Fast Global Alignment Kernels

Marco Cuturi MCUTURI@I.KYOTO-U.AC.JP
Graduate School of Informatics, Kyoto University

Abstract

We propose novel approaches to cast the widely-used family of Dynamic Time Warping (DTW) distances and similarities as positive definite kernels for time series. To this effect, we provide new theoretical insights on the family of Global Alignment kernels introduced by Cuturi et al. (2007) and propose alternative kernels which are both positive definite and faster to compute. We provide experimental evidence that these alternatives are both faster and more efficient in classification tasks than other kernels based on the DTW formalism.

1. Introduction

Kernel methods (Hofmann et al., 2008) have been successfully applied in the last decade on a large variety of datatypes, such as images (Harchaoui & Bach, 2007), graphs (Vishwanathan et al., 2008) or strings on finite alphabets (Sonnenburg et al., 2007) to quote but a few recent references. Paired with well understood tools such as the SVM, the definition of a kernel on structured objects can simplify considerably the analysis of challenging datasets. Our goal in this paper is to follow this line of research and provide practitioners with versatile kernels to compare time series. We argue that this remains an important subject because, despite their ubiquity in science and technology, time series as a general datatype have been comparatively the subject of less study in the kernel literature.

Despite notable attempts by Jebara et al. (2004, §4.5) and Vishwanathan et al. (2007) to use probabilistic arguments to define positive definite (p.d.) kernels, the gold standard to compare time series remains the Dynamic Time Warping (DTW) distance. The DTW framework has been extensively studied since it was first proposed by Sakoe & Chiba (1970) and used since in thousands of application papers. Unfortunately, the DTW distance cannot be used rigorously within the kernel methods framework. Indeed, the DTW distance is not rigorously a distance and is known not to be negative definite since it does not satisfy the triangle inequality (Bahlmann et al., 2002; Müller, 2007, p.72), and as a result cannot be used to define a p.d. kernel. The lack of positive definiteness in a kernel contradicts most of the mathematical foundations of kernel methods, from the theory of reproducing kernel Hilbert spaces to their convex optimization machinery.

The DTW distance has been used nonetheless with kernel machines by many authors (Bahlmann et al., 2002; Shimodaira et al., 2002; Zhou et al., 2010, *etc.*) who correct numerically any deficiency of the Gram matrices produced by DTW distances or more simply consider their square (Gudmundsson et al.). Other authors have taken some liberties with the original definition of the DTW distance in order to define p.d. kernels. For instance, Hayashi et al. (2005) propose to embed time series in Euclidean spaces such that the distance of such representations approximates that of the DTW. Cuturi et al. (2007) use the soft-minimum (rather than the minimum) of the costs of all the alignments that can map a time series onto another to define a positive definite kernel.

None of these references consider the most significant limitation of DTW distances, namely their quadratic computational complexity, which scales in $O(nm)$ with the lengths n and m of the time series to be compared. This paper builds upon the family of Global Alignment (GA) kernels (Cuturi et al., 2007) to propose DTW-inspired kernels that are fast to compute and positive definite. We start this paper with a brief review of global alignment kernels in Section 2 and follow in Section 3 with new results and insights for these kernels. Section 4 presents the faster variations we coin down as triangular global alignment kernels. We conclude this paper with promising experimental results in Section 5.

Appearing in *Proceedings of the 28^{th} International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

2. Review of Global Alignment Kernels

Global alignment kernels have been relatively successful in different application fields (Joder et al., 2009; Ricci et al., 2010; de Vries & van Someren, 2010) and shown to be competitive when compared with other time series kernels.

2.1. Alignments and the DTW framework

Let $\mathcal{X}^{\mathbb{N}}$ be the set of discrete-time time series taking values in an arbitrary space $\mathcal{X}$. An alignment π between two time series $\mathbf{x} = (x_1, \cdots, x_n)$ and $\mathbf{y} = (y_1, \cdots, y_m)$ of lengths n and m respectively is a pair of increasing integral vectors (π_1, π_2) of length $p \leq n+m-1$ such that $1 = \pi_1(1) \leq \cdots \leq \pi_1(p) = n$ and $1 = \pi_2(1) \leq \cdots \leq \pi_2(p) = m$, with unitary increments and no simultaneous repetitions. Namely, for all indices $1 \leq i \leq p-1$, the increment vector of π belongs to a set of 3 elementary moves which can be represented as $\rightarrow, \uparrow$ and $\nearrow$ moves,

$$\begin{pmatrix} \pi_1(i+1) - \pi_1(i) \\ \pi_2(i+1) - \pi_2(i) \end{pmatrix} \in \left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}. \qquad (1)$$

The two coordinates π_1 and π_2 of the alignment π are also known in the DTW literature as warping functions. Note that alignments are only constrained by n and m, the respective lengths of $\mathbf{x}$ and $\mathbf{y}$. We write $\mathcal{A}(n,m)$ for the set of all alignments between two time series of length n and m. In its simplest form the DTW distance between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathrm{DTW}(\mathbf{x},\mathbf{y}) \overset{\text{def}}{=} \min_{\pi \in \mathcal{A}(n,m)} D_{\mathbf{x},\mathbf{y}}(\pi), \qquad (2)$$

where, writing $|\pi|$ for the length of π, the cost

$$D_{\mathbf{x},\mathbf{y}}(\pi) \overset{\text{def}}{=} \sum_{i=1}^{|\pi|} \varphi\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right), \qquad (3)$$

is defined by a local divergence φ that measures the discrepancy between any two points x_i and y_j observed in $\mathbf{x}$ and $\mathbf{y}$. When $\mathcal{X} = \mathbb{R}^d$, φ can be typically defined as the squared Euclidean distance $\varphi(x,y) = \|x-y\|^2$.

2.2. Soft-Minimum of All Alignment Scores

The Global Alignment (GA) kernel is defined as the exponentiated *soft-minimum* of all alignment distances,

$$k_{\mathrm{GA}}(\mathbf{x},\mathbf{y}) \overset{\text{def}}{=} \sum_{\pi \in \mathcal{A}(n,m)} e^{-D_{\mathbf{x},\mathbf{y}}(\pi)}. \qquad (4)$$

Equation (4) can be rewritten using the local similarity function κ induced from the divergence φ as $\kappa \overset{\text{def}}{=} e^{-\varphi}$:

$$k_{\mathrm{GA}}(\mathbf{x},\mathbf{y}) = \sum_{\pi \in \mathcal{A}(n,m)} \prod_{i=1}^{|\pi|} \kappa(x_{\pi_1(i)}, y_{\pi_2(i)}).$$

Cuturi et al. (2007) argue that the similarity described by k_{GA} incorporates the whole spectrum of costs $\{D_{\mathbf{x},\mathbf{y}}(\pi), \pi \in \mathcal{A}(n,m)\}$ and provides thus a richer statistic than the minimum of that set, which is the sole quantity considered by the DTW distance. They also prove or state the following results:

1. k_{GA} is p.d. if $\kappa/(1+\kappa)$ is p.d. on $\mathcal{X}$.
2. k_{GA} is likely to yield diagonally dominant Gram matrices when used on real-life datasets.
3. The computational effort required to compute k_{GA} scales in $O(mn)$, similar to the DTW distance. More precisely, the value of $k_{\mathrm{GA}}(X,Y)$ is equal to $M_{n,m}$ where the coefficients $M_{i,j}$ are defined by the boundary values $M_{0,0} = 1, M_{0,i} = M_{j,0} = 0$ and the recurrence

$$M_{i,j} = \kappa(x_i, y_j)\left(M_{i-1,j-1} + M_{i,j-1} + M_{i-1,j}\right) \quad (5)$$

We study more closely statements 1 and 2 in Section 3 and propose a variation of the GA kernel in Section 4 that has lower complexity than $O(mn)$.

3. On Some Issues Raised by GA Kernels

We recall that a positive definite kernel function κ is infinitely divisible if for all $n \in \mathbb{N}$, $\kappa^{1/n}$ is also p.d. (Berg et al., 1984, §3.2.6). For kernels κ that take positive values, κ is infinitely divisible if and only if $-\log(\kappa)$ is negative definite (n.d.). For a p.d. kernel κ we write $\tilde{\kappa}(x,y)$ for its normalized counterpart $\kappa(x,y)/\sqrt{\kappa(x,x)\kappa(y,y)}$.

3.1. Diagonal dominance of k_{GA}

Cuturi et al. (2007) conjecture that k_{GA} will produce diagonally dominant Gram matrices K on most datasets in the sense that the sum of the magnitude of off-diagonal entries of such Gram matrices is far smaller than their trace. Diagonal dominance of Gram matrices is an undesirable property, since it implies that all points in a training set are orthogonal to each other in the corresponding feature space. This conjecture has shed some doubts on the practical applicability of global alignment kernels (Gudmundsson et al., p.2774) which we would like to dissipate in this section by showing that diagonal dominance can in fact be avoided in most practical cases.

We assume that the divergence φ is null on the diagonal, namely that $\varphi(x,x) = 0$ for any $x \in \mathcal{X}$. Let κ be modified to incorporate an exponent $\lambda > 0$, that is $\kappa \overset{\text{def}}{=} e^{-\lambda\varphi}$. Obviously, $k_{\mathrm{GA}}(\mathbf{x},\mathbf{y}) \underset{\lambda\to 0}{\rightarrow} \operatorname{card} \mathcal{A}(n,m)$.

The cardinal of $\mathcal{A}(n,m)$ is known as the Delannoy number $D(n,m)$ (Sulanke, 2003; Banderier & Schwer, 2005) and thus $k_{\mathrm{GA}}(\mathbf{x},\mathbf{y})$ becomes in the limit a kernel exclusively defined by the lengths of $\mathbf{x}$ and $\mathbf{y}$. Two cases may thus arise when applying k_{GA} on a sample $\{X_1,\cdots,X_p\}$ of time series,

- All time series X_i have same length n. The $p\times p$ Gram matrix generated by k_{GA} varies between I_p as λ goes to infinity, to $D(n,n)\mathbf{1}_{p,p}$ when λ is set to 0. Diagonal dominance, if any, can be easily corrected for with a smaller λ value.
- The lengths of all time series vary freely and are upper-bounded by an arbitrary length n. In this case k_{GA} varies between I_p in the limit $\lambda\to\infty$ and a $p\times p$ matrix whose entries are sampled in the Delannoy matrix, $D(|X_i|,|X_j|), 1\leq i,j\leq n$.

Therefore, diagonal dominance only arises as an issue when the sequences to be compared have different lengths. To illustrate this point, we investigate the case where $p=n$ sequences are compared, with length spanning 1 to n and study the diagonal dominance of the matrix of Delannoy numbers $\mathbf{D}=[D(i,j)]_{1\leq i,j\leq n}$ in Lemma 1 below.

The combinatorics literature provides a few useful results (Sulanke, 2003) to prove this lemma. Among them, the formula $D(i,j)=\sum_{k=1}^{\infty}2^k\binom{i}{k}\binom{j}{k}$ highlights the fact that D is a positive definite kernel on integers, with infinite but sparse feature map $\left(2^{k/2}\binom{i}{k}\right)_{k\in\mathbb{N}}$. Central Delannoy numbers $D_k\overset{\mathrm{def}}{=}D(k,k)$, which appear in the diagonal of $\mathbf{D}$, are also known to follow the recurrence

$$D_{k+1}=\frac{3(2k+1)}{k+1}D_k-\frac{k}{k+1}D_{k-1}.$$

Finally, we will use the inequality $D_{k+1}>3D_k$ which can be derived by considering Equation (5) and setting $\lambda=0$, yielding $D(i+1,j+1)=D(i,j)+D(i+1,j)+D(i,j+1)$. Lemma 1 shows that $\mathbf{D}$ is not diagonally dominant, since the sum of its off-diagonal coefficients is comparable to its trace.

Lemma 1. $$\sum_{i,j=1,i\neq j}^{n}D(i,j)>\left(1-\frac{n}{9n-1}\right)\sum_{i=1}^{n}D_i.$$

Proof. The sum of all coefficients of $\mathbf{D}$ can be expressed as a central Delannoy number,

$$\sum_{i,j=1}^{n}D(i,j)=\sum_{k=1}^{\infty}2^k\cdot\textstyle\sum_{i=1}^{n}\binom{i}{k}\cdot\sum_{j=1}^{n}\binom{j}{k}$$
$$=\sum_{k=1}^{\infty}2^k\binom{n+1}{k+1}^2=\frac{D_{n+1}-1}{2}.$$

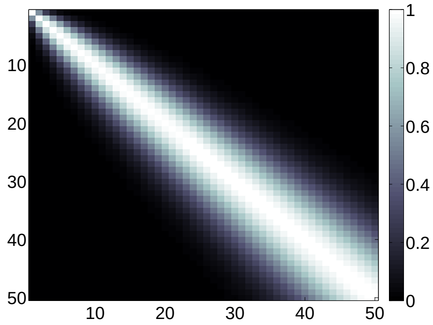

Figure 1. Normalized Delannoy numbers $\tilde{D}(i,j)$, $i,j\leq 50$.

Consider now the ratio ρ between this sum and the trace of the Delannoy matrix $\mathbf{D}$,

$$\rho=\frac{\frac{D_{n+1}-1}{2}}{\operatorname{tr}\mathbf{D}}=\frac{3\left(1-\frac{1}{2(n+1)}\right)D_n-\left(1-\frac{1}{2(n+1)}\right)D_{n-1}}{D_n+D_{n-1}+\sum_{k=1}^{n-2}D_k}.$$

Since $\sum_{k=1}^{n-2}D_k<\sum_{k,l=1}^{n-2}D(k,l)=\frac{1}{2}D_{n-1}-\frac{1}{2}$,

$$\rho>\frac{3\left(1-\frac{1}{2(n+1)}\right)D_n-\left(1-\frac{1}{2(n+1)}\right)D_{n-1}}{D_n+\frac{3}{2}D_{n-1}-\frac{1}{2}}.$$

One can check that the two derivatives of the map f_n defined on $[1,\infty)^2$ as

$$f_n:(x,y)\mapsto\frac{3\left(1-\frac{1}{2(n+1)}\right)x-\left(1-\frac{1}{2(n+1)}\right)y}{x+\frac{3}{2}y-\frac{1}{2}}$$

are positive. f_n is thus strictly increasing in both x and y. Let T be a constant such that $D_{n-1}>T$. Since $D_n>3D_{n-1}>3T$ we have that $f_n(D_n,D_{n-1})>f_n(3T,T)$. Setting $T=n+1$ gives $\rho>2-\frac{n}{9n-1}$ which closes the proof. ■

Numerical evidence shows that the ratio between off-diagonal and diagonal coefficients of the Delannoy matrix is slightly higher than the bound $\approx 8/9$ given here. This ratio is for instance ≈ 1.4 for $n=100$. For illustration purposes we plot in Figure 1 the values of the normalized Delannoy numbers $\tilde{D}(i,j)\overset{\mathrm{def}}{=}D(i,j)/\sqrt{D(i,i)D(j,j)}$ for $i,j\leq 50$ which would appear if one were to use the normalized kernel $\tilde{k}_{\mathrm{GA}}$ with $\lambda=0$ on a dataset of time series of length 1 to n. We thus observe empirically that for $\lambda\approx 0$, $\tilde{k}_{\mathrm{GA}}(\mathbf{x},\mathbf{y})$ is not negligible when $\frac{1}{2}\leq n/m\leq 2$. To conclude this section, we have thus shown that for a properly chosen λ, GA kernels can compare sequences as long as they share similar lengths, in the sense that one is not longer than twice the length of the other.

3.2. Geometric Divisibility of Local Kernels

Cuturi et al. (2007, Theorem 1) prove that a GA kernel k_{GA} defined through a local kernel κ is positive definite if $\kappa/(1+\kappa)$ is positive definite. One may wonder whether this result cannot be generalized to *all* positive definite kernels κ, regardless of the positive definiteness of $\kappa/(1+\kappa)$. We use the theory of mapping kernels to give a negative answer to this question.

Mapping Kernels Shin & Kuboyama (2008) have recently generalized Haussler's (1999) approach to define kernels on discrete structures by introducing mapping kernels, which are kernels of the form

$$k(\mathbf{x},\mathbf{y}) = \sum_{(x,y)\in\mathcal{M}(\mathbf{x},\mathbf{y})} \mathbf{k}(x,y),$$

where $\mathbf{k}$ is a local kernel on substructures of $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{X}$ and mapping sets $\mathcal{M}$ are set-valued functions such that $\mathcal{M}(\mathbf{x},\mathbf{y}) \subset \mathcal{X}^2$. The GA kernel k_{GA} is a mapping kernel when $\mathcal{M}$ and $\mathbf{k}$ are defined with alignments

$$\mathcal{M}_{\text{GA}}(\mathbf{x},\mathbf{y}) = \{(\mathbf{x}_{\pi_1},\mathbf{y}_{\pi_2}) \,|\, \pi = (\pi_1,\pi_2) \in \mathcal{A}(n,m)\},$$

$$\mathbf{k}(\mathbf{x}_{\pi_1},\mathbf{y}_{\pi_2}) = \prod_{i=1}^{|\pi|} \kappa(\mathbf{x}_{\pi_1(i)},\mathbf{y}_{\pi_2(i)}).$$

Shin & Kuboyama prove that a mapping kernel is p.d. for all p.d. kernels κ if and only if $\mathcal{M}(\mathbf{x},\mathbf{y})$ is a transitive set, that is $\forall\,(\mathbf{x},\mathbf{y},\mathbf{z}) \in \mathcal{X}^3$,

$$(x,y) \in \mathcal{M}(\mathbf{x},\mathbf{y}), (y,z) \in \mathcal{M}(\mathbf{x},\mathbf{z}) \Rightarrow (x,z) \in \mathcal{M}(\mathbf{x},\mathbf{z}).$$

Lemma 2. *The mapping set $\mathcal{M}_{\text{GA}}$ is not transitive.*

Proof. Consider two time series $\mathbf{x}$ and $\mathbf{z}$ of length 2 and $\mathbf{y}$ of length 3. It is clear that

$$\left(\begin{smallmatrix}1,1,2\\1,2,3\end{smallmatrix}\right) \in \mathcal{A}(2,3), \left(\begin{smallmatrix}1,2,3\\1,1,2\end{smallmatrix}\right) \in \mathcal{A}(3,2),\ \text{yet} \left(\begin{smallmatrix}1,1,2\\1,1,2\end{smallmatrix}\right) \notin \mathcal{A}(2,2),$$

since the latter alignment does not satisfy the increasing property defined in Equation (1). ■

The mapping kernel theorem thus proves that GA kernels k_{GA} cannot be p.d. for all local kernels κ. This negative result highlights the fact that additional conditions on κ are needed to obtain p.d. kernels k_{GA}. The condition that $\kappa/(1+\kappa)$ is p.d., which we coin down geometric divisibility in Definition 1, can be interpreted as a sufficient condition in this context.

Geometric Divisibility Consider the map τ from nonnegative scalars to $[0,1[$, $\tau : x \mapsto \frac{x}{1+x}$ and note that $\tau^{-1}(x) = \frac{x}{1-x}$. By abuse of notation, for any nonnegative valued function f we write τf for the composition $\tau \circ f$ of τ and f and for any $[0,1[$ valued function g we write $\tau^{-1}g$ for $\tau^{-1} \circ g$.

Definition 1 (Geometric Divisibility)**.** *Let f be a nonnegative valued function on $\mathcal{X} \times \mathcal{X}$. f is said to be geometrically divisible if τf is positive definite.*

For a geometrically divisible (g.d.) function f, $|\tau f| < 1$ and thus f can be written as the geometric series $f = \sum_{i=1}^{\infty} (\tau f)^i$, giving the definition its name. Any g.d. function f is by definition a sum of p.d. kernels, and is thus *necessarily* positive definite.

Geometric divisibility was mentioned in (Cuturi et al., 2007) as a "mild" condition on the local kernel κ to ensure that the resulting GA kernel k_{GA} is positive definite. Contrary to what is suggested by Cuturi et al., the Gaussian kernel is *not* geometrically divisible. Indeed, this can be exhibited numerically by showing that a suitable Gram matrix does not have that property[1]. We conjecture that other related kernels, such as the Laplace kernels[2] are not geometrically divisible either. The next result shows that infinite divisibility is preserved when applying τ^{-1} to a given kernel in order to obtain a geometrically divisible kernel.

Lemma 3. *For any infinitely divisible kernel κ such that $0 < \kappa < 1$, $\tau^{-1}\kappa$ is g.d. and infinitely divisible.*

Proof. Note that

$$\begin{aligned} -\log(\tau^{-1}\kappa) &= -\log\kappa + \log(1-\kappa) \\ &= -\log\kappa + \int_{t=0}^{1} \frac{-\kappa}{1-t\kappa}dt. \end{aligned} \quad (6)$$

Since for each $t \le 1$, $|t\kappa| < 1$ the identity

$$\frac{\kappa}{1-t\kappa} = \sum_{i=0}^{\infty} t^i \kappa^{i+1},$$

holds, the right hand-side of Equation (6) is the sum of a n.d. kernel (by infinite divisibility of κ) and minus a sum of p.d. kernels. $-\log(\tau^{-1}\kappa)$ is thus n.d., which proves $\tau^{-1}\kappa$'s infinite divisibility. ■

Considering the Gaussian kernel κ_σ first, the kernel $\tau^{-1}(\kappa_\sigma/2) = \kappa_\sigma/(2-\kappa_\sigma)$ is thus both infinitely *and* geometrically divisible. Hence its logarithm

$$\phi_\sigma(x,y) \overset{\text{def}}{=} \tfrac{1}{2\sigma^2}\|x-y\|^2 + \log\left(2 - e^{-\frac{\|x-y\|^2}{2\sigma^2}}\right), \quad (7)$$

is a n.d. kernel that can be scaled by a factor λ to define a local kernel $e^{-\lambda\phi_\sigma}$ that can be easily tuned to correct any diagonal dominance as suggested in Section 3.1. Note that a similar construction can be carried out with Laplace kernels.

[1] *e.g.*, matlab code to test this:

```
rand('state',0); o=ones(10,1); A=rand(2,10); B=sum(A.^2);
K=exp(2*A'*A-kron(o,B)-kron(o',B')); eig(K./(1+K))
```

[2] $\exp(-\lambda\|x-y\|^a), 0 < a < 2, \lambda > 0$

4. Triangular Global Alignment Kernels

Itakura (1975) and Sakoe & Chiba (1978) proposed to speed up the computation of DTW distances through additional constraints on alignments. Similar ideas can be applied to GA kernels as shown in this section.

4.1. Constrained Alignments for DTW

Exact DTW distances are expensive to compute for time series of dimension d since they require $O(dnm)$ elementary operations at each evaluation. To cope with this cost when using DTW within nearest neighbor methods, inexpensive lower bounds on $\mathrm{DTW}(\mathbf{x},\mathbf{y})$ can be used to screen and discard all time series $\mathbf{y}$ in a database that have a poor match with a time series of interest $\mathbf{x}$ (Lemire, 2009). Unfortunately such approaches are ineffective with kernel methods since the latter require to compute all pairwise similarities of a database of time series in the training phase. The DTW algorithm can be sped up by only considering a small subset of all alignments as detailed by Rabiner & Juang (1993, §4.7). These constraints can be formulated with weights $\gamma_{i,j}$ that modify the cost function of Equation (3) into

$$D^{\gamma}_{\mathbf{x},\mathbf{y}}(\pi) \overset{\text{def}}{=} \sum_{i=1}^{|\pi|} \gamma_{\pi_1(i),\pi_2(i)}\, \varphi(x_{\pi_1(i)}, y_{\pi_2(i)}), \qquad (8)$$

For instance, Sakoe & Chiba (1978) define the band

$$\gamma_{i,j} = \begin{cases} 1, & \text{if } |i-j| < T \\ \infty, & \text{if } |i-j| \geq T \end{cases}, \qquad (9)$$

which ensures that only alignments π that are close to the diagonal are considered, namely such that $\pi_1(i)$ and $\pi_2(i)$ remain close as i grows. This increase in speed comes at the cost of an approximation, since the resulting optimal path may turn out to be suboptimal.

4.2. GA Kernels with Position Kernels

Weights introduced in Equation (8) can be naturally extended to GA kernels. Indeed, any p.d. kernel ω on $\mathbb{N}$, rather than a weight $\gamma_{i,j}$, can be paired with a kernel κ on $\mathcal{X}$ to form GA kernels:

Remark 1. *Let κ be a kernel on $\mathcal{X}$ and ω a kernel on $\mathbb{N}$. Then using $\tau^{-1}(\omega \otimes \kappa)$ as a local kernel, the kernel k_{GA} is p.d. on time series of points taken in $\mathbb{N} \times \mathcal{X}$.*

With this simple argument, points enumerated in a time series $\mathbf{x}$ can be described with their position as (i, x_i) for $1 \leq i \leq n$. The kernel ω modulates the similarity of two points (x_i, y_j) by taking into account their respective location i and j while κ quantifies the similarity of x_i and y_j. We consider in this paper kernels $\omega(i,j)$ that only depend on $|i-j|$, namely radial basis kernels $\omega(i,j) = \psi(|i-j|)$ where ψ is a real-valued function on $\mathbb{N}$. Such kernels on integers are also known as Toeplitz kernels. In the context of Global Alignment kernels, Toeplitz kernels are more appealing if they are compactly supported as discussed in the next section.

4.3. Compactly Supported Toeplitz Kernels

A Toeplitz kernel ω is compactly supported of order $T \in \mathbb{N}$ if for $q \geq T, \psi(q) = 0$ and $\psi(T-1) \neq 0$. Using such a kernel within GA kernels has obvious advantages:

Theorem 1. *Let κ be a kernel on $\mathcal{X} \times \mathcal{X}$ and ω a compactly supported Toeplitz kernel of order T. Then using $\tau^{-1}(\omega\kappa)$ as a local kernel, $k_{GA}(\mathbf{x},\mathbf{y})$ can be computed with $O(T\min(n,m))$ operations. Furthermore, $k_{GA}(\mathbf{x},\mathbf{y})$ is null when $|n-m| > T$.*

Proof. Since ω has compact support, all elements of the $n \times m$ Gram matrix that are off the diagonal by C are null. The recursive iteration $M_{i+1,j+1} = K_{i,j}\,(M_{i,j} + M_{i,j+1} + M_{i+1,j})$ only has to be applied on the portion of the Gram matrix that is non-zero, which entails up to $(2T-1)\min(n,m) - T(T-1)/2$ updates, as illustrated in Figure 2. ■

The Sakoe & Chiba band defined as $e^{-\gamma_{i,j}} = \delta_{|i-j|<T}$ is symmetric and Toeplitz but not p.d., and cannot be used as a local kernel. Compactly supported kernels were studied by Gneiting (2002) who highlights in particular the triangular kernel for integers,

$$\omega(i,j) = \left(1 - \frac{|i-j|}{T}\right)_+, \qquad (10)$$

which is known to be p.d. in $\mathbb{R}$ but not in higher dimensions. Genton (2002, Figure 1) provides an interesting discussion on its characteristics. We consider this kernel in the following and define Triangular GA (TGA) kernels as GA kernels obtained when pairing the triangular kernel of Equation (10) with any local kernel κ following the construction given in Remark 1.

5. Experiments

5.1. Databases

In addition to different classification tasks on datasets taken from the UCI Machine Learning repository (Frank & Asuncion, 2010) we have compiled a database of freeway traffic, the PEMS database, which we introduce below. All datasets describe multivariate time series of dimension d and variable length n associated with one of many possible classes.

Database	d	n range, **med**(n)	classes	# points
Japanese Vowels	12	7-29, 15	9	640
Libras	2	45	15	945
Handwritten Characters	3	60-182, 122	20	2858
AUSLAN	22	45-136, 55	95	2465
PEMS	963	144	7	440

Table 1. Characteristics of the different databases considered in the benchmark test

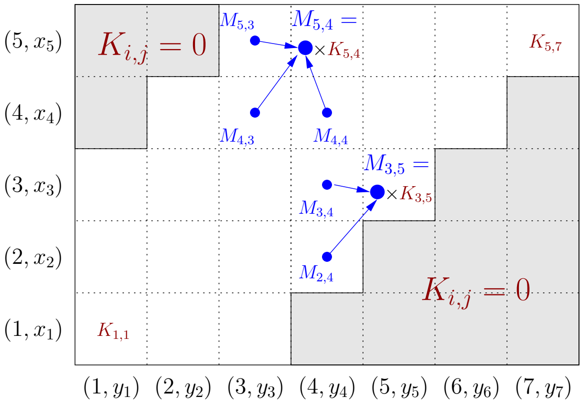

Figure 2. Illustration of the GA kernel recursion when using a Compactly-supported Toeplitz kernel of order $T = 3$. The recursion $M_{i+1,j+1} = K_{i,j}\,(M_{i,j} + M_{i,j+1} + M_{i+1,j})$ implies that the computation of $M_{i,j}$ can be bypassed for indexes such that $K_{i,j} = 0$ where $K_{i,j}$ stands for $\tau^{-1}(\omega \otimes \kappa)\left((i, x_i), (j, y_j)\right)$.

PEMS database We have downloaded 15 months worth of daily data from the California Department of Transportation PEMS website[3]. The data describes the occupancy rate, between 0 and 1, of different car lanes of San Francisco bay area freeways. The measurements cover the period from Jan. 1st 2008 to Mar. 30th 2009 and are sampled every 10 minutes. We consider each day in this database as a single time series of dimension 963 (the number of sensors which functioned consistently throughout the studied period) and length $6 \times 24 = 144$. The task is to classify each day as the correct day of the week, from Monday to Sunday, e.g. label it with an integer in $\{1, 2, 3, 4, 5, 6, 7\}$. We remove public holidays from the dataset, as well as two days with anomalies (March 8th 2009 and March 9th 2008) where all sensors were muted between 2:00 and 3:00 AM. This results in a database of 440 time series.

5.2. Kernels, Parameters and Methodology

We consider the following kernels in our benchmark:

DTW kernel Following Haasdonk & Bahlmann (2004) the DTW distance introduced in Equation (2) can be used as a pseudo n.d. kernel to define the pseudo p.d. kernel $k_{\text{DTW}} = e^{-\frac{1}{t}\text{DTW}}$.

DTW kernel with Sakoe-Chiba Constraints Sakoe & Chiba's constrained DTW distance,

$$\text{DTW}_{\text{SC}}(\mathbf{x}, \mathbf{y}) = \min_{\pi \in \mathcal{A}(n,m)} D^{\gamma}_{\mathbf{x},\mathbf{y}}(\pi),$$

where the weights $\gamma_{i,j}$ are defined Equation (9) can also yield a pseudo p.d. kernel, $k_{\text{SC}} = e^{-\frac{1}{t}\text{DTW}_{\text{SC}}}$. **DTAK kernel** Shimodaira et al. (2002) consider a variant of the DTW to define the pseudo p.d. kernel

$$k_{\text{DTAK}}(\mathbf{x}, \mathbf{y}) = \max_{\pi \in \mathcal{A}(n,m)} \sum_{i=1}^{|\pi|} \kappa_\sigma(x_{\pi_1(i)}, y_{\pi_2(i)}).$$

To be consistent with the definition k_{DTW} and k_{SC} we consider its exponentially scaled expression $(k_{\text{DTAK}})^{\frac{1}{t}}$. **GA Kernel** We use the GA kernel k_{GA} seeded with the local kernel $\kappa = e^{-\phi_\sigma}$ where the negative definite kernel ϕ_σ is given in Equation (7).

TGA Kernel We consider the Gaussian kernel paired with the triangular kernel to define the local kernel

$$\tau^{-1}(\omega \otimes \tfrac{1}{2}\kappa_\sigma)(i, x; j, y) = \frac{\omega(i,j)\kappa_\sigma(x,y)}{2 - \omega(i,j)\kappa_\sigma(x,y)}.$$

The kernels considered in these experiments are renormalized before using `libsvm`'s implementation of support vector machines. We consider a doubly nested CV scheme to obtain estimates of classification error rates. Namely, each dataset is first randomly split into 3 balanced folds. For each kernel, the parameters taken within an adaptive grid described in Table 2 that have the lowest mean classification-error on the training fold (using 3 folds 2 repeats cross-validations and selecting the C constant of `libsvm` in $\{1, 10^2, 10^4\}$) are used to test the accuracy of the kernel on the remaining two folds of data with a SVM trained on the training fold. When the Gram matrix of the training fold is not positive definite, which happens only with and relatively often for $k_{\text{DTW}}, k_{\text{SC}}$ and k_{DTAK}, we regularize it with a ridge to ensure it becomes positive definite. Such a split is repeated 3 times, yielding 3×3 error rates estimates for each database/kernel pair. We report in Figure 3 the mean and the standard deviation of each group of 9 error estimates.

[3] `http://pems.dot.ca.gov`

Kernel	Parameters	Parameter Line/Grid
k_{DTW}	t	$t \in \{0.2, 0.5, 1, 2, 5\} \cdot \mathbf{med}(\mathrm{DTW}(\mathbf{x}, \mathbf{x}))$
k_{SC}	t, T	$t \in \{0.2, 0.5, 1, 2, 5\} \cdot \mathbf{med}(\mathrm{DTW}_{\mathrm{SC}}(\mathbf{x}, \mathbf{y})), \quad T \in \{0.25, 0.5\} \cdot \mathbf{med}(\lvert\mathbf{x}\rvert)$
k_{DTAK}	t, σ	$t \in \{0.2, 0.5, 1, 2, 5\} \cdot \mathbf{med}(-\log k_{\mathrm{DTAK}}(\mathbf{x}, \mathbf{y})), \quad \sigma \in \{0.2, 0.5, 1, 2\} \cdot \mathbf{med}(\Vert x - y\Vert)$
k_{GA}	σ	$\sigma \in \{0.2, 0.5, 1, 2, 5\} \cdot \mathbf{med}(\Vert x - y\Vert) \cdot \sqrt{\mathbf{med}(\lvert\mathbf{x}\rvert)}$
k_{TGA}	σ, T	$\sigma \in \{0.2, 0.5, 1, 2, 5\} \cdot \mathbf{med}(\Vert x - y\Vert) \cdot \sqrt{\mathbf{med}(\lvert\mathbf{x}\rvert)}, \quad T \in \{0.25, 0.5\} \cdot \mathbf{med}(\lvert\mathbf{x}\rvert)$

Table 2. Parameter grid for all kernels. $\mathbf{med}(f)$ stands for the empirical median of f computed on training sets. When the arguments of f are time series $\mathbf{x}$ or $\mathbf{y}$ the median is computed by sampling over the entire training set in both variables if necessary. When the arguments of f are vectors x or y, the vectors are sampled randomly within time series sampled randomly in the training set. The $\cdot$ multiplication is elementwise, *e.g.* $\{1, 2, 3\} \cdot \sigma = \{\sigma, 2\sigma, 3\sigma\}$.

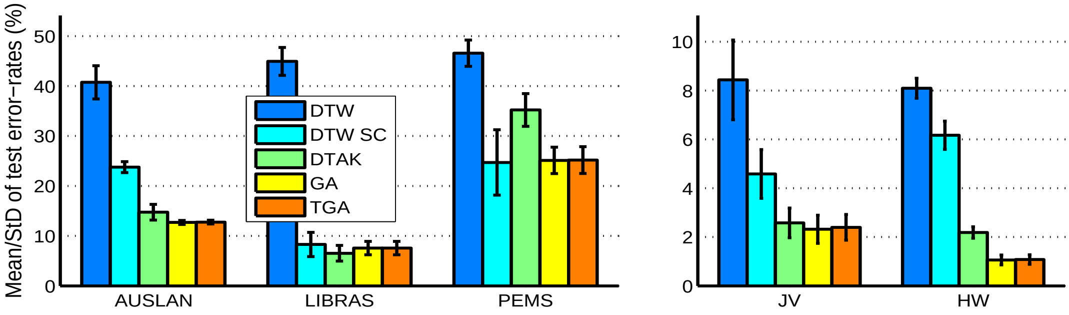

Figure 3. Mean and standard deviations of classification error rates on test folds, using a 3 folds 3 repeats cross validation procedure for each kernel/database pair. Parameters were selected independently for each test iteration by applying an adaptive grid search (Table 2) within each training fold, using a 3 folds 2 repeats cross validation.

5.3. Results and Discussion

Source codes for these experiments are available on the author's webpage. These experiments show that GA and TGA kernels behave similarly and compare favorably to all other DTW kernels considered here, as pictured in Figure 3. The poor performance of the DTW kernel suggests that it should be avoided in most applications, to consider instead DTAK or (T)GA kernels. We report in Figure 4 average error rates along with the average runtime required to compute a single kernel evaluation of TGA kernels as a function of T. Note that for low T, the resulting kernel matrix is sparse and non-zero only for time series of very similar length. When $T = 1$ the TGA kernel becomes de facto a Gaussian kernel (with distance ϕ_σ) that can only compare time series of equal length. As T increases the TGA kernel converges to the GA kernel, with substantial or negligible improvements in performance, depending on the dataset. Yet, for most databases T does not need to be increased significantly to reach a performance that is comparable to that of GA kernels, as illustrated in Figure 3, where T never exceeds half the median length of time series in each database as specified in Table 2. This leads however to faster computations, as detailed in Figure 4.

Acknowledgements We thank anonymous reviewers for helpful comments. This work was supported by JSPS grant 23700172.

References

Bahlmann, C., Haasdonk, B., and Burkhardt, H. Online handwriting recognition with support vector machines-a kernel approach. In *Proc. of 8th Intern. Workshop on Frontiers in Handwriting Recognition*, pp. 49–54, 2002.

Banderier, C. and Schwer, S. Why Delannoy numbers? *Journal of Stat. Plann. and Inf.*, 135(1):40–54, 2005.

Berg, C., Christensen, J.P.R., and Ressel, P. *Harmonic Analysis on Semigroups*. Number 100 in Graduate Texts in Mathematics. Springer Verlag, 1984.

Cuturi, M., Vert, J.-P., Birkenes, Ø., and Matsui, T. A kernel for time series based on global alignments. In *Proceedings of ICASSP*, volume II, pp. 413 – 416, 2007.

de Vries, G. and van Someren, M. Clustering Vessel Trajectories with Alignment Kernels under Trajectory Compression. *Machine Learning and Knowledge Discovery in Databases*, pp. 296–311, 2010.

Frank, A. and Asuncion, A. UCI machine learning repository, http://archive.ics.uci.edu/ml, 2010.

Genton, M.G. Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2:312, 2002.

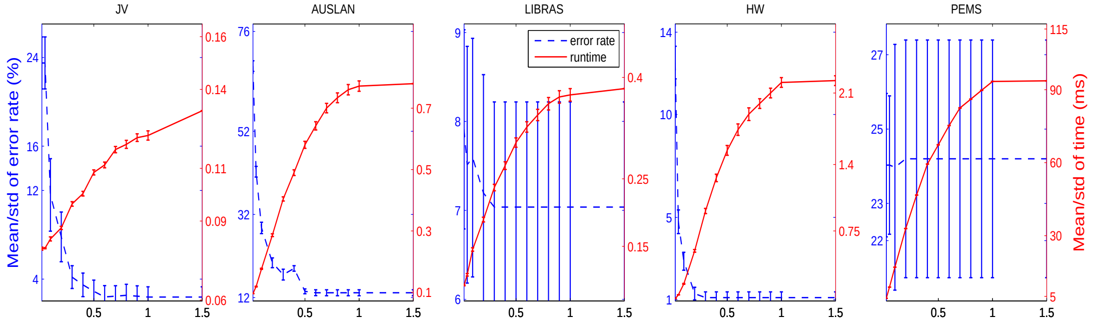

Figure 4. Performance and speed (average runtime of one kernel evaluation in milliseconds) of the TGA kernel as a function of T, shown here as a fraction between 0 and 1 of the median length of all time series in a given dataset. Runtimes for $T = 1.5\,\mathbf{med}|\mathbf{x}|$ are also given for reference.

Gneiting, T. Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83(2):493–508, 2002.

Gudmundsson, S., Runarsson, T.P., and Sigurdsson, S. Support vector machines and dynamic time warping for time series. In *IEEE IJCNN 2008*, pp. 2772 –2776.

Haasdonk, B. and Bahlmann, C. Learning with distance substitution kernels. *Pattern Recognition, Proc. of the 26th DAGM Symposium*, pp. 220–227, 2004.

Harchaoui, Z. and Bach, F. Image classification with segmentation graph kernels. In *CVPR*, 2007.

Haussler, D. Convolution kernels on discrete structures. Technical report, UCSC, 1999. USCS-CRL-99-10.

Hayashi, A., Mizuhara, Y., and Suematsu, N. Embedding time series data for classification. *Machine Learning and Data Mining in Pattern Recognition*, pp. 356–365, 2005.

Hofmann, T., Scholkopf, B., and Smola, A.J. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171, 2008.

Itakura, F. Minimum prediction residual principle applied to speech recognition. *IEEE Trans. on Acoustics, Speech and Sig. Proc.*, 23(1):67 – 72, February 1975.

Jebara, T., Kondor, R., and Howard, A. Probability product kernels. *Journal of Machine Learning Research*, 5: 819–844, 2004.

Joder, C., Essid, S., and Richard, G. Temporal integration for audio classification with application to musical instrument classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):174–186, 2009.

Lemire, D. Faster retrieval with a two-pass dynamic-time-warping lower bound. *Pattern Recognition*, 42(9):2169–2180, 2009.

Müller, M. *Information Retrieval for Music and Motion*. Springer-Verlag, 2007.

Rabiner, L. and Juang, B.H. *Fundamentals of speech recognition*, volume 103. Prentice hall, 1993.

Ricci, E., Tobia, F., and Zen, G. Learning Pedestrian Trajectories with Kernels. In *Proc. of the 20th International Conference in Pattern Recognition*, 2010.

Sakoe, H. and Chiba, S. A similarity evaluation of speech patterns by dynamic programming. *Nat. Meeting of Institute of Electronic Communications Engineers of Japan, p. 136*, 1970.

Sakoe, H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech, and Sig. Proc.*, 26:43–49, 1978.

Shimodaira, H., Noma, K.-I., Nakai, M., and Sagayama, S. Dynamic time-alignment kernel in support vector machine. In Dietterich, T. G., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

Shin, K. and Kuboyama, T. A generalization of Haussler's convolution kernel: mapping kernel. In *Proceedings of the 25th international conference on Machine learning*, pp. 944–951, 2008.

Sonnenburg, S., Rieck, K., and Rätsch, G. Large scale learning with string kernels. In *Large Scale Kernel Machines*, pp. 73–103. MIT Press, 2007.

Sulanke, R.A. Objects counted by the central Delannoy numbers. *J. Integer Seq*, 6(1), 2003.

Vishwanathan, S.V.N., Smola, A.J., and Vidal, R. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *International Journal of Computer Vision*, 73(1):95–119, 2007.

Vishwanathan, S.V.N., Borgwardt, K.M., Kondor, I.R., and Schraudolph, N.N. Graph kernels. *Journal of Machine Learning Research*, 9:1–37, 2008.

Zhou, F., De la Torre, F., and Cohn, J.F. Unsupervised discovery of facial events. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2574–2581, 2010.