMotions9	0.365 ± 0.139	0.325 ± 0.075	0.450 ± 0.202
FingerMovements3	0.496 ± 0.009	0.536 ± 0.051	0.498 ± 0.054
FingerMovements30	0.512 ± 0.029	0.506 ± 0.065	0.522 ± 0.044
FingerMovements9	0.504 ± 0.047	0.510 ± 0.031	0.480 ± 0.041
JapaneseVowels3	0.812 ± 0.158	0.725 ± 0.195	0.857 ± 0.027
JapaneseVowels30	0.419 ± 0.221	0.209 ± 0.081	0.611 ± 0.175
JapaneseVowels9	0.432 ± 0.221	0.556 ± 0.324	0.805 ± 0.041
Wins	0	2	7

Table 3: Classification accuracy for old shapelets and new shapelets on the Speech Commands dataset.

Discrepancy Old	L2
-	-
-	-