Random Warping Series: A Random Features Method for Time-Series Embedding

Lingfei Wu
IBM Research

Ian En-Hsu Yen
Carnegie Mellon University

Jinfeng Yi
IBM Research

Fangli Xu
College of William and Mary

Qi Lei
University of Texas at Austin

Michael J. Witbrock
IBM Research

Abstract

Time series data analytics has been a problem of substantial interests for decades, and Dynamic Time Warping (DTW) has been the most widely adopted technique to measure dissimilarity between time series. A number of global-alignment kernels have since been proposed in the spirit of DTW to extend its use to kernel-based estimation method such as support vector machine. However, those kernels suffer from diagonal dominance of the Gram matrix and a quadratic complexity w.r.t. the sample size. In this work, we study a family of alignment-aware positive definite (p.d.) kernels, with its feature embedding given by a distribution of Random Warping Series (RWS). The proposed kernel does not suffer from the issue of diagonal dominance while naturally enjoys a Random Features (RF) approximation, which reduces the computational complexity of existing DTW-based techniques from quadratic to linear in terms of both the number and the length of time-series. We also study the convergence of the RF approximation for the domain of time series of unbounded length. Our extensive experiments on 16 benchmark datasets demonstrate that RWS outperforms or matches state-of-the-art classification and clustering methods in both accuracy and computational time.

1 Introduction

Over the last two decades, time series classification and clustering have received considerable interests in many applications such as genomic research [Leslie et al., 2002], image alignment [Peng et al., 2015b,a], speech recognition [Cuturi et al., 2007, Shimodaira et al., 2001], and motion detection [Li and Prakash, 2011]. One of the main challenges in time series data stems from the fact that there are no explicit features in sequences [Xing et al., 2010]. Therefore, a number of feature representation methods have been proposed recently, among which the approaches deriving features from phase dependent intervals [Deng et al., 2013, Baydogan et al., 2013], phase independent shapelets [Ye and Keogh, 2009, Rakthanmanon and Keogh, 2013], and dictionary based bags of patterns [Senin and Malinchik, 2013, Schäfer, 2015] have gained much popularity due to their highly competitive performance [Bagnall et al., 2016]. However, since the aforementioned approaches only consider the local patterns rather than global properties, the effectiveness of these features largely depends on the underlying characteristics of sequences that may vary significantly across applications. More importantly, these approaches may typically not be a good first choice for large scale time series due to their quadratic complexity in terms both of the number \( N \) and (or) length \( L \) of time series.

Another family of research defines a distance function to measure the similarity between a pair of time series. Although Euclidean distance is a widely used option and has been shown to be competitive with other more complex similarity measures [Wang et al., 2013], various elastic distance measures designed to address the temporal dynamics and time shifts are more appropriate [Xing et al., 2010, Kate, 2016]. Among them, dynamic time warping (DTW) [Berndt and Clifford, 1994] is the standard elastic distance measure for time series. Interestingly, an 1NN classifier with DTW has been
demonstrated as the gold standard benchmark, and has been proved difficult to beat consistently [Wang et al., 2013, Bagnall et al., 2016]. Recently, a thread of research has attempted to directly use the pair-wise DTW distance as features [Hayashi et al., 2005, Gudmundsson et al., 2008, Kate, 2016, Lei et al., 2017]. However, the majority of these approaches have quadratic complexity in both number and length of time series in terms of the computation and memory requirements.

Despite the successes of various explicit feature design, kernel methods have great promise for learning non-linear models by implicitly transforming a simple representations into a high-dimension feature space [Rahimi and Recht, 2007, Chen et al., 2016, Wu et al., 2016, Yen et al., 2014]. The main obstacle for applying kernel method to time series is largely due to two distinct characteristics of time series, (a) variable length; and (b) dynamic time scaling and shifts. Since elastic distance measures, such as DTW, take into account these two issues, there have been several attempts to apply DTW directly as a similarity measure in a kernel-based classification model [Shimodaira et al., 2001, Gudmundsson et al., 2008]. Unfortunately, the DTW distance does not correspond to a valid positive-definite (p.d.) kernel and thus direct use of DTW leads to an indefinite kernel matrix that neither corresponds to a loss minimization problem nor giving a convex optimization problem [Bahlmann et al., 2002, Cuturi et al., 2007, Cuturi, 2011, Marteau and Gibet, 2015]. However, the effectiveness of the global alignment kernels is impaired by the diagonal dominance of the resulting kernel matrix. Also, the quadratic complexity in both the number and length of time series make it hard to scale.

In this paper, inspired by the latest advancement of kernel learning methodology from distance [Wu et al., 2018], we study Random Warping Series (RWS), a generic framework to generate vector representation of time-series, where we construct a family of p.d. kernels from an explicit feature map given by the DTW between original time series and a distribution of random series. To admit an efficient computation of the kernel, we give a random features approximation that uniformly converges to the proposed kernel using a finite number of random series drawn from the distribution. The RWS technique is fully parallelizable, and highly extensible in the sense that the building block DTW can be replaced by recently proposed elastic distance measures such as CID [Batista et al., 2014] and DTDC [Górecki and Łuczak, 2014]. With a number $R$ of random series, RWS can substantially reduce the computational complexity of existing DTW-based techniques from $O(N^2L^2)$ to $O(NRL)$ and memory consumption from $O(NL + N^2)$ to $O(NR)$. We also extend existing analysis of random features to handle time series of unbounded length, showing that $R = O(1/\epsilon^2)$ suffices for the uniform convergence to $\epsilon$ precision of the exact kernel. We evaluate RWS on 16 real-world datasets on which it consistently outperforms or matches state-of-the-art baselines in terms of both testing accuracy and runtime. In particular, RWS often achieves orders-of-magnitude speedup over other methods to achieve the same accuracy.

# DTW and Global Alignment Kernels

We first introduce the widely-used technique DTW and nearest-neighbor DTW (1NN-DTW), and then illustrate the existing global alignment kernels for time series and their disadvantages.

**Time Series Alignment and 1NN-DTW.** Let $\mathcal{X}$ be the domain of input time series, and $\{x_i\}_{i=1}^N$ be the set of time series, where the length of each time series $|x_i| \leq L$, taking numeric values in $\mathbb{R}$. A special challenge in time series lies in the fact that the series could have different lengths, and a signal could be generated with time shifts and different scales, but with a similar pattern. To take these factors into account, an alignment (also called a warping function) is often introduced to provide a better distance/similarity measure between two time series $x_i = (x_{i1}, \ldots, x_{in})$ and $x_j = (x_{j1}, \ldots, x_{jm})$ of lengths $n$ and $m$ respectively. Specifically, an alignment $a = (a_1, a_2)$ of length $|a| = p$ between two time series $x_i$ and $x_j$ is a pair of increasing vectors $(a_1, a_2)$ such that $1 = a_1(1) \leq \ldots \leq a_1(p) = n$ and $1 = a_2(1) \leq \ldots \leq a_2(p) = m$ with unitary increments and no simultaneous repetitions. The set of all alignments between $x_i$ and $x_j$ is defined as $A(x_i, x_j)$.

In the literature of DTW [Berndt and Clifford, 1994], the DTW distance between $x_i$ and $x_j$ is defined as follows in its simplest form:

$$S(x_i, x_j) = \min_{a \in A(x_i, x_j)} \sum_{t=1}^{|a|} \tau(x_i(a_1(t)), x_j(a_2(t))).$$

Here $\tau(x_i, x_j; a)$ is a dissimilarity measure between $x_i$ and $x_j$ under alignment $a$. Typically, Dynamic Programming (DP) is employed to find the optimal alignment $a^*$ and then compute DTW distance. The dissimilarity function $\tau$ could be defined as any commonly used distance such as the squared Euclidean distance. To accelerate the computation and improve the performance, a Sakoe and Chiba band is often used
to constrain the search window size for DTW [Sakoe and Chiba, 1978, Rakthanmanon et al., 2012].

DTW has been widely used for time series classification in combination with the 1NN algorithm, and this combination has been shown to be exceptionally difficult to beat [Wang et al., 2013, Bagnall et al., 2016]. However, there are two disadvantages of 1NN-DTW. First, this method incurs the high computational cost of $O(N^2)$ complexity for computing DTW similarity between all pairs of time series, where each evaluation of DTW without constraints takes $O(L^2)$ computation. Second, Nearest-Neighbor methods often suffer from the problems of high variance. For example, if a class label is determined by a small portion of time series, a Nearest-Neighbor identification on the basis of similarity between two series will be ineffective due to noise and irrelevant information.

**Existing Global Alignment Kernels.** To take the advantage of DTW in other prediction methods based on *Empirical Risk Minimization* (ERM) such as SVM and Logistic Regression, a thread of research has been trying to derive a valid *p.d. kernel* that resembles $S(x_i, x_j)$. A framework for designing such kernel is the time series global-alignment kernel proposed in [Cuturi et al., 2007] and further explored in [Cuturi, 2011]. The kernel replaces the minimum in (1) with a soft minimum that sums over all possible DTW alignments between two series $x_i, x_j$:

\[
k(x_i, x_j) := \sum_{a \in \mathcal{A}(x_i, x_j)} \exp(-\tau(x_i, x_j; a))
\]

\[
= \sum_{a \in \mathcal{A}(x_i, x_j)} \prod_{t=1}^{\infty} \kappa(x_i[a_1(t)], x_j[a_2(t)])
\]

(2)

where $\kappa(\cdot, \cdot)$ is some local similarity function induced from the divergence $\tau$ as $\kappa = \exp(-\tau)$. The function (2) is a p.d. kernel when $\kappa(\cdot, \cdot)$ satisfies certain conditions [Cuturi et al., 2007]. However, it is known that a soft minimum can be orders of magnitude larger than the minimum when summing over exponentially many terms, which results in a serious *diagonally dominant problem* for the kernel (2). In other words, the kernel value between a series to itself $k(x_i, x_i)$ is orders of magnitude larger than other values $k(x_i, x_j)$. Thus in practice, one must take the log of the kernel (2) even though such operation is known to break the p.d. property [Cuturi, 2011]. In addition, the evaluation of kernel (2) requires running DP over all pairs of samples and thus gives a high complexity of $O(N^2L^2)$.

Figure 1: Example of the DTW alignment between original time series of length $L = 512$ and random time series of length $D = 10$.

### 3 Novel Time-Series Kernels via Alignments to Random Series

In this section, we study a new approach to build a family of p.d. kernels for time series based on DTW, inspired by the latest advancement of kernel learning methodology from distance [Wu et al., 2018].

Formally, the kernel is defined by integrating a feature map over a distribution of random time series $p(\omega)$, with each feature produced by alignments between original time series $x$ and random series $\omega$:

\[
k(x, y) = \int_{\omega} p(\omega)\phi_{\omega}(x)\phi_{\omega}(y)d\omega,
\]

where $\phi_{\omega}(x) := \sum_{a \in \mathcal{A}(\omega, x)} p(a|\omega)\tau(\omega, x; a)$. (3)

The kernel (3) enjoys several advantages. First, (3) is a *p.d. kernel* by its construction.

**Proposition 1.** The kernel (3) is positive definite, that is, $\sum_{i=1}^{N} \sum_{j=1}^{N} c_ic_j k(x_i, x_j) \geq 0$ for any $\{c_i : c_i \in \mathbb{R}\}_{i=1}^{N}$ and any $\{x_i \in \mathcal{X}\}_{i=1}^{N}$.

**Proof.** By definition (3), we have

\[
\sum_{i=1}^{N} \sum_{j=1}^{N} c_ic_j k(x_i, x_j) = \sum_{i=1}^{N} \sum_{j=1}^{N} c_ic_j \int_{\omega} p(\omega)\phi_{\omega}(x_i)\phi_{\omega}(y_j)d\omega
\]

\[
= \int_{\omega} p(\omega) \left( \sum_{i=1}^{N} \sum_{j=1}^{N} c_ic_j \phi_{\omega}(x_i)\phi_{\omega}(x_j) \right) d\omega
\]

\[
= \int_{\omega} p(\omega) \left( \sum_{i=1}^{N} c_i \phi_{\omega}(x_i) \right)^2 d\omega \geq 0.
\]

Secondly, by choosing

\[
p(a|\omega) = \begin{cases} 1, & a^* = \arg \min_a \tau(\omega, x; a) \\ 0, & \text{o.w.} \end{cases}
\]

(4)
one can avoid the diagonal dominance problem of the kernel matrix, since the kernel value between two time series \( k(x, y) \) depends only on the correlation of \( \tau(\omega, x; a_2) \) and \( \tau(\omega, y; a_2) \) under their optimal alignments. It thus avoids the dominance of the diagonal terms \( k(x, x) \) caused by the summation over exponentially many alignments. We can interpret the random series \( \omega \) of length \( D \) as the possible shapes of a time series, defined by \( D \) segments, each associated with a random number. Figure 1 gives an example of a random series \( \omega \) of length \( D = 10 \), which divides a time series \( x \) into \( D \) segments and outputs a dissimilarity score as the feature \( \phi_\omega(x) \). The third advantage of (3) is its computational efficiency due to a simple random features approximation. Although the kernel function (3) seems hard to compute, we show that there is a low-dimensional representation of each series \( T(x) \), by which one can efficiently find an approximate solution to that of the exact kernel (3) within \( \epsilon \) precision. This is in contrast to the global-alignment kernel (2), where although one can evaluate the kernel matrix exactly in \( O(N^2L^2) \) time, it is unclear how to efficiently find a low-rank approximation.

### 3.1 Computation of Random Warping Series

Although the kernel (3) does not yield a simple analytic form, it naturally yields a random approximation of the form using a simple MC method,

\[
k(x, y) \approx \langle T(x), T(y) \rangle = \frac{1}{R} \sum_{i=1}^{R} \langle \phi_{\omega_i}(x), \phi_{\omega_i}(y) \rangle.
\]

The feature vector \( T(x) \) is computed using dissimilarity measure \( \tau(\{\omega_i\}_{i=1}^{R}, x) \), where \( \{\omega_i\}_{i=1}^{R} \) is a set of random series of variable length \( D \) with each value drawn from a distribution \( p(\omega) \). In particular, the function \( \tau \) could be any elastic distance measure but without loss of generality we consider DTW as our similarity measure since it has proved to be the most successful metric for time series [Wang et al., 2013, Xi et al., 2006].

Algorithm 1 summarizes the procedure to generate feature vectors for raw time series. There are several comments worth making here. First of all, the distribution of \( p(\omega) \) plays an important role in capturing the global properties of original time-series. Since we explicitly define a kernel from this distribution, it is flexible to search for the best distribution that fits data well for underlying applications. In our experiments, we find the Gaussian distribution is generally applicable for time series from various applications. Specifically, the parameter \( \sigma \) stems from a distribution \( p(\omega) \) that should well capture the characteristics of time series \( \{x_i\}_{i=1}^{N} \). Second, as shown in Figure 1, a short random warping series could typically identify the local patterns as well as global patterns in raw time series. It suggests that there are some optimal alignments that allow short random series to segment raw time series to obtain discriminatory features. In practice, there is no prior information for this optimal alignment and thus we choose to uniformly sample the length of random series between \( [D_{\min}, D_{\max}] \) to give an unbiased estimate of \( D \), where \( D_{\min} = 1 \) is used in our experiments. Additional benefits lie in the fact that random series with variable lengths may simultaneously identify multi-scale patterns hidden in the raw time series.

In addition to giving a practical way to approximate the proposed kernel, applying these random series also enjoys the double benefits of reduced computation and memory consumption. Compared to the family of global alignment kernels [Cuturi et al., 2007, Cuturi, 2011], computing the dense kernel matrix \( K \in \mathbb{R}^{N \times N} \) requires \( O(N^2) \) times evaluation of DTW which usually takes \( O(L^2) \) complexity based on DP. It also needs \( O(NL + N^2) \) to store the original time series and resulting kernel matrix. In contrast, our RWS approximation only requires linear complexity of \( O(NRL) \) computation and \( O(NR) \) storage size, given \( D \) is a small constant. This dramatic reduction in both computation and memory storage empowers much more efficient training and testing when combining with ERM classifiers such as SVM.

### 3.2 Convergence of Random Warping Series

In the following, we extend standard convergence analysis of Random Features [Rahimi and Recht, 2007] from a kernel between two fixed-dimensional vectors to a kernel function measuring similarity between two time series of variable lengths. Note [Wu et al., 2018] has proposed a general analysis for any distance-based kernel through covering number w.r.t. the distance, which however, does not apply directly here since DTW is not a distance metric.

Let \( (A, B) \) be \( l \times D \) and \( l \times L \) matrices that map each
element of ω and x to an element of a DTW alignment path. The feature map of RWS can be expressed as

\[ \phi_ω(x) := \min_{(A,B) \in A(ω,x)} \tau(Aω, Bx) = \frac{1}{n} \sum_{i=1}^{n} \tau([Aω]_i, [Bx]_i). \]

(5)

Note that in practice one can often convert a similarity function into a dissimilarity function to fit into the above setting. The goal is to approximate the kernel \( k(x, y) := \int X p(\omega) \phi_ω(x) \phi_ω(y) dω \) via a sampling approximation \( s_R(x, y) = \frac{1}{R} \sum_{i=1}^{R} \phi_ω(x) \phi_ω(y) \) with \( ω_i \sim p(\omega) \). Note we have \( E[s_R(x, y)] = E_ω[\phi_ω(x) \phi_ω(y)] = k(x, y) \). The question is how many samples \( R \) are needed to guarantee

\[ |s_R(x, y) - k(x, y)| \leq \epsilon \quad \forall x, y \in X \] (6)

In the standard analysis of RF, the required sample size is \( \Omega(d^2 \log \frac{2d \dim(X)}{ε}) \) where \( X \) comprises all \( d \)-dimensional vectors of diameter \( \dim(X) \). The standard analysis does not apply to our case for two reasons: (a) our domain \( X \) contains time series of different lengths, and (b) our kernel involves a minimization (5) over all possible DTW alignments, and thus is not shift-invariant as required in [Rahimi and Recht, 2007]. To obtain a uniform convergence bound that could potentially handle time series of unbounded length, we introduce the notion of minimum shape-preserving length.

**Definition 1.** The Minimum Shape-Preserving Length (MSPL) \( d_ε \) of tolerance \( ε \) is the smallest \( L \) such that \( \forall x \in X, \exists \bar{x} \in \mathbb{R}^L \)

\[ \min_{(A,B) \in A(\bar{x}, x): B = I} \| A\bar{x} - Bx \| \leq ε \] (7)

where \( A(\bar{x}, x) \) is the set of possible alignments between \( \bar{x} \) and \( x \) considered by DTW, and \( I \) is an identity matrix.

In other words, \( d_ε \) defines the smallest length one can compress a time series to with approximation error no more than \( ε \), measured by DTW in the \( l_2 \) distance. Then the following gives the number of RWS required to guarantee an \( ε \) uniform convergence over all possible inputs \( x, y \in X \).

**Theorem 1.** Assume the ground metric \( τ(Aω, Bx) \) satisfies \( |τ(., .)| ≤ γ \) and is Lipschitz-continuous w.r.t. \( x \) with parameter \( β(ω) \) where \( \text{Var}[β(ω)] \leq σ_2^2 \). The RWS approximation with \( R \) features satisfies

\[ P \left[ \max_{x,y \in X} |s_R(x, y) - k(x, y)| ≥ 3ε \right] \leq 8r^2 \left( \frac{4γσ_r}{ε} \right)^2 e^{\frac{ρr^2}{32γ^2(1+δ)}}. \] (8)

where \( r \) is the radius of time series domain \( X \) in the \( l_∞ \) norm and \( d_ε \) is the MSPL with precision \( ε \).

**Proof Sketch.** Let \( f(x, y) := s_R(x, y) - k(x, y) \). We have \( E[f(x, y)] = 0 \) and \( |f(x, y)| ≤ 2γ^2 \) by the boundedness of function \( τ(., .) \). Then by Hoeffding inequality, we have

\[ P[|f(x, y)| ≥ t] ≤ 2 \exp(-Rt^2/8γ^4) \] (9)

for a given pair \( (x, y) \in X \times X \). To get a uniform bound that holds for all pairs of series \((x, y) \in X \times X \), consider the pair of series \((\bar{x}, \bar{y})\) of minimum shape-preserving length \( d(ε) \) under precision \( ε \). We have an \( ε \)-net \( E \) with \( |E| = \frac{(2d)^d}{ε^d} \) that covers the \( d \)-dimensional \( l_∞ \)-ball of radius \( r \). Then through union bound and (9), we have

\[ P \left[ \max_{\tilde{x}, \tilde{y} \in E} |f(\tilde{x}, \tilde{y})| ≥ t \right] ≤ 2|E|^2 \exp(-Rt^2/8γ^4). \] (10)

Let \( B_∞(d) \) be the \( d \)-dimensional \( l_∞ \)-ball. Given any time series \( x, y \in X \) of arbitrary length, we can first find \( \bar{x}, \bar{y} \in B_∞(d) \) with \( ||A\bar{x} - x|| ≤ \epsilon, ||A\bar{y} - y|| ≤ \epsilon \) and then find \( \tilde{x}, \tilde{y} \in E \) such that \( ||\bar{x} - \tilde{x}|| ≤ \epsilon, ||\bar{y} - \tilde{y}|| ≤ \epsilon \). By the result of Lemma 1 (see appendix 6.1), the closeness of \((x, y)\) to \((\tilde{x}, \tilde{y})\) implies the closeness of \( f(x, y) \) to \( f(\tilde{x}, \tilde{y}) \), which leads to

\[ P[|f(x, y) - f(\tilde{x}, \tilde{y})| ≥ 2t] ≤ \frac{8γ^2σ_r^2ε^2}{t^2}. \] (11)

Combining (10) and (11), we have

\[ P \left[ \max_{x,y \in X} |f(x, y)| ≥ 3t \right] ≤ 2 \left( \frac{2γ}{ε} \right)^{2d} e^{-\frac{ρr^2}{32γ^2}} + \frac{16γ^2σ_r^2ε^2}{t^2}. \] (12)

This is of the form \( \kappa_1ε^{-2d} + \kappa_2ε^2 \). Choosing \( ε = (\kappa_1/\kappa_2)^{1/(2+2d)} \) to balance the two terms in (12), the RHS becomes \( 2\kappa_1^{1/(1+d)}d_0^{d/(1+d)} \). This yields the result

\[ P \left[ \max_{x,y \in X} |f(x, y)| ≥ 3t \right] ≤ 8r^2 \left( \frac{4γσ_r}{t} \right)^2 e^{-\frac{ρr^2}{32γ^2(1+δ)}}. \]

The above theorem 1 shows that, to guarantee \( \sup_{x,y \in X} |s_R(x, y) - k(x, y)| ≤ \epsilon \) with probability \( 1 - δ \), it suffices to have \( R = \Omega(d_ε^2 \log \frac{2d \dim(X)}{ε}) \). In practice, the constants \( r, γ \) are not particularly large due to the normalization on series \( x, y \in X \) and dissimilarity function \( τ(., .) \). The main factor determining the rate of convergence is the shape-preserving length \( d_ε \). Note that for problems with time series length bounded by \( L \), we always have \( d_ε ≤ L \), which means the number of features required would be only of order \( R = Ω(L/ε^2) \).

**4 Experiments**

We conduct experiments to demonstrate the efficiency and effectiveness of the RWS, and compare against 9
baselines on 16 real-world datasets from the widely-used UCR time-series classification archive [Chen et al., 2015] as shown in Table 1. We evaluate RWS on the datasets with variable number and length to achieve these goals: 1) competitive or better accuracy for small problems; 2) matches or outperforms other methods in terms of both performance and runtime for middle or large scale tasks. We implement our method in Matlab and use C Mex function 1 for computationally expensive component of DTW. For other methods we use the same routine to promote a fair runtime comparison, similar to [Lei et al., 2017, Paparrizos and Gravano, 2015]. More details about datasets and parameter settings are in Appendix 6.2.

Table 1: Properties of the datasets. The number and the length of time series are sorted increasingly.

<table>
<thead>
<tr>
<th>Name</th>
<th>C:Classes</th>
<th>N:Train</th>
<th>M:Test</th>
<th>L:length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beef</td>
<td>5</td>
<td>30</td>
<td>30</td>
<td>470</td>
</tr>
<tr>
<td>DPTW</td>
<td>6</td>
<td>400</td>
<td>130</td>
<td>80</td>
</tr>
<tr>
<td>IPD</td>
<td>2</td>
<td>67</td>
<td>1,029</td>
<td>24</td>
</tr>
<tr>
<td>PPOAG</td>
<td>3</td>
<td>400</td>
<td>205</td>
<td>80</td>
</tr>
<tr>
<td>MPOC</td>
<td>2</td>
<td>600</td>
<td>291</td>
<td>80</td>
</tr>
<tr>
<td>POC</td>
<td>2</td>
<td>1,800</td>
<td>858</td>
<td>80</td>
</tr>
<tr>
<td>LKA</td>
<td>3</td>
<td>375</td>
<td>375</td>
<td>720</td>
</tr>
<tr>
<td>IWBS</td>
<td>11</td>
<td>220</td>
<td>1,080</td>
<td>256</td>
</tr>
<tr>
<td>TWOP</td>
<td>4</td>
<td>1,000</td>
<td>1,000</td>
<td>128</td>
</tr>
<tr>
<td>ECGST</td>
<td>5</td>
<td>500</td>
<td>4,500</td>
<td>140</td>
</tr>
<tr>
<td>CHCO</td>
<td>3</td>
<td>467</td>
<td>3,840</td>
<td>166</td>
</tr>
<tr>
<td>Wafer</td>
<td>2</td>
<td>1,000</td>
<td>6,174</td>
<td>152</td>
</tr>
<tr>
<td>MALLAT</td>
<td>8</td>
<td>55</td>
<td>2,345</td>
<td>1,024</td>
</tr>
<tr>
<td>FordB</td>
<td>2</td>
<td>3636</td>
<td>810</td>
<td>500</td>
</tr>
<tr>
<td>NIFECG</td>
<td>42</td>
<td>1,800</td>
<td>1,965</td>
<td>750</td>
</tr>
<tr>
<td>HO</td>
<td>2</td>
<td>370</td>
<td>1,000</td>
<td>2,700</td>
</tr>
</tbody>
</table>

4.1 Effects of $\sigma$, $R$ and $D$ on RWS

Setup. We first perform experiments to investigate the characteristics of the RWS method by varying the kernel parameter $\sigma$, the rank $R$ and the length $D$ of random series. Due to limited space, we only show typical results and see Appendix 6.3 for complete ones.

Effects of $\sigma$. It is well known that the choice of the kernel parameter $\sigma$ determines the quality of various kernels. Figure 2 shows that in most cases the training and testing performance curves agree well in the sense that they consistently increase at the beginning, stabilize around $\sigma = 1$ (which corresponds to the standard distribution), and finally decrease in the end. In a few cases like NIFECG, the optimal performance is slightly shifted from $\sigma = 1$. This observation is favorable since it suggests that one may easily tune our approach over a smaller interval around $\sigma = 1$ for good performance.

Effects of $R$. We evaluate the training and testing performance when varying the rank $R$ from 4 to 512 with fixed $\sigma$ and $D$. Figure 3 shows that the training and testing accuracy generally converge almost exponentially when increasing $R$ from very small number ($R = 4$) to a relative large number ($R = 64$), and then slowly saturate to the optimal performance. Empirically, this feature is the most favorable because the performance of RWS is relatively stable even for small $R$. More importantly, this confirms our analysis in Theorem 1 that our RWS approximation can guarantee (rapid) convergence to the exact kernel.

Effects of $D$. We investigate the effect of the length $D$ of the random series on training and testing performance. As hinted at earlier, a key insight behind the proposed time-series kernel depends on the assumption that a random series of short length can effectively segment raw time series in a way that captures its patterns. Figure 4 shows that although testing accuracy seems to fluctuate when varying $D_{\text{max}}$ from 10 to 100, it is clear that the near-peak performance can be achieved when $D_{\text{max}}$ is small in the most of cases.

4.2 Comparing Feature Representations

Baselines and Setup. We compare our approach with two recently developed methods: 1) TSEigen [Hayashi et al., 2005]: learn a low-rank feature representation for a similarity matrix computed using DTW distance through Singular Value Decomposition [Wu and Stathopoulos, 2015, Wu et al., 2017]; 2) TSMC [Lei et al., 2017]: a recently proposed similarity preserving representation for a similarity matrix computed using DTW. We set $R = 32$ for all methods. We employ a linear SVM implemented in LIBLINEAR [Fan et al., 2008] since it can separate the effectiveness of the feature representation from the power of the nonlinear learning solvers.

Results. Table 2 clearly demonstrates the significant advantages of our approach compared to other representations in terms of both classification accuracy and computational time. Indeed, TSMC improves the

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accu</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>RWS</td>
<td>0.733</td>
<td>0.3</td>
</tr>
<tr>
<td>TSEigen</td>
<td>0.79</td>
<td>0.5</td>
</tr>
<tr>
<td>TSMC</td>
<td>0.969</td>
<td>0.3</td>
</tr>
<tr>
<td>MPOC</td>
<td>0.868</td>
<td>0.4</td>
</tr>
<tr>
<td>POC</td>
<td>0.711</td>
<td>0.8</td>
</tr>
<tr>
<td>CHCO</td>
<td>0.792</td>
<td>7.3</td>
</tr>
<tr>
<td>IWBS</td>
<td>0.619</td>
<td>8.9</td>
</tr>
<tr>
<td>TWOP</td>
<td>0.999</td>
<td>9.6</td>
</tr>
<tr>
<td>ECGST</td>
<td>0.933</td>
<td>10.6</td>
</tr>
<tr>
<td>CHCO</td>
<td>0.572</td>
<td>6.3</td>
</tr>
<tr>
<td>Wafer</td>
<td>0.993</td>
<td>9.6</td>
</tr>
<tr>
<td>MALLAT</td>
<td>0.937</td>
<td>33.9</td>
</tr>
<tr>
<td>FordB</td>
<td>0.727</td>
<td>43.3</td>
</tr>
<tr>
<td>NIFECG</td>
<td>0.907</td>
<td>19.8</td>
</tr>
<tr>
<td>HO</td>
<td>0.843</td>
<td>43.3</td>
</tr>
</tbody>
</table>
Figure 2: Train (Blue) and test (Red) accuracy when varying $\sigma$ with fixed $D$ and $R$.

Figure 3: Train (Blue) and test (Red) accuracy when varying $R$ with fixed $\sigma$ and $D$.

Figure 4: Train (Blue) and test (Red) accuracy when varying $D$ with fixed $\sigma$ and $R$.

computational efficiency compared to TSEigen without compromising large loss of the accuracy as claimed in [Lei et al., 2017]. However, RWS is corroborated to achieve both higher accuracy and faster train and testing time compared to TSMC and TSEigen. The improved accuracy of RWS suggests that a truly p.d. time series kernel admits better feature representations than those obtained from a similarity or kernel (not p.d.) matrix. In addition, improved computational time illustrates the effectiveness of using random series to approximate the exact kernel.

4.3 Comparing Time-Series Classification

Baselines. We now compare our method with other state-of-the-art time series classification methods that also take advantage of DTW distance or employ DTW-like kernels: 1) 1NN-DTW: use window size $min(L/10, 40)$; 2) 1NN-DTW$_{opt}$: use optimal window size using leave-one-out cross validation from test data in [Chen et al., 2015] 3) DTWF [Kate, 2016]: a recently proposed method that combines DTW without and with constraints and SAX [Lin et al., 2007] as features; 4) TGAK [Cuturi, 2011]: a fast triangular global alignment kernel for time-series; 5) RWS(LR): RWS with large rank that achieves the best accuracy with more computational time; 6) RWS(SR): small rank that obtains comparable accuracy in less time. We conduct grid search for important parameters in each method suggested in [Kate, 2016, Cuturi, 2011].

Results. Table 3 corroborates that RWS consistently outperforms or matches other state-of-the-art methods in terms of testing accuracy while requiring significantly less computational time. First, RWS(SR) can achieve better or similar performance compared to 1NN-DTW and 1NN-DTW$_{opt}$ for all datasets. This is a strong sign that our learned feature representation is very effective, since, using it, even a linear SVM can beat the well-recognized benchmark. Meanwhile, the clear computational advantages of RWS over 1NN-DTW can be observed when the number or the length of time series samples become large. This is not surprising since
Table 3: Classification performance comparison among methods using DTW or DTW-like kernels.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RWS(LR)</th>
<th>RWS(SR)</th>
<th>1NN-DTW</th>
<th>1NN-DTW*</th>
<th>TGAK</th>
<th>DTWF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beef</td>
<td>0.767</td>
<td>0.8</td>
<td>0.733</td>
<td>0.8</td>
<td>0.663</td>
<td>24.7</td>
</tr>
<tr>
<td>DPTW</td>
<td>0.865</td>
<td>4.2</td>
<td>0.80</td>
<td>0.2</td>
<td>0.738</td>
<td>27.9</td>
</tr>
<tr>
<td>IPD</td>
<td>0.985</td>
<td>1.0</td>
<td>0.962</td>
<td>0.4</td>
<td>0.947</td>
<td>55.3</td>
</tr>
<tr>
<td>PPOAG</td>
<td>0.868</td>
<td>0.3</td>
<td>0.859</td>
<td>0.2</td>
<td>0.776</td>
<td>2.0</td>
</tr>
<tr>
<td>MPOC</td>
<td>0.773</td>
<td>6.8</td>
<td>0.708</td>
<td>0.8</td>
<td>0.635</td>
<td>4.4</td>
</tr>
<tr>
<td>POC</td>
<td>0.815</td>
<td>38.2</td>
<td>0.746</td>
<td>4.7</td>
<td>0.721</td>
<td>36.9</td>
</tr>
<tr>
<td>LKA</td>
<td>0.84</td>
<td>5.9</td>
<td>0.816</td>
<td>13.6</td>
<td>0.712</td>
<td>97.7</td>
</tr>
<tr>
<td>IWBS</td>
<td>0.641</td>
<td>132.4</td>
<td>0.619</td>
<td>8.8</td>
<td>0.504</td>
<td>70.9</td>
</tr>
<tr>
<td>TWOP</td>
<td>1</td>
<td>16.1</td>
<td>0.999</td>
<td>4.4</td>
<td>1</td>
<td>222.2</td>
</tr>
<tr>
<td>ECGST</td>
<td>0.94</td>
<td>9.2</td>
<td>0.933</td>
<td>4.9</td>
<td>0.928</td>
<td>137.8</td>
</tr>
<tr>
<td>CHCO</td>
<td>0.777</td>
<td>189.1</td>
<td>0.683</td>
<td>48.1</td>
<td>0.627</td>
<td>160.8</td>
</tr>
<tr>
<td>Wafer</td>
<td>0.995</td>
<td>143.6</td>
<td>0.993</td>
<td>9.6</td>
<td>0.986</td>
<td>412.3</td>
</tr>
<tr>
<td>MALLAT</td>
<td>0.952</td>
<td>72.8</td>
<td>0.937</td>
<td>33.8</td>
<td>0.937</td>
<td>150.3</td>
</tr>
<tr>
<td>FordB</td>
<td>0.783</td>
<td>543.8</td>
<td>0.62</td>
<td>5.6</td>
<td>0.589</td>
<td>1476</td>
</tr>
<tr>
<td>NIFECCG</td>
<td>0.936</td>
<td>140.2</td>
<td>0.903</td>
<td>20.0</td>
<td>0.845</td>
<td>2699</td>
</tr>
<tr>
<td>HO</td>
<td>0.871</td>
<td>336.9</td>
<td>0.834</td>
<td>41.9</td>
<td>0.816</td>
<td>4883</td>
</tr>
</tbody>
</table>

Table 4: Clustering performance comparison among different methods.

<table>
<thead>
<tr>
<th>Clustering</th>
<th>RWS(LR)</th>
<th>RWS(SR)</th>
<th>KMeans-DTW</th>
<th>CLDS</th>
<th>K-Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>NMI</td>
<td>Time</td>
<td>NMI</td>
<td>Time</td>
<td>NMI</td>
</tr>
<tr>
<td>Beef</td>
<td>0.29</td>
<td>1.1</td>
<td>0.27</td>
<td>1.0</td>
<td>0.25</td>
</tr>
<tr>
<td>DPTW</td>
<td>0.52</td>
<td>0.6</td>
<td>0.56</td>
<td>0.5</td>
<td>0.50</td>
</tr>
<tr>
<td>PPOAG</td>
<td>0.56</td>
<td>0.5</td>
<td>0.54</td>
<td>0.5</td>
<td>0.44</td>
</tr>
<tr>
<td>IWBS</td>
<td>0.43</td>
<td>43.9</td>
<td>0.36</td>
<td>6.3</td>
<td>0.37</td>
</tr>
<tr>
<td>TWOP</td>
<td>0.23</td>
<td>11.2</td>
<td>0.3</td>
<td>4.7</td>
<td>0.12</td>
</tr>
<tr>
<td>ECGST</td>
<td>0.46</td>
<td>25.7</td>
<td>0.4</td>
<td>7.0</td>
<td>0.48</td>
</tr>
<tr>
<td>MALLAT</td>
<td>0.92</td>
<td>48.2</td>
<td>0.91</td>
<td>25.4</td>
<td>0.72</td>
</tr>
<tr>
<td>NIFECCG</td>
<td>0.71</td>
<td>346.1</td>
<td>0.68</td>
<td>43.7</td>
<td>0.63</td>
</tr>
</tbody>
</table>

RWS reduces both number and length of time series from quadratic complexity to linear complexity. Second, RWS is much better than another family of time series kernels represented by TGAK, which probably indicates that considering the soft-minimum of all alignment distances does not capture well hidden patterns of time series. Third, DTWF shows significant performance difference compared to 1NN-DTW, which is consistent with the reported results in [Kate, 2016]. However, compared to DTWF, RWS(LR) can still show clear advantages in accuracy among 11 cases out of the total 16 datasets while achieving one or two orders of magnitude speedup. More importantly, RWS can support a trade-off between the accuracy and run-time. This feature is highly desirable in real applications that may have a variety of priorities and constraints.

4.4 Comparing Time-Series Clustering

Baselines. We compare our method against several time-series clustering baselines: 1) KMeans-DTW [Petitjean et al., 2011, Paparrizos and Gravano, 2015]: accelerate computation with lower bounding approach LBKeogh [Keogh, 2002]; 2) CLDS [Li and Prakash, 2011]: learns a feature representation with R hidden variables through complex-valued linear dynamical systems; 3) K-Shape [Paparrizos and Gravano, 2015]: recently proposed clustering method demonstrated to outperform state-of-the-art clustering approaches in accuracy and computational time; 4) RWS(LR); 5) RWS(SR). We combine our learned feature representation with the classic KMeans algorithm [Hartigan and Wong, 1979]. We employ a commonly used clustering metric, the normalized mutual information (NMI) scaling between 0 and 1) to measure the performance, where higher values indicate better accuracy.

Results. Table 4 shows that RWS provides similar or better performance and typically is substantially faster than KMeans-DTW when the number or the length of time-series become large. In addition, RWS can consistently outperform CLDS in terms of both accuracy and runtime. Interestingly, even compared to the state-of-the-art method K-Shape, RWS can still yield a clear advantage in terms of accuracy; RWS yields 5 wins, 1 even, and 2 loses over K-Shape for 8 datasets. Besides its accuracy, the better computational efficiency of RWS over K-Shape is also corroborated.

5 Conclusions and Future Work

In this work, we have studied an effective and scalable time-series (p.d.) kernel for large-scale time series problems based on RWS approximation, and the feature embedding generated by the technique is generally applicable to most of learning problems. There are several interesting directions of future work, including: i) studying the effects of different random time-series distribution p(u) and ii) exploring more elastic dissimilarity measure between time series such as CID and DTDC.
References


Xi Peng, Junzhou Huang, Qiong Hu, Shaoing Zhang, Ahmed Elgammal, and Dimitris Metaxas. From cir-
Random Warping Series: A Random Features Method for Time-Series Embedding


