Times series averaging from a probabilistic interpretation of time-elastic kernel

Pierre-François Marteau

Times series averaging from a probabilistic interpretation of time-elastic kernel

Pierre-Francois Marteau, *Member, IEEE,*
E-mail: see http://people.irisa.fr/Pierre-Francois.Marteau/

Abstract—In the light of regularized dynamic time warping kernels, this paper re-considers the concept of time elastic centroid (TEC) for a set of time series. From this perspective, we show that TEC can be readily addressed as a preimage problem. However, this non-convex problem is ill-posed, and obtaining a sub-optimal solution may involve heavy computational costs, especially for long time series. We then derive two new algorithms based on a probabilistic interpretation of kernel alignment matrices that expresses the result in terms of probabilistic distributions over sets of alignment paths. The first algorithm is an agglomerative iterative heuristic procedure inspired from a state-of-the-art DTW barycentre averaging algorithm. The second proposed algorithm uses a progressive agglomerative heuristic method to perform classical averaging of the aligned samples but also averages the times of occurrence of the aligned samples. By comparing classification accuracies for 45 time series datasets obtained by first nearest centroid/medoid classifiers we show that: i) centroid-based approaches significantly outperform medoid-based approaches, ii) for the considered datasets, the second algorithm which combines averaging in the sample space and along the time axes, emerges as the most significantly robust heuristic model for time-elastic averaging with a promising noise reduction capability.

Index Terms—Time series averaging Time elastic kernel Dynamic Time Warping Time series clustering and classification.

1 Introduction

Since Maurice Fréchet's pioneering work [1] in the early 1900s, *time-elastic* matching of time series or symbolic sequences has attracted much attention from the scientific community in numerous fields such as information indexing and retrieval, pattern analysis, extraction and recognition, data mining, etc. This approach has impacted a very wide spectrum of applications relating to a multitude of socio-economic issues such as the environment, industry, health, energy, defense and so on.

Among other time elastic measures, Dynamic Time Warping (DTW) was widely popularized during the 1970s with the advent of speech recognition systems [2], [3] and numerous variants that have since been proposed to match time series with a certain degree of time distortion tolerance.

The main issue addressed here is time series or shape averaging in the context of a time elastic distance. This is a long-standing issue that is currently becoming increasingly prevalent; it is relevant for summarizing subsets of time series, defining significant prototypes, identifying outliers, performing data mining tasks (mainly exploratory data analysis such as clustering) and speeding up classification, as well as regression or data analysis processes in a big data context.

In this paper, we specifically tackle the question of averaging subsets of time series, not from considering the DTW measure itself as has been already largely explored, but from the perspective of the so-called regularized DTW kernel (KDTW) that ensures positive definiteness. From this new perspective, the estimation of a time series average or centroid can be readily addressed as a preimage (inverse) problem. However, this approach has some theoretical and practical limitation that are discussed in the following sections. A more promising direct approach is developed here, which is based on a probabilistic interpretation of kernel alignment matrices, allowing a precise definition of the average of a pair of time series from the expected value of local alignments of samples. The tests carried out so far demonstrate the robustness and the efficiency of this approach comparison to the state-of-the art approach.

The structure of this paper is as follows: after an introduction, the second section summarizes the most relevant related studies on time series averaging as well as DTW kernelization. In the third section, we show how the estimation of a time-elastic centroid can be addressed as a preimage problem in the context of the DTW regularized kernel (KDTW). In the fourth section, we derive a probabilistic interpretation from the kernel alignment matrices evaluated on a pair of time series. In the fifth section, we define the average of a pair of time series, and based on this pairwise averaging procedure, we propose two sub-optimal algorithms designed for the averaging of any subset of time series.

2 Related works

Time series averaging in the context of (multiple) time elastic distance alignments has been mainly addressed in the scope of the Dynamic Time Warping (DTW) measure [2], [3]. Although other time elastic distance

• *P.-F. Marteau is with UMR CNRS IRISA, Université de Bretagne Sud, F-56000 Vannes, France.*

measures such as the Edit Distance With Real Penalty (ERP) [4] or the Time Warp Edit Distance (TWED) [5] could be considered instead, without loss of generality, we remain focused throughout this paper on DTW and its kernelization.

2.1 DTW and time elastic centroid of a pair of time series

A classical formulation of DTW can be given as follows. If d is a fixed positive integer, we define a time series of length T as a multidimensional sequence $v = v(i)$, such that, $\forall i \in \{1, .., T\}$, $v(i) \in \mathbb{R}^d$.

Definition 2.1: If u and v are two time series with respective lengths T_1 and T_2, an *alignment path* $\pi = (\pi_k)$ of length $p = |\pi|$ between u and u is represented by a sequence

$$\pi : \{1, \ldots, p\} \rightarrow \{1, \ldots, T_1\} \times \{1, \ldots, T_2\}$$

such that $\pi_1 = (1, 1)$, $\pi_p = (T_1, T_2)$, and (using the notation $\pi_k = (i_k, j_k)$, for all $k \in \{1, \ldots, p-1\}$, $\pi_{k+1} = (i_{k+1}, j_{k+1}) \in \{(i_k + 1, j_k), (i_k, j_k + 1), (i_k + 1, j_k + 1)\}$.

We define $\forall k\ \pi_k(1) = i_k$ and $\pi_k(2) = j_k$, as the index access functions at step k of the mapped elements in the pair of aligned time series.

In other words, a warping path defines a way to travel along both time series simultaneously from beginning to end; it cannot skip a point, but it can advance one time step along one series without advancing along the other, thereby justifying the term *time-warping*.

If δ is a distance on $\mathbb{R}^d$, the global *cost* of a warping path π is the sum of distances (or squared distances or local costs) between pairwise elements of the two time series along π, i.e.:

$$\text{cost}(\pi) = \sum_{(i_k, j_k) \in \pi} \delta(v_{i_k}, w_{j_k})$$

A common choice of distance on $\mathbb{R}^d$ is the one generated by the L^2 norm:

$$\delta(x, y) = \|x - y\|_2^2 = \sum_{l=1}^{d} (x_l - y_l)^2.$$

Definition 2.2: For a finite time series, any warping path has a finite length, and thus the number of existing warping paths is finite. Hence, there exists at least one path π^* whose cost is minimal, so we can define $\text{DTW}(u, v)$ as the minimal cost taken over all existing warping paths. Hence

$$\text{DTW}(u, v) = \min_{\pi} \text{cost}(\pi(u, v)) = \text{cost}(\pi^*(u, v)). \quad (1)$$

Definition 2.3: From the DTW measure, it is straightforward to define the time elastic centroid $c(u, v)$ of a pair of time series u and v as the time series (c_k) whose elements are $c_k = \textit{Centroid}(u(\pi_k^*(1)), v(\pi_k^*(2)))$, $\forall k \in 1, \cdots, |\pi^*|$, where *Centroid* corresponds to the usual definition in Euclidean space.

2.2 Time elastic centroid of a set of time series

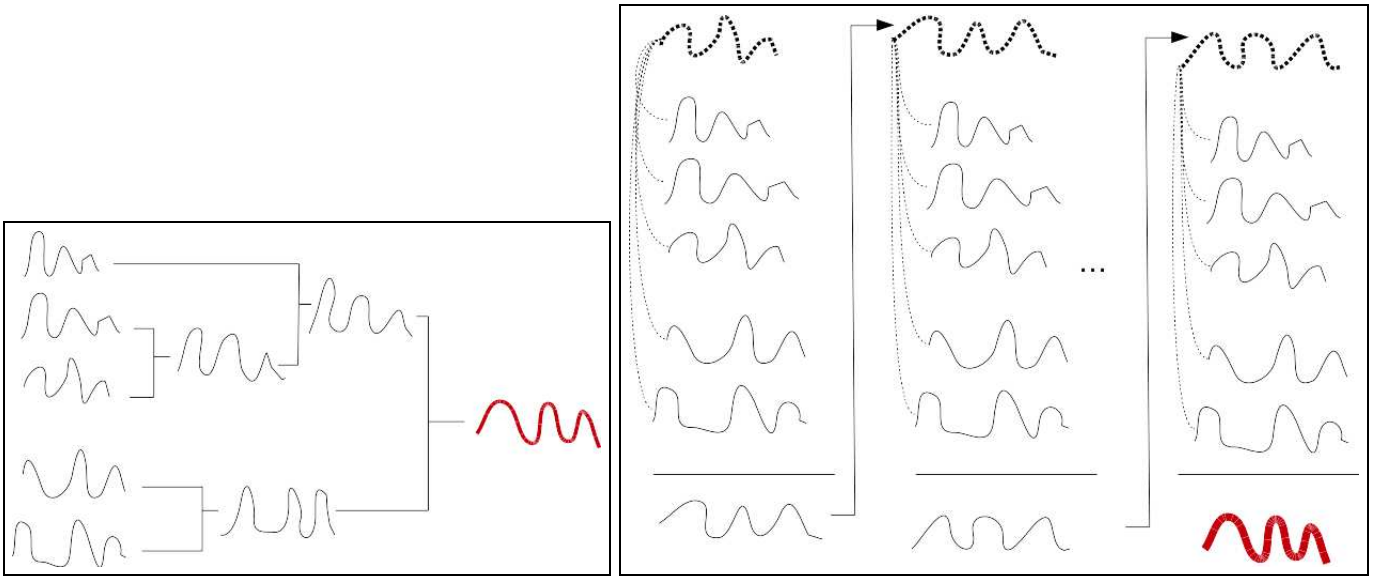
A single alignment path is required to calculate the time elastic centroid of a pair of time series (Def. 2.3). However, multiple path alignments need to be considered to evaluate the centroid of a larger set of time series. Multiple alignments have been widely studied in bioinformatics [6], and it has been shown that the computational complexity of determining the optimal alignment of a set of sequences under the sum of all pairs (SP) score scheme is a NP-complete problem [7] [8]. The time and space complexity of this problem is $O(L^k)$, where k is the number of sequences in the set and L is the length of the sequences when using dynamic programming to search for an optimal solution [9]. This latter result applies to the estimation of the time elastic centroid of a set of k time series with respect to the DTW measure. Since the search for an optimal solution becomes rapidly intractable with increasing k, sub-optimal heuristic solutions have been subsequently proposed, most of them falling into one of the following three categories.

2.2.1 Progressive heuristics

Progressive heuristic methods estimate the time elastic centroid of a set of k time series by combining pairwise centroids (Def. 2.3). This kind of approach constructs a binary tree whose leaves correspond to the time series of the data set, and whose nodes correspond to the calculation of a local pairwise centroid, such that, when the tree is complete, the root is associated with the estimated data set centroid. The proposed strategies differ in the way the tree is constructed. One popular approach consists of providing a random order for the leaves, and then constructing the binary tree up to the root using this ordering [10]. Another approach involves constructing a dendrogram (a hierarchical ascendant clustering) from the data set and then using this dendrogram to calculate pairwise centroids starting with the closest pairs of time series and progressively aggregating series that are farther away [11] as illustrated on the left of Fig. 1. Note that these heuristic methods are entirely based on the calculation of a pairwise centroid, so they do not explicitly require the evaluation of a DTW centroid for more than two time series. Their degree of complexity varies linearly with the number of time series in the data set.

2.2.2 Iterative heuristics

Iterative heuristics are based on an iterated three-step process. For a given temporary centroid candidate, the first step consists of calculating the inertia, i.e. the sum of the DTW distances between the temporary centroid



(a) Progressive agglomeration (b) Iterative agglomeraation with refinement

Fig. 1. Progressive hierarchical with similar first agglomeration (left) v.s. iterative agglomeration (right) strategies. Final centroid approximations are presented in red bold color. Temporary estimations are presented using a bold dotted black line

and each time series in the data set. The second step evaluates the best pairwise alignment with the temporary centroid for each time series $u_j(i)$ in the data set ($j \in \{1 \cdots n\}$). A new time series $\tilde{u}_j(i)$ is thus constructed that contains all the samples of time series $u_j(i)$, but with time being stretched or compressed according to the best alignment path. The third step consists of producing a new temporary centroid candidate $c(i)$ from the set $\{\tilde{u}_j(i)\}$ by successively averaging (in the sense of the Euclidean centroid), the samples at every timestamp i of the $\tilde{u}_j(i)$ time series. Basically, $c(i) = \sum_{j=1..n_i} \tilde{u}_j(i).\mathbb{1}(i,j) / \sum_{j=1..n_i} \mathbb{1}(i,j))$, where $\mathbb{1}(i,j)$ is an indicator function equal to 1 if time series $\tilde{u}_j$ is defined for timestamp i, but which is otherwise 0.

Thus, the new centroid candidate replaces the previous one and the process is iterated until the inertia is no longer reduced or the maximum number of iterations is reached. Generally, the first temporary centroid candidate is taken as the DTW medoid of the considered data set. This process is illustrated on the right of Fig. 1. The three steps of this heuristic method were first proposed in [12]. The iterative aspect of this heuristic approach was initially introduced by [13] and refined by [14]. Note that, in contrast to the progressive method, this kind of approach needs to evaluate, at each iteration, all the alignments with the current centroid candidate. The complexity of the iterative approach is higher than the progressive approach, the extra computational cost being linear with the number of iterations. More sophisticated approaches have been proposed to escape some local minima. [15] have evaluated a genetic algorithm for managing a population of centroid candidates, thus improving with some success the straightforward iterative heuristic methods.

2.2.3 *Optimization approaches*

Given the entire set of time series $\mathbb{S}$ and a subset of n time series $S = \{u_j\}_{j=1\cdots n} \subseteq \mathbb{S}$, optimization approaches attempt to estimate the centroid of S from the definition of an optimization problem, which is generally expressed by Eq. 2 given below:

$$c = \underset{s \in \mathcal{S}}{\operatorname{argmin}} \sum_{j=1}^{n} \mathrm{DTW}(s, u_j) \tag{2}$$

To our knowledge, the first attempt to use this kind of direct approach for the estimation of time elastic centroid estimation was recently described in [16].

These authors (op.cit.) derived a solution of their original non-convex constrained optimization problem, by integrating a temporal weighting of local sample alignments to highlight the temporal region of interest in a time series data set, thus penalizing the other temporal regions. Two time elastic measures were specifically addressed: i) a dynamic time warping measure between a time series and a weighted time series (representing the centroid estimate) and ii) an (indefinite) kernel DTW called DTAK [17]. Their results are very promising: although the number of parameters to optimize is linear with the size and the dimensionality of the time series, the two steps gradient-based optimization process they derived is very computationally efficient and shown to outperform the state of the art approaches on some challenging scalar and multivariate data sets. However, as numerous local *optima* exist in practice, the method is not guaranteed to converge toward the best possible centroid, which is anyway the case in all other approaches.

2.3 Discussion and motivation

According to the state of the art in time elastic centroid estimation, an exact centroid, if it exists, can be calculated by solving a NP-complete problem whose complexity is exponential with the number of time series to be averaged. Heuristic methods with increasing time complexity have been proposed since the early 2000s. Simple pairwise progressive aggregation is a less complex approach, but which suffers from its dependence on initial conditions. Iterative aggregation is reputed to be more efficient, but entails a higher computational cost. It could be combined with ensemble methods or soft optimization such as genetic algorithms. The non-convex optimization approach has the merit of directly addressing the mathematical formulation of the centroid problem in a time elastic distance context. This approach nevertheless involves a higher complexity and must deal with a relatively large set of parameters to be optimized (the weights and the sample of the centroid). Its scalability could be questioned, specifically for high dimensional multivariate time series.

It should also be mentioned that some criticisms of these heuristic methods have been made in [18]. Among other drawbacks, the fact that DTW is not a metric (the triangle inequality is not satisfied) could explain the occurrence of unwanted behaviour such as centroid drift outside the time series cluster to be averaged. We should also be borne in mind that keeping a single best alignment (even though several may exist, without mentioning the *good* ones) can increase the dependence of the solution on the initial conditions. It may also increase the aggregating order of the time series proposed by the chosen method, or potentially enhance the convergence rate.

In this study, we do not directly address the issue of time elastic centroid estimation from the DTW perspective, but rather from the point of view of the regularized dynamic time warping kernel (KDTW) [19]. This perspective allows us to consider centroid estimation as a preimage problem, which is in itself another optimization perspective. More importantly, the KDTW alignment matrices can be used to derive a probabilistic interpretation of the pairwise alignment of time series. This leads us to propose a robust interpolation scheme jointly along the time axis and in the sample space. We do not claim that using KDTW and its probabilistic interpretation can solve all or even any of the fundamental questions raised earlier: since the problem tackled here is NP-complete, an exact solution requires exponentially complex computations and any heuristic method must handle numerous local minima. Our aim is to throw some new light on the problem as well as obtain new quantitative results showing, in this difficult context, that the proposed alternative approach is worth considering.

2.4 Time elastic kernels and their regularization

Dynamic Time Warping (DTW), [2], [3] as defined in Eq.1 can be recursively evaluated as

$$\begin{aligned} d_{dtw}(X_p, Y_q) &= d_E^2(x(p), y(q)) \\ &+ \text{Min} \begin{cases} d_{dtw}(X_{p-1}, Y_q) & sup \\ d_{dtw}(X_{p-1}, Y_{q-1}) & sub \\ d_{dtw}(X_p, Y_{q-1}) & ins \end{cases} \end{aligned} \tag{3}$$

where $d_E(x(p), y(q)$ is the Euclidean distance (eventually, the square of the Euclidean distance) defined on $\mathbb{R}^k$ between the two positions/?points in sequences X and Y taken at times p and q, respectively.

Apart from the fact that the triangular inequality does not hold for the DTW distance measure, it is furthermore not possible to define a positive definite kernel directly from this distance. Hence, the optimization problem, which is inherent to the learning of a kernel machine, is no longer quadratic and, at least for some tasks, could be a source of limitation.

Regularized DTW: recent studies [20], [19] lead us to propose new guidelines to ensure that kernels constructed from elastic measures such as DTW are positive definite. A simple instance of such a regularized kernel, derived from [19], can be expressed in the following form, which makes use of two recursive terms:

$$\text{KDTW}(X_p, Y_q) = K_{dtw}^{xy}(X_p, Y_q) + K_{dtw}^{xx}(X_p, Y_q)$$

$$K_{dtw}^{xy}(X_p, Y_q) = \tfrac{1}{3} e^{-\nu d_E^2(x(p), y(q))} \sum \begin{cases} h(p-1, q) K_{dtw}^{xy}(X_{p-1}, Y_q) \\ h(p-1, q-1) K_{dtw}^{xy}(X_{p-1}, Y_{q-1}) \\ h(p, q-1) K_{dtw}^{xy}(X_p, Y_{q-1}) \end{cases}$$

$$K_{dtw}^{xx}(X_p, Y_q) = \tfrac{1}{3} \sum \begin{cases} h(p-1, q) K_{dtw}^{xx}(X_{p-1}, Y_q) e^{-\nu d_E^2(x(p), y(p))} \\ \Delta_{p,q} h(p, q) K_{dtw}^{xx}(X_{p-1}, Y_{q-1}) e^{-\nu d_E^2(x(p), y(q))} \\ h(p, q-1) K_{dtw}^{xx}(X_p, Y_{q-1}) e^{-\nu d_E^2(x(q), y(q))} \end{cases} \tag{4}$$

where $\Delta_{p,q}$ is the Kronecker symbol, $\nu \in \mathbb{R}^+$ is a *stiffness* parameter which weights the local contributions, i.e. the distances between locally aligned positions, and $d_E(.,.)$ is a distance defined on $\mathbb{R}^k$.

The initialization is simply $K_{dtw}^{xy}(X_0, Y_0) = K_{dtw}^{xx}(X_0, Y_0) = 1$.

The main idea behind this regularization is to replace the operators $\min$ and $\max$ (which prevent symmetrization of the kernel) by a summation operator ($\sum$). This allows us to consider the best possible alignment, as well as all the best (or nearly the best) paths by summing their overall cost. The parameter ν is used to check what is termed as nearly-the-best alignment, thus penalizing alignments that are too far away from the optimal ones. This parameter can be

easily optimized through a cross-validation.

3 KDTW CENTROID AS A PREIMAGE PROBLEM

In this section, we tackle the centroid estimation question from a *kernelized centroid* point of view, the kernel of interest being KDTW.

The Moore-Aronszajn theorem [21] establishes that a reproducing kernel Hilbert space (RKHS) exists uniquely for every positive definite kernel and *vice-versa*. Let $\mathcal{H}$ be the RKHS associated to kernel κ defined on a set $\mathcal{X}$, and let $\langle .,. \rangle_{\mathcal{H}}$ be the inner product defined on $\mathcal{H}$. In addition, the representer property of the evaluation functional in $\mathcal{H}$ is expressed as: for any $\psi \in \mathcal{H}$ and any $x_j \in \mathcal{X}$, $\psi(x_j) = \langle \psi(.), \kappa(., x_j) \rangle_{\mathcal{H}}$.

Denoting $\phi(.)$ as the map that assigns the kernel function $\kappa(., x)$ to each input $x \in \mathcal{X}$, the reproducing property of the kernel implies that for any $(x_i, x_j) \in \mathcal{X}^2$, $\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$.

Furthermore, $D_{\mathcal{H}}(x_i, x_j)^2 = ||\phi(x_i) - \phi(x_j)||^2_{\mathcal{H}} = \langle \phi(x_i), \phi(x_i) \rangle_{\mathcal{H}} + \langle \phi(x_j), \phi(x_j) \rangle_{\mathcal{H}} - 2.\langle \phi(x_i), \phi(x_) \rangle_{\mathcal{H}}$ is the generalization of the squared Euclidean distance defined in the feature space $\mathcal{H}$, which can be expressed in kernel terms as: $D_{\mathcal{H}}(x_i, x_j)^2 = \kappa(x_i, x_i)^2 + \kappa(x_j, x_j)^2 - 2.\kappa(x_i, x_j)$ (the so-called kernel trick).

Finally, the representer theorem [22] states that any function $\varphi(.)^*$ of a RKHS $\mathcal{H}$ minimizing a regularized cost functional of the form:

$$\sum_{i=1}^{n} \mathbf{J}(\varphi(x_i), y_i) + g(||\varphi||^2_{\mathcal{H}})$$

with predicted output $\varphi(x_i)$ for input x_i and desired output y_j, where $g(.)$ is a strictly monotonically increasing function on $\mathbb{R}^+$-, is equivalent to a kernel expansion expressed in terms of available data ($\{(x_i, y_i)\}$)

$$\varphi^*(.) = \sum_{i=1}^{n} \gamma_i \kappa(x_i, .), \text{ where } \forall i, \gamma_i \in \mathbb{R}. \tag{5}$$

Hence, a direct definition of the kernelized centroid of the set $\{x_i, i = 1..n\}$ expressed in the RKHS $\mathcal{H}$ feature space associated with kernel κ can be written as:

$$\begin{aligned} \varphi^*(.) &= \arg\min_{\varphi(.) \in \mathcal{H}} \sum_{i=1}^{n} ||\varphi(.) - \kappa(., x_i)||^2_{\mathcal{H}} \\ &= \arg\min_{\varphi(.) \in \mathcal{H}} n \cdot ||\varphi(.)||^2_{\mathcal{H}} - 2 \cdot \sum_{j=1}^{n} \langle \varphi(.), \kappa(., x_j) \rangle_{\mathcal{H}} \end{aligned} \tag{6}$$

The representer theorem applies and thus $\varphi^*(.)$ takes the form given in Eq. 5, which allows us to rewrite Eq. 6 as follows:

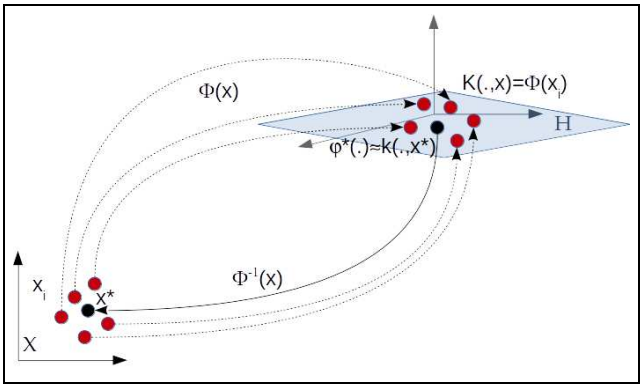


Fig. 2. Centroid estimation viewed as a preimage problem.

$$\begin{aligned} \varphi^*(.) = \arg\min_{\{\lambda_i\}_{i=1 \cdots n}} & \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_i \gamma_j \kappa(x_i, x_j) \\ & - 2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_j \kappa(x_i, x_j) \end{aligned} \tag{7}$$

Unfortunately, if the kernelized centroid is related to a well-defined quadratic optimization problem in the RKHS space $\mathcal{H}$ (Eq. 7), it is an ill-posed problem in set $\mathcal{X}$. This is known as the preimage problem, since the pre-image of $\phi(.)^*$ might not exist. Instead, we are seeking the best approximation, namely $x^* \in \mathcal{X}$ whose map $\phi(x^*) = \kappa(., x^*)$ is as close as possible to $\varphi(.)^*$, as illustrated in Fig.2.

Hence, if we remove the term that does depend upon x, the optimization problem becomes:

$$\begin{aligned} x^* &= \operatorname*{argmin}_{x \in \mathcal{X}} n \cdot ||\kappa(., x)||^2_{\mathcal{H}} - 2 \cdot \sum_{j=1}^{n} \langle \kappa(., x), \kappa(., x_j) \rangle_{\mathcal{H}} \\ &= \operatorname*{argmin}_{x \in \mathcal{X}} n \cdot \kappa(x, x) - 2 \cdot \sum_{j=1}^{n} \kappa(x, x_j) \end{aligned} \tag{8}$$

For KDTW, the non-convex optimization problem cannot be straightforwardly addressed using gradient-based approaches mainly because the derivative cannot be determined analytically. Moreover, the number of variables (linear with the length of the time series and with the dimensionality of each sample) is generally high so this approach often encounters combinatorial difficulties related to the number of local minima. A derivative-free method could nevertheless be applied for local modelling of the functional to be optimized. In an attempt to carry out such a preimage formulation to estimate the time elastic centroid for a set of time series, we applied the state-of-the-art BOBYQA algorithm developed for bound constrained optimization without using derivatives [23]. Fig.3 and Fig.4 give the centroid estimations for each category of the CBF and Trace datasets, respectively [24]. On the top left diagram of the figures, the values of the function to be minimized

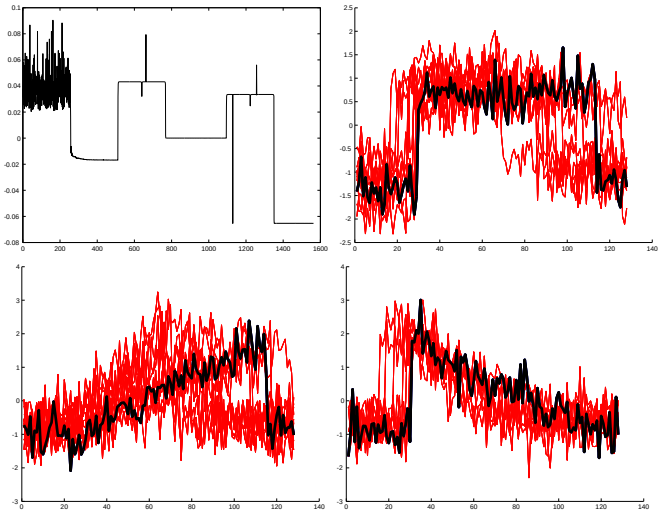


Fig. 3. Centroid estimation for the three categories contained in the CBF dataset as a solution of the preimage problem. In bold, the centroid time series; in light red, the time series of the averaged dataset. At top left of figure, the value of the minimized functional is plotted on a log-scale versus the iteration index.

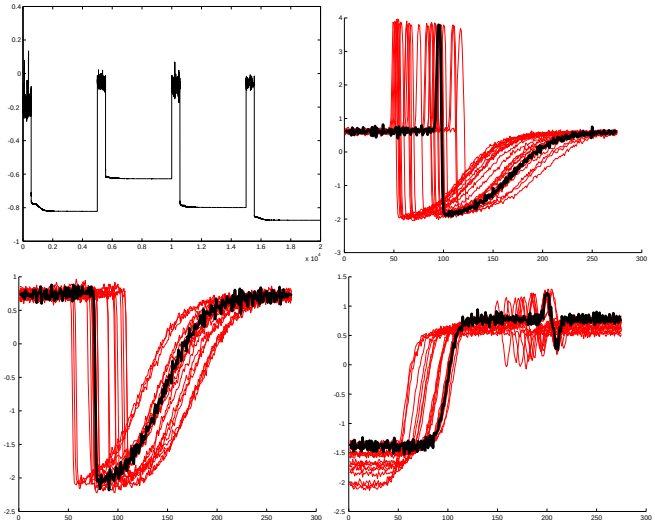


Fig. 4. Centroid estimations of the first three categories (out of four) contained in the Trace dataset as a solution of the preimage problem. In bold blue, the centroid time series; in light red, the time series of the averaged dataset. Top left diagram of figure shows value of the minimized functional expressed on a log-scale, plotted against the iteration index.

are plotted against the number of iterations. The optimization process is initialized using the medoid for each category. We show that the required number of iterations is quite high and depends on the number of variables. For the CBF dataset, the time series is made up of 128 samples while there are 275 samples for the Trace dataset. The convergence rate is roughly ten times slower for the Trace data set compared with the CBF dataset, mainly because KDTW complexity is quadratic with the length of the time series. The iteration cost becomes somewhat prohibitive for long time series or large time series datasets. Although this approach could be possibly optimized, the parameters need to be carefully set up (basically, definition of the trust region) and, in any case, as stated above, the optimum so provided remains an estimation of the centroid that is sought. Finally, note that the functional starts to decrease after attaining the number of iterations (in this case, twice the length of the time series) initially required for local estimation of the functional.

4 Probabilistic Interpretation of Time Elastic Kernel Alignment Matrices

In this section, we consider the recursive term $K_{dtw}^{xy}(.,.)$ that is used in Eq. 4. When evaluating the similarity between two time series X_p and Y_q with respective lengths of p and q, this recursion allows the construction of an alignment matrix $AM(i,j)$ with $i \in \{1 \cdots p\}$ and $i \in \{1 \cdots q\}$. The cell at location (i,j) contains the summation of the global costs of all alignment paths, as defined in *definition* 2.1, that connect cell $(1,1)$ with cell (i,j). For any alignment path π, the global cost is expressed as:

$$cost(\pi) = \prod_{k=1}^{|\pi|} e^{-\nu d_E^2(X(\pi_k(1)),Y(\pi_k(2)))} \quad (9)$$

i.e. the product along the path of the local alignment costs. We can give a probabilistic interpretation of these local costs $exp(-\nu d_E^2(X(\pi_k(1)),Y(\pi_k(2))))$: basically, we can assume that these local costs correspond (within the magnitude of the scalar multiplication constant) to the local *a priori* probability of aligning sample $X(\pi_k(1))$ with sample $Y(\pi_k(2))$. By making this assumption, we eventually attach a probability distribution to the set of all alignment paths, with the $cost(\pi)$ corresponding (within the magnitude of the scalar multiplication constant) to the probability attached to alignment path π.

Hence, the cell (i,j) of matrix AM, contains the sum of the probabilities (within the magnitude of the scalar multiplication constant) of the paths that connect cell $(1,1)$ to cell (i,j).

Similarly, if, instead of X and Y, we evaluate the similarity between X_r and Y_r derived from X and Y by reversing the temporal index, we obtain an alignment matrix AM_r whose cell (i,j) contains the sum of the probabilities (to within a multiplicative scalar constant) of the paths that connect cell (p,q) tp cell (i,j).

Finally, multiplying properly cells of AM with cells of AM_r yields the Alignment Matrix Average (AMA) defined as:

$$AMA(i,j) = AM(i,j) \cdot AM_r(p-i+1, q-j+1) \quad (10)$$

and whose cell (i,j) contains the sum of the probabilities (upto the normalization constant) of the paths that connect cell $(1,1)$ to cell (p,q) while going through the cell (i,j).

From this path probability distribution, we can now derive an alignment probability distribution between the samples of X and the samples of Y as follows:

- For all i, the probability of aligning sample $X(i)$ is $P(i) = 1$; all samples need to be aligned.
- Similarly, for all j, the probability of aligning sample $Y(j)$ is $P(j) = 1$.
- The probability of aligning sample $X(i)$ with sample $Y(j)$ is $P(i,j) = P(i|j).P(j) = P(i|j)$. $P(i|j)$ is the probability that sample $X(i)$ is aligned with sample $Y(j)$ given that the alignment process is in state j. The estimation of $P(i|j)$ is obtained by using matrix AMA:
$$P(i|j) = \frac{AMA(i,j)}{\sum_{i=1}^{p} AMA(i,j)}$$
- Furthermore, the probability of aligning sample $X(i)$ with sample $Y(j)$ is also $P(i,j) = P(j|i).P(i) = P(j|i)$. Similarly, the estimation of $P(j|i)$ is obtained by using matrix AMA:
$$P(j|i) = \frac{AMA(i,j)}{\sum_{j=1}^{q} AMA(i,j)} \quad (11)$$

Note that the normalization constant mentioned above is eliminated.

Since $P(i,j) = P(i|j) = P(j|i)$, we can finally estimate the probability of aligning sample $X(i)$ with sample $Y(j)$ as follows:

$$P(i,j) = \frac{1}{2} \cdot \left(\frac{AMA(i,j)}{\sum_{i=1}^{p} AMA(i,j)} + \frac{AMA(i,j)}{\sum_{j=1}^{q} AMA(i,j)} \right) \quad (12)$$

Eq. 12 forms the basis of our pairwise time elastic time series averaging algorithm given below.

As an example, Fig 5 presents the AMA matrix corresponding to the alignment of a positive halfwave with a sinus wave. The three potential alignment pathes are clearly identified in the light blue and red colors.

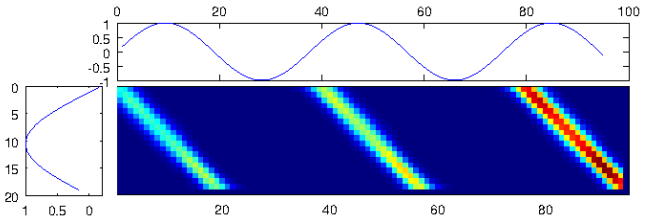


Fig. 5. AMA matrix for the alignment of a positive halfwave with a sinus wave.

5 Time elastic centroid based on the AMA alignment matrix

Based on the structure of the KDTW kernel and the AMA matrix, and by using the so-called DtwBarycenter Averaging (DBA) method developed by [12], [14], [13], we first present the KernelDtwNarycenter Averaging (KDBA) algorithm for estimating a time elastic centroid for a set of time series according to an iterative agglomerative procedure as shown in Fig. 1b. Secondly, we detail the concept of a time elastic average for a pair of time series (KDTW-PWA), and then develop the progressive heuristic approach presented in Fig. 1a that uses KDTW-PWA to estimate another kind of time elastic centroid (KDTW-C1) for a set of time series of any cardinal.

5.1 KDTW-Centroid of a set of time series based on KDBA algorithm

Following the DBA algorithmic approach [12], [13], we present here the development of our kernelized version called KDBA. KDBA directly applies the definition of the alignment matrix average (AMA) as given in Eq.10 and its probabilistic interpretation Eq.12.

Algorithm 1 KDBA

procedure KDBA(R,S, ν)
  // R: a reference time series
  // S: a set of time series $\{S_1, \cdots, S_N\}$
  // ν: the stiffness parameter of KDTW kernel
  Double AMA(.,.);
  Vector-Of-SetOfSamples SampleAssociations(L);
  Ts $A(|R|)$; //Create a D dimensional
    //time series of length L;
  for Int $i = 1$ to $|R|$ **do** SampleAssociations(i)={};
  for Int $n = 1$ to $|S|$ **do**
    Evaluate AMA matrix for R, S_n with ν;
    Ts ts//containing L "zeroed" samples;
    Double $normFactor(|R|)$;
    for Int $i = 1$ to $|R|$ **do**
      normFactor(i)=0;
      for Int $j = 1$ to $|S_n|$ **do**
        $ts(i) = ts(i) + S_n(j) * AMA(i,j)$;
        $normFactor(i) = normFactor(i) +$
          $AMA(i,j)$;
      $ts(i) = ts_1(i)/normFactor(i)$;
      SampleAssociations(i)=$(ts(i))$;
  for Int $i = 1$ to $|R|$ **do**
    A(i)=barycenter(SampleAssociations(i));
  return A

Let us consider a set S of N time series, $S = \{S_1, S_2, \cdots, S_N\}$, and R a reference time series. Let $|R|$ and $|S_n|$ be the lengths of R and S_n, respectively. $P_n(i,j)$, with $i = 1\{1, ...|S_n|\}$ and $j = 1\{1, ...|R|\}$, is obtained from the AMA matrix resulting from the alignment of S_n with R, according to Eq.12. Algorithm 1 computes an average

Algorithm 2 iKDBA

procedure iKDBA(C,S, ν)
  //C: a reference time series
  //S: a set of time series
  //maxIter: maximum number of iterations
  //ν: the stiffness parameter of KDTW kernel
  Ts A; //a D dimensional Timeseries
  Double inertia = computeInertia(C, S);
  Boolean Continue=True;
  Int $i = 0$;
  while Continue **do**
    A=C;
    C=KDTW-C2(C,S, ν);
    Double new_inertia = computeInertia(C, S);
    if new_inertia $>$ inertia OR $i > maxIter$ **then**
      Continue = False;
    i=i+1;
  return A

time series A according to the following equation:

$$\forall i \in \{1, \cdots, |r|\},\ \mathtt{A}(i) = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{|S_n|} P_n(i,j) S_n(j) \quad (13)$$

Note that the iterative average of time series produced by algorithm 1 has the same size as the reference time series R.

The algorithm 1 can be refined by iterating until no further improvement is obtained [14]. An improvement is observed when the sum of the distances (resp. similarities) between the current average R and the new pairwise average provided by KDBA, A, is lowered (resp. increased). Algorithm 2 implements this iterative strategy, which will necessarily find a local minimum or will stop when a maximum number of iterations has been reached.

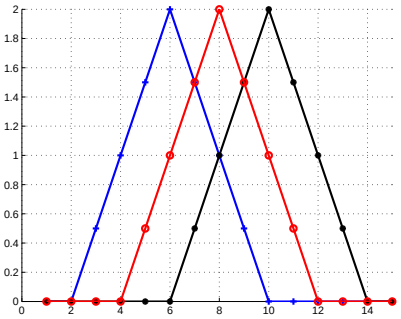


Fig. 6. Expected time location for the centroid (in red circles) of two triangular-shaped time series shifted in time (in blue '+' and black '*')

5.2 KDTW average of a pair of time series (KDTW-PWA)

Algorithm 3 KDTW-PWA

procedure KDTW-PWA(X,Y, AMA)
  //X,Y: two time series of D dimensional samples
  //AMA: the average alignment matrix for X,Y
  Int $p = |X|$, $q = |Y|$, $L = \max\{p, q\}$;
  Ts $A(L)$, $B(L)$; //Create 2 D dimensional
    //time series of length L;
  Double α;
  Double $N_A(L), N_B(L)$; //two double arrays
  for Int $i = 1$ to L **do**
    for d=1 to D **do**
      $A(i,d) = 0$, $B(i,d) = 0$;
    $N_A(i) = 0$, $N_B(i) = 0$;
  for Int $i = 1$ to L **do**
    if $i < p$ **then**
      for Int $j = 1$ to q **do**
        $\alpha = (i+j)/2 - \lfloor (i+j)/2 \rfloor$;
        for d=1 to D **do**
          $A(\lfloor (i+j)/2 \rfloor, d) +=$
          $\alpha \cdot (X(i,d) + Y(j,d)) \cdot AMA(i,j)$;
          $A(\lceil (i+j)/2 \rceil, d) +=$
          $(1-\alpha) \cdot (X(i,d) + Y(j,d)) \cdot AMA(i,j)$;
        $N_A(\lfloor (i+j)/2 \rfloor) += \alpha * AMA(i,j)$;
        $N_A(\lceil (i+j)/2 \rceil) += (1-\alpha) * AMA(i,j)$;
    if $i < q$ **then**
      for Int $j = 1$ to p **do**
        $\alpha = (i+j)/2 - \lfloor (i+j)/2 \rfloor$;
        for d=1 to D **do**
          $B(\lfloor (i+j)/2 \rfloor, d) +=$
          $\alpha \cdot (X(j,d) + Y(i,d)) \cdot AMA(j,i)$;
          $B(\lceil (i+j)/2 \rceil, d) +=$
          $(1-\alpha) \cdot (X(j,d) + Y(i,d)) \cdot AMA(j,i)$;
        $N_B(\lfloor (i+j)/2 \rfloor) += \alpha * AMA(j,i)$;
        $N_B(\lceil (i+j)/2 \rceil) += (1-\alpha) * AMA(j,i)$;
  for Int $i = 1$ to L **do**
    for d=1 to D **do**
      $A(i,d) = (A(i,d)/N_A(i) + B(i,d)/N_B(i))/4$;
  return A

Similarly to DBA, the KDBA algorithm averages a set of time series in the sample space but not along the time axis. Basically, let us suppose we are averaging two triangular-shaped time series such as represented by the blue crosses and black dots on Fig.5.1. When using DBA or KDBA algorithms with one of the two time series acting as the reference, then the calculated average would be the reference distribution itself. However, we would also expect to average the time shift between the two series, thus obtaining the distribution indicated by the red dots in Fig.fig:time-shift. This is precisely our main motivation for the deriving the following Pair Wise Averaging (KDTW-PWA) algorithm designed to average a pair of time series in the sample space but also along

the time axis.

Algorithm 3 provides the KDTW-PWA average (A) of the two time series X and Y according to Eq.14.

$$\forall k = 1 \cdots L, \quad A(k) = \sum_{i,j|\frac{i+j}{2}=k} \left(P(i,j) \cdot \frac{X(i)+Y(j)}{2} \right) = \sum_{i,j|\frac{i+j}{2}=k} \left(\frac{P(i|j)+P(j|i)}{2} \cdot \frac{X(i)+Y(j)}{2} \right) \quad (14)$$

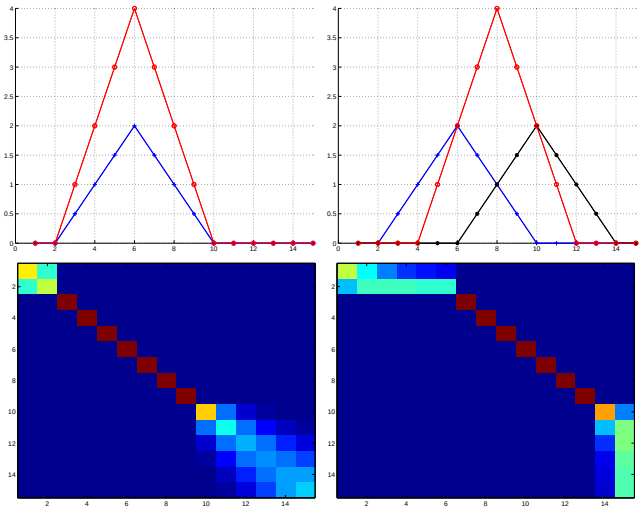


Fig. 7. Averaging triangular-shaped time series. On the left, the two time series (in blue) are identical (superimposed) and the centroid (red) is amplified by a factor of two. On the right, the two time series (in blue) have the same shape but have been shifted in time. The KDTW-PWA average is given in red, still amplified by a factor of two. The corresponding (normalized) AMA alignment matrices are given at the bottom.

As the time indices are considered discrete (integer values), the time averaging $(i+j)/2$ is smoothed between the floor and cell integer values, using the smoothing coefficient α (line 17 of the algorithm).

Thus, the KDTW-PWA jointly averages the sample values of the two time series and their time locations. Eq. 14 allows us to interpret the centroid of a pair of time series as the mathematical expectation of aligning the two sequences of samples.

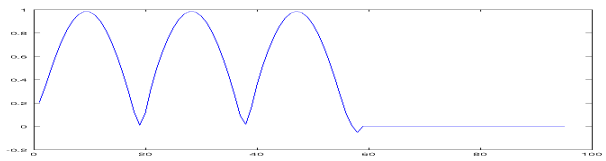


Fig. 8. Centroid corresponding to the pairwise alignment for the sinus experiment depicted in Fig 5.

As an example, the centroid corresponding to the pairwise alignment of the sinus experiment depicted in Fig 5 is presented in Fig 8. Notice that in the centroid, the negative halfwaves of the sine wave have been *filtered*. This is because the negative halfwaves do not match with the positive halfwave that is aligned with the sine wave.

In Fig 7, we present a very simple experiment that consists of averaging two identical triangular-shaped time series (on left of figure) and two time series with identical triangular shapes but shifted in time. At the bottom of the figure, the corresponding AMA matrices are presented. The KDTW-PWA distributions, presented in red, are multiplied by a factor of two to facilitate reading of the figure. We can see that, for both situations, the centroid is precisely located at the correct averaged time of occurrence of the two time series, whether or not they are shifted in time. The most likely alignment areas on the AMA matrices are shown in in red and the less likely alignment areas in blue. The time shift is clearly visible on the right-hand figure.

5.3 KDTW-Centroid of a set of time series based on KDTW-PWA

Algorithm 4 pKDTW-PWA

```
1:  procedure pKDTW-PWA(S, ν)
2:      //S: a set of time series of D dimensional samples
3:      //ν: the stiffness parameter of KDTW kernel
4:      Ts A; //a D dimensional time series
5:      SetOfTimeSeries S_0;
6:      while |S| > 1 do
7:          S_0 = ∅
8:          while |S| > 1 do
9:              Let ts_1, ts_2 the first two time series in S;
10:             Evaluate the AMA matrix for ts_1 and ts_2
11:                     with ν as the stiffness parameter
12:             A = KDTW-PWA(ts_1, ts_2, AMA);
13:             S_0 = S_0 ∪ {A};
14:             S = S \ {ts_1, ts_2};
15:         S = S_0 ∪ S;
16:     Let A be the single element of S;
17:     return A
```

To average a larger set of time series using the pairwise average KDTW-PWA, we simply adopt the progressive agglomerative approach presented in Fig.1a. This heuristic approach, detailed in Algorithm 4 has $O(n)$ complexity, n being the size of the considered set of time series.

The figures presented in Table 1 compare the centroid estimates provided by the iterated DBA, iKDBA and pKDTW-PWA algorithms. For the experiment, The DBA and iKDBA were iterated at most 20 times. Although the DBA and iKDBA estimates appear quite similar, the centroid estimates provided by the pKDTW-PWA algorithm is much smoother. This is a general property of the latter algorithm, which implements a time averaging principle based on the time expectation of sample occurrences, thus somehow allowing it to filter *noisy* data. Note also

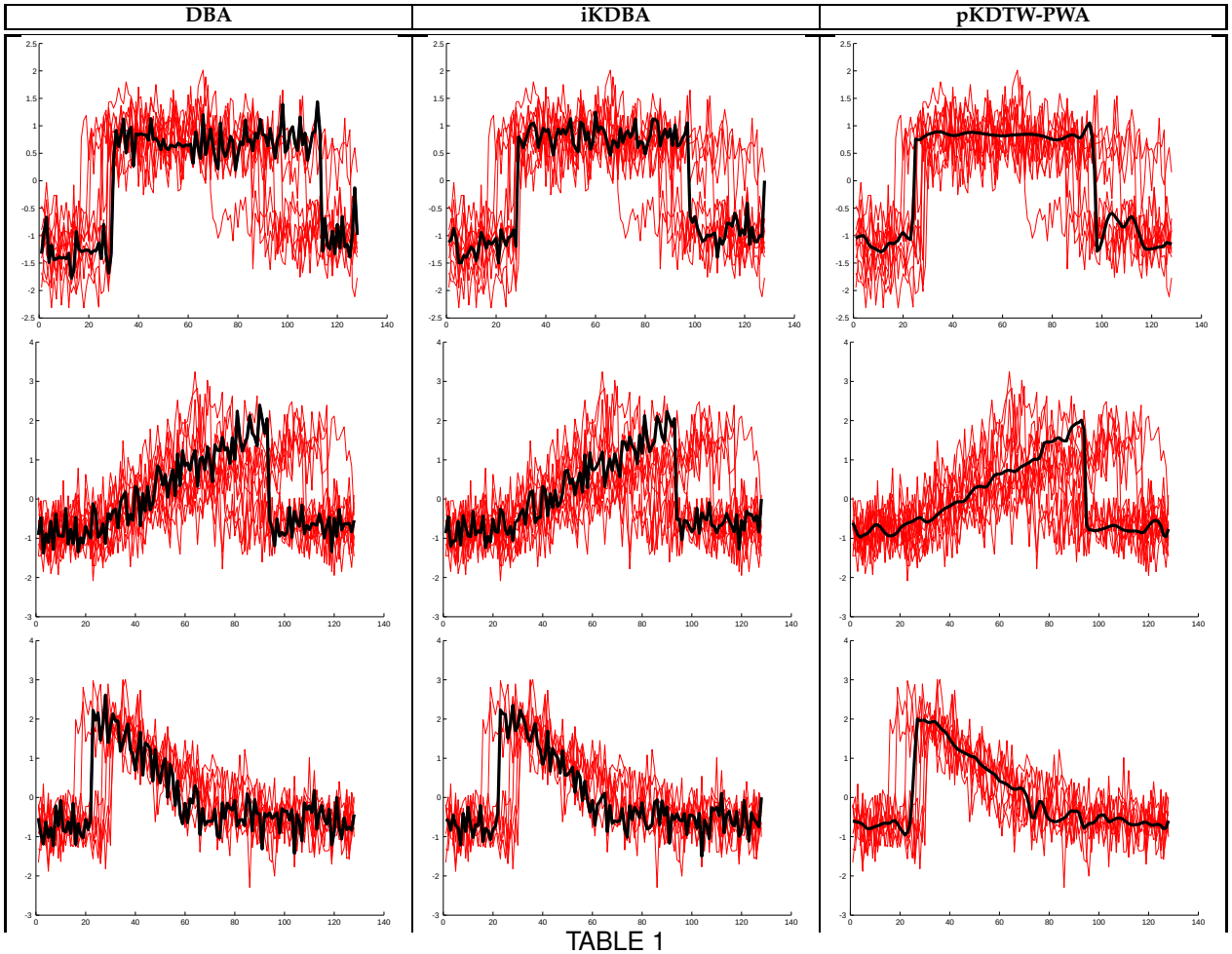


TABLE 1
Centroid estimation for the three categories of the CBF dataset. The centroid estimation is indicated as a bold black line superimposed on of the time series (in light red) that are averaged. The centroid estimates provided by the DBA algorithm are given on the left side, the estimates provided by the iKDBA algorithm in the centre and the estimates provided by the pKDTW-PWA algorithm on the right side.

that the DBA and iKDBA estimates for the CBF data set are close to the results provided by the preimage approach (Fig.3).

6 Experimentation

The purpose of this experiment is to evaluate the effectiveness of the proposed time elastic averaging methods against a double baseline, namely k-medoid-based approaches and the DBA algorithm. The first baseline allow us to compare centroid-based with medoid-based approaches. The second baseline highlights the advantages we can expect from using p.d elastic kernels instead of indefinite kernels such as DTW in the context of time series averaging. DBA is also currently considered as a state of the art method to average a set of sequences consistently with DTW.

For this purpose, we empirically evaluate the effectiveness of the methods using a first nearest centroid/medoid (1-NC) classification task on a set of time series derived from widely diverse fields of application. The task consists of representing each category contained in a training data set by estimating its medoid or centroid and then evaluating the error rate of a 1-NC classifier on an independent testing data set. Hence, the classification rule consists of assigning to the tested time series the category which corresponds to the closest (or most similar) medoid or centroid according to DTW or KDTW measures.

In [25] a nice generalized k-NC task is described. The authors demonstrate that by selecting the appropriate number k of centroids (using DBA and k-means), they achieve, without loss, a 70% speed-up in average, compared to the original k-Near Neighbor task. Although, in general, the classification accuracies is improved when several centroids are used to represent the TRAIN datasets, our main purpose is to highlight and amplify the discrimination between time series averaging methods: this is why stick here with the 1-NC task.

DBA and iKDBA iterative centroid methods are

iterated at most 20 times and yield local estimates of the centroid. The pKDTW-PWA progressive agglomerative centroid method is only processed once, and hence is roughly 20 times faster than iKDBA and about 10 times faster than DBA.

A collection of 45 data sets is used to assess the proposed algorithms. The collection includes synthetic and real data sets, as well as univariate and multivariate time series. These data sets are distributed as follows:

- 42 of these data sets are available at the UCR repository [24]. Basically, we used all the data sets except for *StarLightCurves*, *Non-Invasive Fetal ECG Thorax1* and *Non-Invasive Fetal ECG Thorax2*. Although these last three data sets are still tractable, their computational cost is high because of their size and the length of the time series they contain. All the data sets are composed of scalar time series.
- One data set, uWaveGestureLibrary_3D was constructed from the uWaveGestureLibrary_X—Y—Z scalar data sets to compose a new set of multivariate (3D) time series.
- One data set, CharTrajTT, is available at the UCI Repository [26] under the name *Character Trajectories Data Set*. This data set contains multivariate (3D) time series and is divided into two equal sized data sets (TRAIN and TEST) for the experiment.
- The last data set, *PWM2*, which stands for Pulse Width Modulation [27], was specifically defined to demonstrate a weakness in dynamic time warping (DTW) pseudo distance. This data set is composed of artificial scalar time series.

For each dataset, a training subset (TRAIN) is defined as well as an independent testing subset (TEST). We use the training sets to extract single medoids or centroid estimates for each of the categories defined in the data sets.

Furthermore, for KDTW$_{Medoid}$, iKDBA and pKDTW-PWA, the ν parameter is optimized using a *leave-one-out* (LOO) procedure carried out on the TRAIN data sets. The ν value is selected within the discrete set $\{.05, .1, .25, .5, 1, 2, 5, 10, 25, 50, 100\}$. The value that minimizes the LOO classification error rate on the TRAIN data is then used to provide the error rates that are estimated on the TEST data.

The classification results are given in Table 2. It can be seen from this experiment, that

i) Centroid-based methods outperform medoid-based methods: DBA yields lower error rates compared to DTW$_{Medoid}$, as do iKDBA and pKDTW-PWA compared to KDTW$_{Medoid}$.
ii) iKDBA outperforms DBA: under the same experimental conditions (maximum of 20 iterations), the kernalized version of the DTW measure leads to better classification accuracy. To some extent, this confirms previous results obtained for SVM classification [19] on such kinds of datasets.
iii) pKDTW-PWA outperforms iKDBA: this results seems to show that joint averaging in the sample space and along the time axis improves the classification accuracy. As pKDTW-PWA provides a centroid estimation in a single agglomerative step, we can conjecture that this method converges faster toward a satisfactory centroid candidate.

The average ranking for all five tested methods, which supports our preliminary conclusion, is given at the bottom of Table 2.

Following the study of [28] on statistical tests available to evaluate the significance of differences in error rate between classifiers over multiple data sets, we conducted a Friedman's significance test, a sort of non-parametric counterpart of the well-known ANOVA. This test ranks the algorithms for each data set separately, the best performing algorithm being given a rank of 1, the second best rank 2, etc.

According to this test, the null hypothesis is rejected (with a $P - value < 2.2e - 16$). Post-hoc tests can then be carried out to compare pairwise algorithms using the Wilcoxon-Nemenyi-McDonald-Thompson test [29]. For this purpose, we use the R code provided by [30] to generate the parallel coordinate plots and boxplots presented in Fig.9 as well as the results reported in Table 3.

TABLE 3
Significance test: $Algorithm_1$ is considered to be significantly better than $Algorithm_2$ according to the Friedman's test if the P-value (in bold characters) associated with the pairwise test is less than 0.05.

$Algorithm_1$	$Algorithm_2$	P-value
DBA	DTW$_{Medoid}$	**1.98e-05**
KDTW$_{Medoid}$	DTW$_{Medoid}$	**2.99e-03**
iKDBA	DTW$_{Medoid}$	**1.38e-10**
pKDTW-PWA	DTW$_{Medoid}$	**1.09e-12**
KDTW$_{Medoid}$	DBA	7.84e-01
iKDBA	DBA	3.10e-01
pKDTW-PWA	DBA	**2.60e-02**
iKDBA	KDTW$_{Medoid}$	**1.90e-02**
pKDTW-PWA	KDTW$_{Medoid}$	**4.07e-04**
pKDTW-PWA	iKDBA	8.36e-01

Table 3 reports the P-values for each pair of tested algorithms. This post-hoc analysis partially confirms our previous analysis of the classification results. If we consider that the null hypothesis is rejected when the P-value is less than 0.05, the post-hoc analysis shows that centroid-based approaches perform significantly better than medoid-based approaches. Furthermore, KDTW$_{Medoid}$ appears to be significantly better than DTW$_{Medoid}$.

TABLE 2
Comparative study using the UCR and UCI data sets: classification error rates evaluated on the TEST data set (in %) obtained using the first nearest neighbour classification rule for DTW$_{Medoid}$, DBA (centroid), KDTW$_{Medoid}$, $iKDBA$ (centroid) and $pKDTW-PWA$ (centroid). A single medoid/centroid extracted from the TRAIN data set represents each category.

DATASET	# Cat | L	DTW$_{Medoid}$	DBA	KDTW$_{Medoid}$	iKDBA	pKDTW-PWA
Synthetic_Control	6|60	3.00	**2.00**	3.33	**2.00**	4.67
Gun_Point	2|150	44.00	32.00	52.00	**25.33**	**25.33**
CBF	3|128	7.89	5.33	8.11	**4.67**	5
Face_(all)	14|131	25.21	18.05	20.53	17.34	**17.04**
OSU_Leaf	6|427	64.05	56.20	53.31	**52.89**	54.54
Swedish_Leaf	15|128	38.56	30.08	31.36	30.24	**24.00**
50Words	50|270	48.13	41.32	23.40	20.44	**19.34**
Trace	4|275	5.00	7.00	23.00	20.00	**2.00**
Two_Patterns	4| 128	1.83	1.18	1.17	**1.03**	1.12
Wafer	2|152	64.23	33.89	43.92	**12.11**	31.96
Face_(four)	4|350	12.50	13.64	17.05	**6.82**	10.23
Lightning-2	2|637	34.43	37.70	29.51	29.51	**22.95**
Lightning-7	7|319	27.40	27.40	19.18	**17.81**	20.55
ECG200	2|96	32.00	28.00	29.00	28.00	**27.00**
Adiac	37|176	57.54	52.69	**40.67**	72.12	41.43
Yoga	2|426	47.67	47.87	**47.53**	49.80	49.90
Fish	7|463	38.86	30.29	20.57	19.42	**17.14**
Beef	5|470	60.00	**53.33**	56.67	**53.33**	**53.33**
Coffee	2|286	57.14	32.14	32.14	32.14	**21.43**
OliveOil	4|570	26.67	16.67	30	20.00	**13.33**
CinC_ECG_torso	4|1639	74.71	53.55	66.67	59.85	**49.64**
ChlorineConcentration	3|166	65.96	68.15	**65.65**	67.94	65.78
DiatomSizeReduction	4|345	22.88	5.88	11.11	5.56	**1.96**
ECGFiveDays	2|136	47.50	30.20	**10.92**	19.75	17.88
FacesUCR	14|131	27.95	18.44	20.73	16.63	**15.61**
Haptics	5|1092	68.18	64.61	63.64	59.74	**57.47**
InlineSkate	7|1882	78.55	76.55	78.36	**74.73**	75.82
ItalyPowerDemand	2|24	31.68	20.99	**5.05**	6.31	6.22
MALLAT	8|1024	6.95	6.10	6.87	4.22	**3.58**
MedicalImages	10|99	67.76	58.42	58.68	**58.03**	61.71
MoteStrain	2|84	15.10	13.18	12.70	13.58	**9.42**
SonyAIBORobot_SurfaceII	2|65	26.34	**21.09**	26.230	23.29	25.81
SonyAIBORobot_Surface	2|70	38.10	19.47	39.77	15.31	**7.65**
Symbols	6|398	7.64	4.42	3.92	3.82	**3.62**
TwoLeadECG	2|82	24.14	**13.17**	27.04	17.65	22.39
WordsSynonyms	25|270	70.85	64.26	64.26	63.32	**58.15**
Cricket_X	12|300	67.69	**52.82**	61.79	57.17	61.28
Cricket_Y	12|300	68.97	52.82	46.92	**44.61**	54.87
Cricket_Z	12|300	73.59	**48.97**	56.67	51.79	59.74
uWaveGestureLibrary_X	8|315	38.97	33.08	34.34	**32.94**	33.42
uWaveGestureLibrary_Y	8|315	49.30	44.44	42.18	40.31	**40.14**
uWaveGestureLibrary_Z	8|315	47.40	**39.25**	41.96	40.39	39.84
uWaveGestureLibrary_3D	8|315	10.11	**6.00**	13.74	25.65	8.43
CharTrajTT_3D	20|178	6.58	5.18	**4.20**	11.83	4.34
PWM2	3|128	43.00	35.00	21.00	20.33	**11.67**
# Best Scores	-	0	8	6	13	**22**
# Uniquely Best Scores	-	0	6	3	10	**20**
Average rank	-	4,29	2,8	3,16	2.22	**1,89**

Furthermore, pKDTW-PWA is evaluated as significantly better than DBA but not significantly better than iKDBA in this experiment. Note also that DBA is not shown to perform significantly better than KDTW$_{Medoid}$.

This post-hoc analysis is summarized in Fig.10 which shows the ranking graph for the five algorithms tested

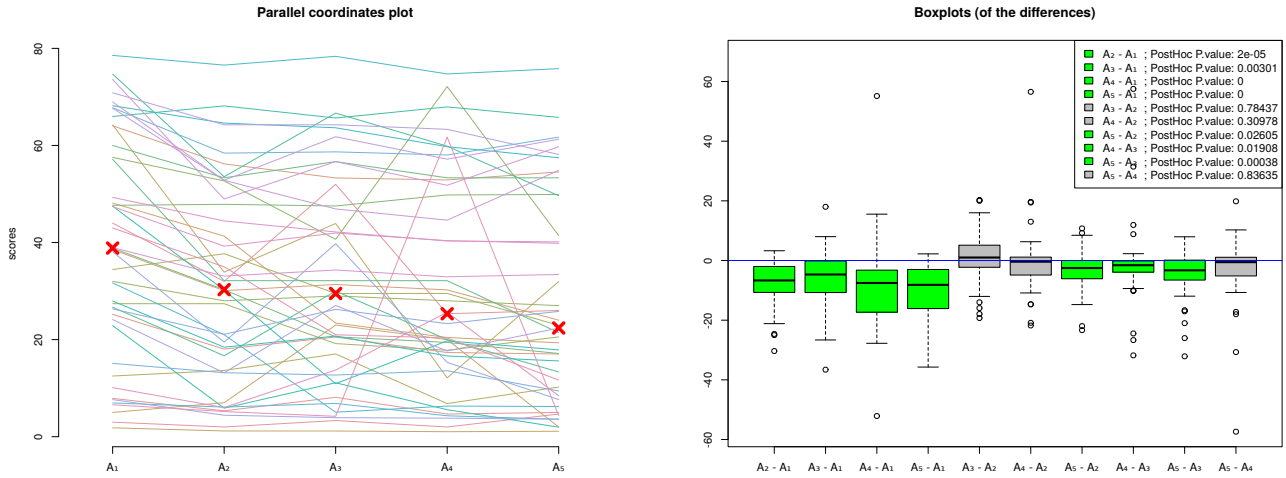


Fig. 9. *Post hoc* analysis of the Friedman's test: (A_1) DTW$_{Medoid}$, (A_2) DBA, (A_3) KDTW$_{Medoid}$, (A_4) iKDBA and (A_5) pKDTW-PWA.

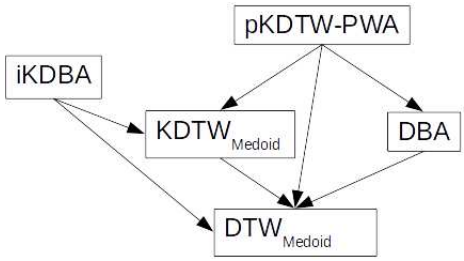


Fig. 10. Dominance graph for the five tested algorithms, according to the significance relation corresponding to Table 3 with a P-value threshold set at .05.

in our experiments.

7 Conclusion

In this paper, we address the reputedly difficult problem of averaging a set of time series in the context of a time elastic distance measure such as Dynamic Time Warping. The new perspective provided by the kernelization of the elastic distance firstly allows us to consider the averaging of time series as a preimage problem. This latter is unfortunately an ill-posed non-convex problem that could suffer from combinatorial number of local *optima* when dealing with long multidimensional time series. Furthermore, this kind of preimage problem can only be resolved using gradient-free optimization procedures that are computationally very costly (since extensive functional evaluation is required).

However, this new kernelization approach allows a re-interpretation of pairwise kernel alignment matrices as distributions of probability over alignment paths. Based on this re-interpretation, we propose two distinct algorithms, iKDBA and pKDTW-PWA, based on iterative and progressive agglomerative heuristic methods, respectively, that are developed to compute approximate solutions to the multi-alignment of time series.

We present an extensive experiment carried out on synthetic and real data sets, mostly containing univariate but also some multivariate time series. Our results show that centroid-based methods significantly outperform medoid-based methods in the context of a first nearest neighbour classification task. Most strikingly, the pKDTW-PWA algorithm, which integrates joint averaging in the sample space and along the time axis, is significantly better than the state-of-the art DBA algorithm, with a potentially lower computational cost. Indeed, the simple one-pass progressive agglomerative heuristic procedure is used in the pKDTW-PWA algorithm can be further optimized.

Acknowledgments

The authors thank the French Ministry of Research, the Brittany Region, the General Council of Morbihan and the European Regional Development Fund that partially funded this research. The authors also thank the promoters of the UCR and UCI data repositories for providing the time series data sets used in this study.

References

[1] M. Fréchet, *Sur quelques points du calcul fonctionnel*, ., Ed. Thèse, Faculté des sciences de Paris., 1906.

[2] V. M. Velichko and N. G. Zagoruyko, "Automatic recognition of 200 words," *International Journal of Man-Machine Studies*, vol. 2, pp. 223–234, 1970.

[3] H. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *Proceedings of the 7th International Congress of Acoustic*, 1971, pp. 65–68.

[4] L. Chen and R. Ng, "On the marriage of lp-norms and edit distance," in *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, ser. VLDB '04. VLDB Endowment, September 2004, pp. 792–803.

[5] P.-F. Marteau, "Time warp edit distance with stiffness adjustment for time series matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 306–318, Feb 2009.

[6] K. H. Fasman and S. S. L., "An introduction to biological sequence analysis," in *Computational Methods in Molecular Biology,*. In Salzberg, S.L., Searls, D.B., and Kasif, S., eds., Elsevier, 1998, pp. 21–42.

[7] L. Wang and T. Jiang, "On the complexity of multiple sequence alignment." *Journal of Computational Biology*, vol. 1, no. 4, pp. 337–348, 1994.

[8] W. Just and W. Just, "Computational complexity of multiple sequence alignment with sp-score," *Journal of Computational Biology*, vol. 8, pp. 615–623, 1999.

[9] H. Carrillo and D. Lipman, "The multiple sequence alignment problem in biology," *SIAM J. Appl. Math.*, vol. 48, no. 5, pp. 1073–1082, Oct. 1988. [Online]. Available: http://dx.doi.org/10.1137/0148063

[10] L. Gupta, D. Molfese, R. Tammana, and P. Simos, "Nonlinear alignment and averaging for estimating the evoked potential," *Biomedical Engineering, IEEE Transactions on*, vol. 43, no. 4, pp. 348–356, April 1996.

[11] V. Niennattrakul and C. Ratanamahatana, "Shape averaging under time warping," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2009. ECTI-CON 2009. 6th International Conference on*, vol. 02, May 2009, pp. 626–629.

[12] W. Abdulla, D. Chow, and G. Sin, "Cross-words reference template for dtw-based speech recognition systems," in *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region*, vol. 4, Oct 2003, pp. 1576–1579 Vol.4.

[13] V. Hautamaki, P. Nykanen, and P. Franti, "Time-series clustering by approximate prototypes," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, Dec 2008, pp. 1–4.

[14] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recogn.*, vol. 44, no. 3, pp. 678–693, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2010.09.013

[15] F. Petitjean and P. Gançarski, "Summarizing a set of time series by averaging: From Steiner sequence to compact multiple alignment," *Journal of theoretical computer science*, vol. 414, no. 1, pp. 76–91, Jan. 2012.

[16] S. Soheily-Khal, A. Douzal-Chouakria, and E. Gaussier, "Time series centroid estimation under weighted and kernel dynamic time warping," *Personal communication (under submission)*, 2014.

[17] H. Shimodaira, K. I. Noma, M. Nakai, and S. Sagayama, "Dynamic Time-Alignment Kernel in Support Vector Machine," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002.

[18] V. Niennattrakul and C. Ratanamahatana, "Inaccuracies of shape averaging method using dynamic time warping for time series data," in *Computational Science – ICCS 2007*, ser. Lecture Notes in Computer Science, Y. Shi, G. van Albada, J. Dongarra, and P. Sloot, Eds. Springer Berlin Heidelberg, 2007, vol. 4487, pp. 513–520. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72584-8_68

[19] P.-F. Marteau and S. Gibet, "On Recursive Edit Distance Kernels with Application to Time Series Classification," *IEEE Trans. on Neural Networks and Learning Systems*, pp. 1–14, Jun. 2014. [Online]. Available: http://hal.inria.fr/hal-00486916

[20] M. Cuturi, J.-P. Vert, O. Birkenes, and T. Matsui, "A kernel for time series based on global alignments," in *IEEE ICASSP 2007*, vol. 2, April 2007, pp. II–413–II–416.

[21] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, 1950.

[22] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, ser. COLT '01/EuroCOLT '01. London, UK, UK: Springer-Verlag, 2001, pp. 416–426. [Online]. Available: http://dl.acm.org/citation.cfm?id=648300.755324

[23] M. J. D. Powell, "The bobyqa algorithm for bound constrained optimization without derivatives," Aug. 2009.

[24] E. J. Keogh, X. Xi, L. Wei, and C. Ratanamahatana, "The UCR time series classification-clustering datasets," 2006, http://wwwcs.ucr.edu/ eamonn/time_series_data/.

[25] F. Petitjean, G. Forestier, G. Webb, A. Nicholson, Y. Chen, and E. Keogh, "Dynamic time warping averaging of time series allows faster and more accurate classification," in *Proceedings of the 14th IEEE International Conference on Data Mining*, 2014, pp. 470–479. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2014.27

[26] M. Lichman, "Uci machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[27] P.-F. Marteau, "Pulse width modulation data sets," 2007. [Online]. Available: http://people.irisa.fr/Pierre-Francois.Marteau/PWM/

[28] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1248547.1248548

[29] M. Hollander and D. Wolfe, *Nonparametric Statistical Methods*, ser. Wiley Series in Probability and Statistics. Wiley, 1999. [Online]. Available: https://books.google.fr/books?id=RJAQAQAAIAAJ

[30] T. Galili, "R code for the friedman test post hoc analysis." february 2010. [Online]. Available: http://www.r-statistics.com/2010/02/post-hoc-analysis-for-friedmans-test-r-code/