Time Series Alignment with Global Invariances

Titouan Vayer, Laetitia Chapel, Nicolas Courty, Rémi Flamary, Yann Soullard, Romain Tavenard

***Abstract*—Multivariate time series are ubiquitous objects in signal processing. Measuring a distance or similarity between two such objects is of prime interest in a variety of applications, including machine learning, but can be very difficult as soon as the temporal dynamics and the representation of the time series, *i.e.* the nature of the observed quantities, differ from one another. In this work, we propose a novel distance accounting both feature space and temporal variabilities by learning a latent global transformation of the feature space together with a temporal alignment, cast as a joint optimization problem. The versatility of our framework allows for several variants depending on the invariance class at stake. Among other contributions, we define a differentiable loss for time series and present two algorithms for the computation of time series barycenters under this new geometry. We illustrate the interest of our approach on both simulated and real world data and show the robustness of our approach compared to state-of-the-art methods.**

***Index Terms*—Time series alignment, global invariance, dynamic time warping**

I. Introduction

Time series are subject to a number of variabilities that make their processing difficult in practice. One of the most well-known example is the temporal shift, usually handled using the celebrated Dynamic Time Warping (DTW, [1]) that aligns, in times, two time series and is invariant to any monotonically increasing temporal map. It has been initially introduced for speech processing applications and is now widely used in a variety of contexts such as human activity recognition [2], satellite image analysis [3] or medical applications [4, 5].

Another source of variability in time series is feature space alterations, that may occur due to a permutation of sensors, changes in sensor properties or even different number of sensors. This problem of heterogeneous representations, also called distribution shifts in machine learning, has been studied mostly on non-structured data. It has been shown that algorithms are notoriously weak [6] when it comes to generalizing to out-of-distribution samples, as they rely on the correlations that are found in the training data [7]. Consequently, dedicated paradigms such as domain adaptation [8] directly take into account this problem in the learning process. Another approach is to learn with respect to some invariance classes (based on some prior knowledge) in order to be more robust to irrelevant feature transformations [9, 10]. In this work, we aim at tackling this problem in the time series context through the definition of similarity measures that naturally encode desirable invariances. More precisely, we introduce similarity measures that are able to deal with both temporal and feature space transformations.

There exist many frameworks to register different spaces under some classes of invariance. In the shape analysis community, matching objects under rigid transformations is a widely studied problem. Iterative Closest Point (ICP, [11]) is a standard algorithm for such a task. It acts by alternating two simple steps: (i) matching points using nearest neighbor search and (ii) registering shapes together based on the obtained matches, which is known as the orthogonal Procrustes problem that has a closed form solution [12]. This idea is further explored in [13, 14], where optimal transport is used to match points in the first step, and a recent extension to objects with a hierarchical structure has been introduced in [15] that considers a dedicated invariance class for the registration step.

This heterogeneous setting has also been investigated in the time series context, where the goal is to align series of features lying in different spaces. One of the most salient track of research in this setting is the Canonical Time Warping (CTW) method. CTW [16] has been introduced for human motion alignment under rigid space transformations. It consists of temporal alignment (using DTW) of transformed time series, using Canonical Correlation Analysis (CCA) to define the feature space transform. Few extensions to CTW have been proposed. GTW [17] parametrizes CTW temporal alignments in continuous time instead of relying on DTW. Deep CTW [18] extends CTW by learning a feature space embedding (in the form of a neural network) before performing CTW. Finally, Canonical soft Time Warping applies the CTW methodology to soft alignments (see Sec. II for more details on soft alignments). In the same vein as CTW, [19] learns an invariant subspace based on DTW alignments. More recently, GromovDTW [20] has been introduced as an extension of the Gromov-Wasserstein distance measure between heterogeneous distributions to the time series context. GromovDTW relies on time series self-similarities as a way to circumvent the need to compute distances across feature spaces. Compared to these approaches, our method works by optimizing a map between feature spaces, hence allowing one to (i) add prior information in the form of constraints on the set of allowed maps and (ii) use the computed map for downstream application, as illustrated in our experiments on MoCap data (as described in Sec. IV).

In more details, we aim at tackling both temporal and feature space invariances. To do so, we state the problem as a joint optimization over temporal alignments and feature space transformations, as depicted in Figure 1. Our general framework allows the use of either DTW or its smoothed counterpart softDTW as an alignment procedure. Similarly, though rigid transformations of the feature space seem a reasonable invariance class, we show that our method can be used in conjunction with other families of transformations. Such a framework allows considering the case when time series differ both in length and feature space dimensionality. We introduce two different optimization procedures that could be used to tackle this problem and show experimentally that they lead to effectively invariant similarity measures. Our method can also be used to compute meaningful barycenters even when time series at stake do not lie in the same feature space. Finally,

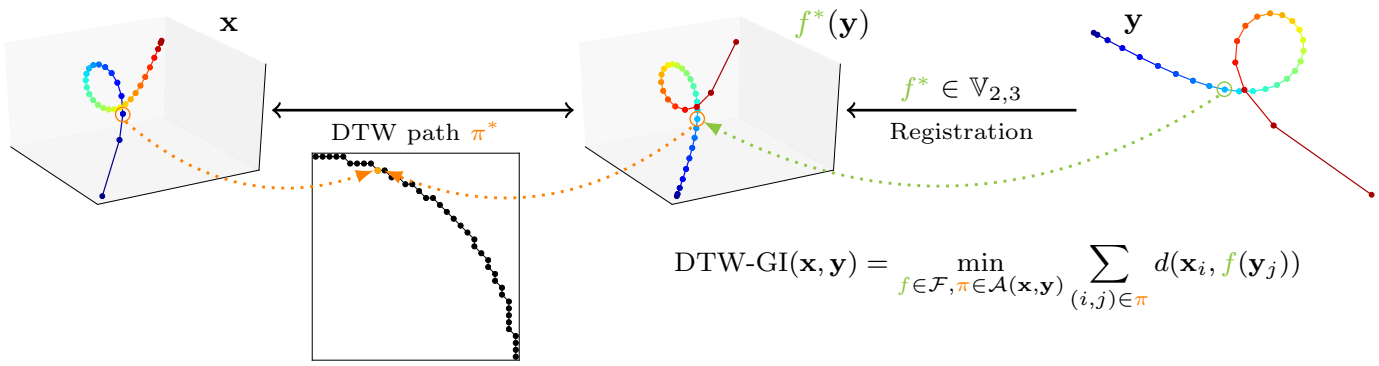

Fig. 1. DTW-GI aligns time series by optimizing on temporal alignment (through Dynamic Time Warping) and feature space transformation (denoted f here). Time series represented here are color-coded trajectories, whose starting (resp. end) point is depicted in blue (resp. red).

we showcase the versatility of our method and the importance of jointly learning feature space transformations and temporal alignments on two real-world applications that are time series forecasting for human motion and cover song identification.

II. Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW, [1]) is an algorithm used to assess similarity between time series, with extensions to multivariate time series proposed in [21, 22]. In its standard form, given two multivariate time series $\mathbf{x} \in \mathbb{R}^{T_x \times p}$ and $\mathbf{y} \in \mathbb{R}^{T_y \times p}$ of the same dimensionality p, DTW is defined as:

$$\text{DTW}(\mathbf{x}, \mathbf{y}) = \min_{\pi \in \mathcal{A}(\mathbf{x},\mathbf{y})} \sum_{(i,j)\in\pi} d(\mathbf{x}_i, \mathbf{y}_j) \tag{1}$$

where $\mathcal{A}(\mathbf{x}, \mathbf{y})$ is the set of all admissible alignments between $\mathbf{x}$ and $\mathbf{y}$ and d is a ground metric. In most cases, d is the squared Euclidean distance, *i.e.* $d(\mathbf{x}_i, \mathbf{y}_j) = \|\mathbf{x}_i - \mathbf{y}_j\|^2$.

An alignment π is a sequence of pairs of time frames which is considered to be admissible iff (i) it matches first (and respectively last) indexes of time series $\mathbf{x}$ and $\mathbf{y}$ together, (ii) it is monotonically increasing and (iii) it is connected (*i.e.* every index from one time series must be matched with at least one index from the other time series). Efficient computation of the above-defined similarity measure can be performed in quadratic time using dynamic programming, relying on the following recurrence formula:

$$\begin{aligned} DTW(\mathbf{x}_{\to t_1}, \mathbf{y}_{\to t_2}) &= d(\mathbf{x}_{t_1}, \mathbf{y}_{t_2}) \\ &+ \min \begin{cases} DTW(\mathbf{x}_{\to t_1}, \mathbf{y}_{\to t_2 - 1}) \\ DTW(\mathbf{x}_{\to t_1 - 1}, \mathbf{y}_{\to t_2}) \\ DTW(\mathbf{x}_{\to t_1 - 1}, \mathbf{y}_{\to t_2 - 1}) \end{cases} \end{aligned} \tag{2}$$

Many variants of this similarity measure have been introduced. For example, the set of admissible alignment paths can be restricted to those lying around the diagonal using the so-called Itakura parallelogram or Sakoe-Chiba band, or a maximum path length can be enforced [23]. Most notably, a differentiable variant of DTW, coined softDTW, has been introduced in [24] and is based on previous works on alignment kernels [25]. It replaces the min operation in Equation (2) by a soft-min operator $\min^\gamma$ whose smoothness is controlled by a parameter $\gamma > 0$, resulting in the DTW_γ distance:

$$\begin{aligned} \text{DTW}_\gamma(\mathbf{x}, \mathbf{y}) &= \min_{\pi \in \mathcal{A}(\mathbf{x},\mathbf{y})}{}^{\gamma} \sum_{(i,j)\in\pi} d(\mathbf{x}_i, \mathbf{y}_j) \\ &= -\gamma \log \left(\sum_{\pi \in \mathcal{A}(\mathbf{x},\mathbf{y})} e^{-\sum_{(i,j)\in\pi} d(\mathbf{x}_i, \mathbf{y}_j)/\gamma} \right). \end{aligned} \tag{3}$$

In the limit case $\gamma = 0$, $\min^\gamma$ reduces to a hard min operator and DTW_γ is defined as equivalent to the DTW algorithm.

III. DTW with Global Invariances

Despite their widespread use, DTW and softDTW are not able to deal with time series of different dimensionalities or to encode feature transformations that may arise between time series. In the following, we introduce a new similarity measure aiming at aligning time series in this complex setting and provide ways to compute associated alignments. We also derive a Fréchet mean formulation that allows computing barycenters under this new geometry.

A. Definitions

Let $\mathbf{x} \in \mathbb{R}^{T_x \times p_x}$ and $\mathbf{y} \in \mathbb{R}^{T_y \times p_y}$ be two time series. In the following, we assume $p_x \geq p_y$ without loss of generality. In order to allow comparison between time series $\mathbf{x}$ and $\mathbf{y}$, we optimize on a family of functions $\mathcal{F}$ that map $\mathbf{y}$ onto the feature space in which $\mathbf{x}$ lies. More formally, we define Dynamic Time Warping with Global Invariances (DTW-GI) as the solution of the following joint optimization problem:

$$\text{DTW-GI}(\mathbf{x}, \mathbf{y}) = \min_{f \in \mathcal{F}, \pi \in \mathcal{A}(\mathbf{x},\mathbf{y})} \sum_{(i,j)\in\pi} d(\mathbf{x}_i, f(\mathbf{y}_j)), \tag{4}$$

where $\mathcal{F}$ is a family of functions from $\mathbb{R}^{p_y}$ to $\mathbb{R}^{p_x}$. Note that this problem can also be written as:

$$\text{DTW-GI}(\mathbf{x}, \mathbf{y}) = \min_{f \in \mathcal{F}, \pi \in \mathcal{A}(\mathbf{x},\mathbf{y})} \langle W_\pi, C(\mathbf{x}, f(\mathbf{y})) \rangle \tag{5}$$

where $f(\mathbf{y})$ is a shortcut notation for the transformation f applied to all observations in $\mathbf{y}$, $\langle ., . \rangle$ denotes the Frobenius inner product, W_π is defined as:

$$\forall i \leq T_x, j \leq T_y,\ (W_\pi)_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in \pi \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

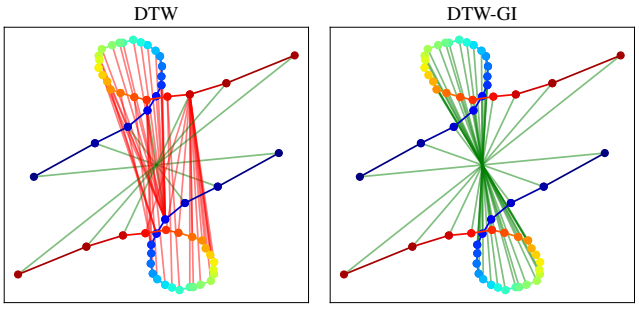

Fig. 2. Example alignments between 2D time series (trajectories in the plane). Color coding corresponds to timestamps. Our DTW-GI method jointly estimates temporal alignment and global rotation between time series. On the contrary, standard DTW alignment fails at capturing feature space distortions and therefore produces mostly erroneous alignment (matching in red), except at the beginning and end of the time series, whose alignments are preserved thanks to DTW border constraints (cf. Section II).

and $C(\mathbf{x}, f(\mathbf{y}))$ is the cross-similarity matrix of squared Euclidean distances between samples from $\mathbf{x}$ and $f(\mathbf{y})$, respectively. This definition can be extended to the softDTW case of Equation (3) as proposed in the following:

$$\begin{aligned}\text{DTW}_\gamma\text{-GI}(\mathbf{x},\mathbf{y}) &= \min_{f\in\mathcal{F}} \min_{\pi\in\mathcal{A}(\mathbf{x},\mathbf{y})}{}^{\gamma} \langle W_\pi, C(\mathbf{x}, f(\mathbf{y}))\rangle \\ &= \min_{f\in\mathcal{F}} -\gamma \log \sum_{\pi\in\mathcal{A}(\mathbf{x},\mathbf{y})} e^{-\langle W_\pi, C(\mathbf{x},f(\mathbf{y}))\rangle/\gamma}\end{aligned} \tag{7}$$

Note that, due to the use of a soft-min operator, Equation (7) is no longer a joint optimization.

These similarity measures estimate both temporal alignment and feature space transformation between time series simultaneously, allowing the alignement of time series when the similarity should be defined up to a global transformation. For instance, one can see in Figure 2 two temporal alignments between two series in 2D that have been rotated in their feature space. In this case DTW-GI, whose invariant is the space of rotations, recovers the proper alignment whereas DTW fails.

Properties of DTW-GI : By definition, DTW-GI and softDTW-GI are invariant under any global transformation $T(\cdot)$ such that $\{f \circ T \,|\, f \in \mathcal{F}\} = \mathcal{F}$ (*i.e.* $\mathcal{F}$ is stable under T), which motivates the name (soft)DTW with Global Invariances. It is also straightforward to see that DTW-GI$(\mathbf{x}, \mathbf{x}) = 0$ for any time series $\mathbf{x}$ as soon as $\mathcal{F}$ contains the identity map.

B. Optimization

Optimization on the above-defined losses can be performed in several ways, depending of the nature of $\mathcal{F}$. We now present one optimization scheme for each loss.

1) Gradient descent: We first consider the optimization on the softDTW-GI loss (7) in the case where $\mathcal{F}$ is a parametric family of functions, here denoted f_θ, that are differentiable with respect to their parameters θ. The optimization can be done with a gradient descent on the parameters of f_θ. Since softDTW is smooth (contrary to DTW), this strategy can be used to compute gradients of DTW$_\gamma$-GI *w.r.t.* θ.

Complexity for this approach is driven by (i) that of a softDTW computation and (ii) that of computing $f_\theta(\mathbf{y})$. If we denote the latter c_f, overall complexity for this approach is hence $O(n_{\text{iter}}(T_x T_y p_x + c_f))$. Note that when Riemannian optimization is involved, an extra complexity term has to be added, corresponding to the cost of projecting gradients onto the considered manifold. This cost is $O(p_y^3)$ for example when optimization is performed on the Stiefel manifold [26], which is an important case for our applications, as discussed in more details in the following.

2) Block Coordinate Descent (BCD): When DTW-GI (5) is concerned, we introduce another strategy that consists in alternating minimization over (i) the temporal alignment and (ii) the feature space transformations. We will refer to this strategy as Block Coordinate Descent (BCD) in the following.

Optimization over the alignment path given a fixed transformation f solely consists in a DTW alignment, as described in Section II. For a fixed alignment path π, the optimization problem then becomes:

$$\min_{f\in\mathcal{F}} \langle W_\pi, C(\mathbf{x}, f(\mathbf{y}))\rangle\,. \tag{8}$$

Recall that C is a matrix of squared distances, which means that the problem above is a weighted least square problem. Depending on $\mathcal{F}$, there can exist a closed form solution for this problem (*e.g.* when $\mathcal{F}$ is the set of affine maps with no further constraints). Let us first note that the matrix C can be rewritten as:

$$C(\mathbf{x}, f(\mathbf{y})) = \mathbf{u}_\mathbf{x} + \mathbf{v}_{f(\mathbf{y})}^T - 2\mathbf{x}f(\mathbf{y})^T \tag{9}$$

where $\mathbf{u}_\mathbf{x} = (\|x_1\|^2, \dots, \|x_{T_x}\|^2)^T$ and $\mathbf{v}_{f(\mathbf{y})} = (\|f(y_1)\|^2, \dots, \|f(y_{T_y})\|^2)^T$. Note that the optimization problem reduces to the linear term on the right if $\mathcal{F}$ is a set of norm preserving operations.

3) Estimating f in the Stiefel manifold: Let us consider the special case where $\mathcal{F}$ is the set of linear maps whose linear operator is an orthonormal matrix, hence lying on the Stiefel manifold that we denote $\mathbb{V}_{p_y,p_x}$ in the following. This invariance class encodes rigid transformations of the features. In this case, the optimization problem becomes:

$$\min_{\mathbf{P}\in\mathbb{V}_{p_y,p_x}} \left\langle W_\pi, \mathbf{u}_\mathbf{x} + \mathbf{v}_{f(\mathbf{y})}^T - 2\mathbf{x}\mathbf{P}\mathbf{y}^T \right\rangle \tag{10}$$

and we have $\mathbf{v}_{f(\mathbf{y})} = (\|y_1\|^2, \dots, \|y_{T_y}\|^2)^T = \mathbf{v}_\mathbf{y}$ since the considered applications are norm-preserving. Overall, we get the following optimization problem:

$$\min_{\mathbf{P}\in\mathbb{V}_{p_y,p_x}} \left\langle W_\pi, \mathbf{u}_\mathbf{x} + \mathbf{v}_\mathbf{y}^T \right\rangle - 2\left\langle W_\pi, \mathbf{x}\mathbf{P}\mathbf{y}^T \right\rangle \tag{11}$$

whose solution is equivalent to solving:

$$\max_{\mathbf{P}\in\mathbb{V}_{p_y,p_x}} \left\langle W_\pi, \mathbf{x}\mathbf{P}\mathbf{y}^T \right\rangle = \max_{\mathbf{P}\in\mathbb{V}_{p_y,p_x}} \left\langle \mathbf{x}^T W_\pi \mathbf{y}, \mathbf{P} \right\rangle \tag{12}$$

since the term $\left\langle W_\pi, \mathbf{u}_\mathbf{x} + \mathbf{v}_\mathbf{y}^T \right\rangle$ does not depend in $\mathbf{P}$.

As described in [27], the latter problem can be solved exactly using Singular Value Decomposition (SVD): if $U\Sigma V^T = M$ is the SVD of a matrix M of shape (p_y, p_x), then $S^\star = UV^T$ is a solution to the linear problem $\sup_{S\in\mathbb{V}_{p_y,p_x}} \langle S, M\rangle$. Note that this method can also tackle the case where $\mathcal{F}$ is an affine map whose linear part lies in the Stiefel manifold by realigning time series means, as discussed for example in [28]. A sketch of

Algorithm 1 Block-Coordinate Descent for DTW-GI with Stiefel registration

$\mathbf{P} \leftarrow I_{p_x,p_y}$
repeat
  $W_\pi \leftarrow$ Alignment matrix from $DTW(\mathbf{x}, \mathbf{y}\mathbf{P}^T)$
  $M \leftarrow \mathbf{x}^T W_\pi \mathbf{y}$ (cf. Eq. (12))
  $U, \Sigma, V^T \leftarrow SVD(M)$
  $\mathbf{P} \leftarrow UV^T$
until convergence

the algorithm (for the simplified case where time series means do not have to be realigned) is presented in Algorithm 1.

Interestingly, this optimization strategy where we alternate between time series alignment, *i.e.* time correspondences between both time series, and feature space transform optimization can be seen as a variant of the Iterative Closest Point (ICP) method in image registration [11], in which nearest neighbors are replaced by matches resulting from DTW alignment. Its overall complexity is then $O(n_{\text{iter}}(T_x T_y p_x + p_x^2 p_y))$. This complexity is equal to that of the gradient-descent when $p_x = O(p_y)$. However, in practice, the number of iterations required is much lower for this BCD variant, making it a very competitive optimization scheme, as discussed in Section IV.

The algorithms presented above are mainly focused on optimization on the Stiefel manifold. Note however that they are not strictly restricted to this case. Typically, (projected) gradient descent based optimization could be performed on any family of functions parametrized by a neural network. Regarding the BCD algorithm, it requires a numerically efficient way to compute the optimal feature space transform for a fixed alignment. In our experiment, we illustrate this in the context of cover song identification, for which aligning song keys is a well-known registration step (see Sec. IV-E for details).

C. Barycenters

Let us now assume we are given a set $\{\mathbf{x}^{(i)}\}_i$ of time series of possibly different lengths and dimensionalities. A barycenter of this set in the DTW-GI sense is a solution to the following optimization problem:

$$\min_{\mathbf{b} \in \mathbb{R}^{T \times p}} \sum_i w_i \min_{f_i \in \mathcal{F}} DTW(\mathbf{x}^{(i)}, f_i(\mathbf{b})), \tag{13}$$

where weights $\{w_i\}_i$ as well as barycenter length T and dimensionality p are provided as input to the problem. Note that, with this formulation, when $\mathcal{F}$ is the Stiefel manifold, p is supposed to be lower or equal to the dimensionality of any time series in the set $\{\mathbf{x}^{(i)}\}_i$.

In terms of optimization, as for similarity estimation, two schemes can be used. First, softDTW-GI barycenters can be estimated through gradient descent (and when the set of series to be averaged is large, a stochastic variant relying on minibatches can easily be implemented). Second, when BCD is used for time series alignment, barycenters can be estimated using a similar approach as DTW Barycenter Averaging (DBA, [29]), that would consist in alternating between barycentric coordinate estimation and DTW-GI alignments.

IV. Experiments

In this section, we provide an experimental study of DTW-GI (and its soft counterpart) on simulated data and real-world datasets. Unless otherwise specified, the set $\mathcal{F}$ of feature space transforms is the set of affine maps whose linear part lies in the Stiefel manifold. In all our experiments, tslearn [30] implementation is used for baseline methods and gradient descent on the Stiefel manifold is performed using GeoOpt [31, 32] in conjunction with PyTorch [33]. Open source code of our method will be released upon publication.

A. Timings

We are first interested in a quantitative evaluation of the temporal complexity of our methods. Note that the theoretical complexity of DTW and softDTW are the same, hence any difference observed in this series of experiments between DTW-GI and softDTW-GI would be solely due to their optimization schemes discussed in Section III-B. In these experiments, the number of iterations for BCD as well as the number of gradient steps for the gradient descent optimizer are set to 5,000. The BCD algorithm used for DTW-GI is stopped as soon as it reaches a local minimum, while early stopping is used for the gradient-descent variant with a patience parameter set to 100 iterations.

We first study the computation time as a function of the length of the time series involved. To do so, we generate random time series in dimension 8 and vary their lengths from 8 to 1,024 timestamps.

Figure 3 (left) shows a clear quadratic trend for all methods presented, except GromovDTW whose complexity is cubic *w.r.t.* the length of the time series due the tensor-matrix multiplication that is involved at each step of its pseudo-Frank-Wolfe algorithm [20]. Note that DTW-GI and its BCD optimizer clearly outperform the gradient descent strategy used for softDTW-GI because the latter requires more iterations before early stopping can be triggered. Building on this, we now turn our focus on the impact of feature space dimensionality p (with a fixed time series length of 32). DTW and softDTW baselines are asymptotically linear with respect to p. Similarly, since GromovDTW relies on pre-computed self-similarity matrices, it only linearly depends in the feature space dimensionality for the computation of these self-similarity matrices. Since feature space registration is performed through optimization on the Stiefel manifold, both our optimization schemes rely on Singular Value Decomposition, which leads to an $O(p^3)$ complexity that can also be observed for both methods in Figure 3 (right). Note also that the CTW baseline is slightly more computationally expensive than DTW-GI in practice, even if asymptotic complexities are the same as for DTW-GI.

B. Rotational invariance

We now evaluate the ability of our method to recover invariance to rotation. To do so, we rely on a synthetic dataset of noisy spiral-like 2d trajectories. For increasing values of an angle α, we generate pairs of spirals rotated by α with additive gaussian noise. Alignments between a reference time

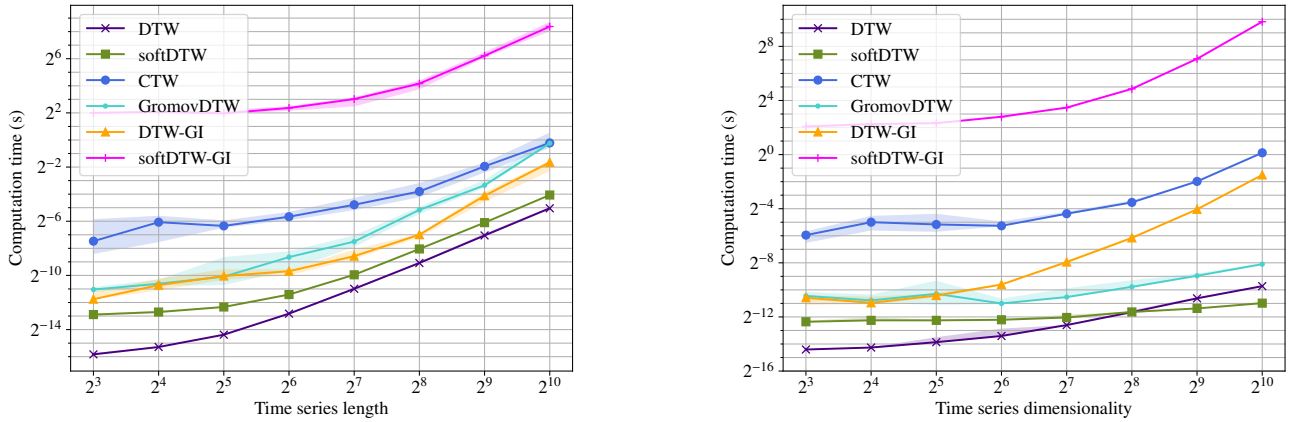

Fig. 3. Computing time as a function of time series length (left) and dimensionality (right). Solid lines correspond to median values and shaded areas correspond to 20th (resp. 80th) percentiles.

series and variants that are subject to an increasing rotation are computed and repeated 50 times per angle. The ratio of each distance to the distance when $\alpha = 0$ is reported in Figure 4 (left). One can clearly see that the GI counterparts of DTW and softDTW are invariant to rotation in the 2d feature space, while DTW and softDTW are not. Interestingly, CTW and GromovDTW, that should be invariant to rotation, still exhibit an increase in the loss with the angle α, suggesting that their algorithm has more difficulties reaching a global minimum in practice. Also, when varying the noise level in Figure 4 (right), one can notice that (soft-)DTW-GI are slightly more robust to high levels of noise than the CTW baseline, while GromovDTW is very sensitive to this noise level (the GW loss is a quadratic function of its input).

C. Barycenter computation

So as to better grasp the notion of similarity captured by our methods, we compute barycenters using the strategy presented in Section III-C. Barycenters are computed for 3 different datasets: the first two are made of 2d trajectories of rotated and noisy spirals or folia, and the third one is composed of both 2- and 3-dimensional spirals (see samples in the left part of Figure 5). For each dataset, we provide barycenters obtained by three baseline methods. DTW Barycenter Averaging (DBA, [29]) is used for DTW while softDTW resorts to a gradient-descent scheme to compute the barycenters. «««< HEAD Their GI counterpart use the same algorithms but rely on the alignments obtained from DTW-GI and softDTW-GI respectively. Finally, GromovDTW is optimized by alternating between computation of the barycenter self-similarity matrix and alignments, as done in [20]. Note that the DTW and softDTW baselines cannot be used for the third dataset since features of the time series do not lie in the same ambient space. We would like to emphasize that the barycenter based on GromovDTW only finds a pairwise distance matrix from which the positions of the points must be inferred, for example by applying multidimensional scaling (MDS) [34] (as done here and in [20]).

For the 2d spiral dataset, all the reconstructed barycenters can be considered as meaningful. Note however that the outer loop of the spiral (the one that suffers the most from the rotation) is better reconstructed using DTW-GI and softDTW-GI variants. When it comes to the folia trajectories, that are more impacted by rotations, baseline barycenters fail to capture the inherent structure of the trajectories at stake, while both our methods generate smooth and representative barycenters. DTW-GI and softDTW-GI are even able to recover barycenters when datasets are made of series that do not lie in the same space, as shown in the third row of Figure 5. Finally, in all three settings considered, temporal alignments successfully capture the irregular sampling from the samples to be averaged (denser towards the center of the spiral / loop of the folium).

D. Time series forecasting

To further illustrate the benefit of our approach, we consider a time series forecasting problem [35], where the goal is to infer the future of a partially observed time series. In this setting, we suppose that we have access to a training set of full time series $\mathbf{X}$, with $\mathbf{x}^{(i)} \in \mathbf{X}$ a time series of length T and dimensionality p_x, and another test set of partial time series $\mathbf{Y}$ where each $\mathbf{y} \in \mathbf{Y}$ is of length $T' < T$ and dimensionality p_y. The goal is to predict the values for timestamps T' to T for each test time series. We will denote by $\mathbf{x}_{\to T'}$ the beginning of the time series $\mathbf{x}$ (up to time T') and $\mathbf{x}_{T' \to}$ its end (from time T' to time T).

Let $d(\mathbf{y}, \mathbf{x}^{(i)})$ denote a dissimilarity measure between time series $\mathbf{y}$ and $\mathbf{x}^{(i)}$ associated with a transformation $f_i \in \mathcal{F} : \mathbb{R}^{p_x} \to \mathbb{R}^{p_y}$ that maps the features of $\mathbf{x}^{(i)}$ onto the features of $\mathbf{y}$. This function aims at capturing the desired invariances in the feature space, as described in the previous section. A typical example is when d is the softDTW-GI cost, then the f_i are the Stiefel linear maps which capture the possible rigid transformation between the features. We propose to predict the future of a time series $\mathbf{y}$ as follows:

$$\hat{\mathbf{y}}_{T' \to} = \sum_i a_d \left(\mathbf{y}_{\to T'}, \mathbf{x}^{(i)}_{\to T'} \right) f_i \left(\mathbf{x}^{(i)}_{T' \to} \right) \qquad (14)$$

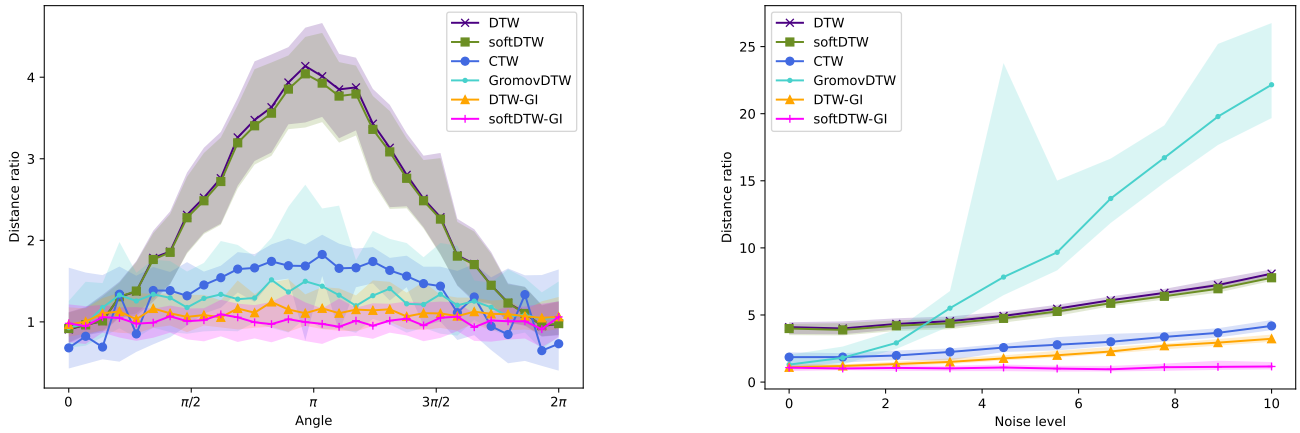

Fig. 4. Illustration of the rotation invariance provided by DTW-GI. On the left, ratio of the distance to that of a non-rotated pair of spirals is presented as a function of the rotation angle for a fixed noise level. On the right, ratio of the distance to that of a non-rotated pair of spirals is presented as a function of the noise level for a fixed rotation angle π. Median distance ratios are reported as solid lines and shaded areas correspond to 20th (resp. 80th) percentiles.

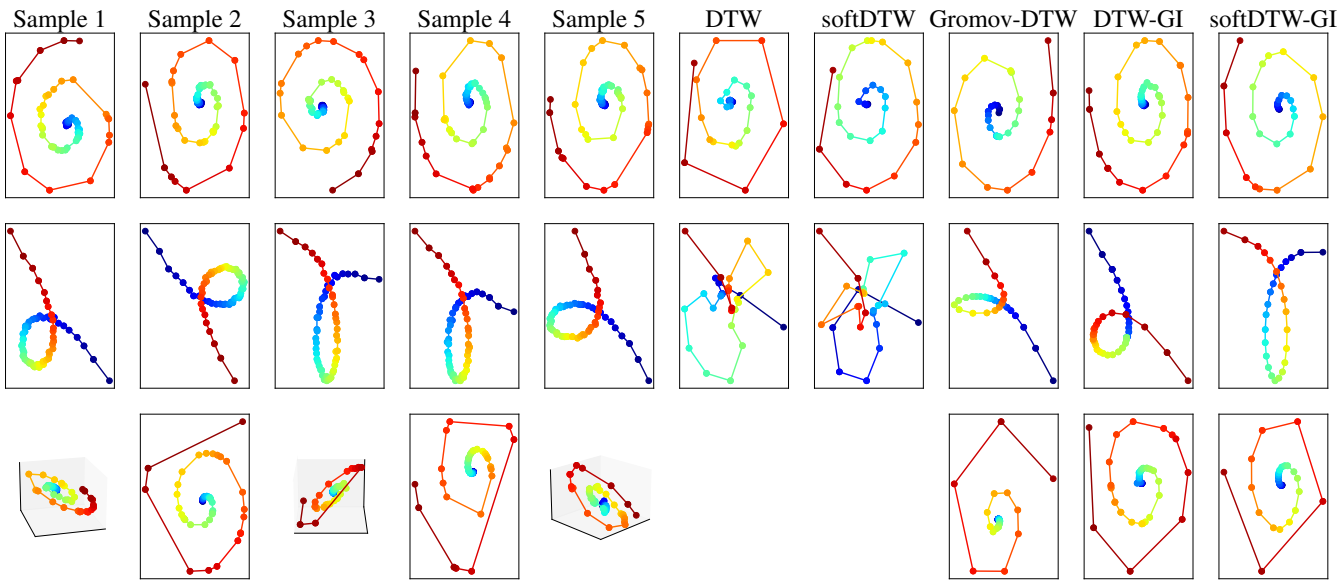

Fig. 5. Barycenter computation using (i) DTW and softDTW baseline approaches, (ii) their rotation-invariant counterparts DTW-GI and softDTW-GI and (iii) the alternative Gromov-DTW method. Each row correspond to a different dataset, and the latter one contains both 2d and 3d trajectories, hence cannot be tackled by DTW nor softDTW. Trajectories are color-coded from blue (beginning of the series) to red (end of the series).

where a_d is the attention kernel:

$$a_d(\mathbf{y}, \mathbf{x}_i) = \frac{e^{-\lambda d(\mathbf{y}, \mathbf{x}_i)}}{\sum_j e^{-\lambda d(\mathbf{y}, \mathbf{x}_j)}} \tag{15}$$

with $\lambda > 0$. The prediction is based on the known timestamps for the time series of the training set and on transformations f_i that aim at capturing the latent transformation between training and test time series. The attention kernel gives more importance to time series that are close to the time series we want to forecast *w.r.t.* the notion of dissimilarity d. Note that for large values of λ, the softmax in Equation (15) converges to a hard max and the proposed approach corresponds to a nearest neighbor imputation.

1) Dataset and methodology: We use the *Human3.6M* dataset [36] which consists of 3.6 million video frames of human movements recorded in a controlled indoor motion capture setting. This dataset is composed of 7 actors performing 15 activities ("Walking", "Sitting" ...) twice. We are interested in forecasting the 3D positions of the subject joints evolving over time. We follow the same data partition as [37]: the training set has 5 subjects (S6, S7, S8, S9 and S11) and the remaining 2 subjects (S1 and S5) compose the test set. In our experiments, 1) we split the limit frames as follows: we keep the first $T' = 300$ timestamps to calculate the coefficient $a_d\left(\mathbf{y}_{\to T'}, \mathbf{x}^{(i)}_{\to T'}\right)$ and the transformations f_i 2) we find the hyperparameter λ which gives the best prediction for $t \in [T', T_0]$ (where $T_0 = 400$) 3) the remaining times $[T_0, T']$ are used for the test set. We set the last limit frame as $T' = 1100$ which corresponds to predicting $T' - T_0 = 700$ timestamps, that is predicting 14 seconds of motion given the initial 8 seconds. To emulate

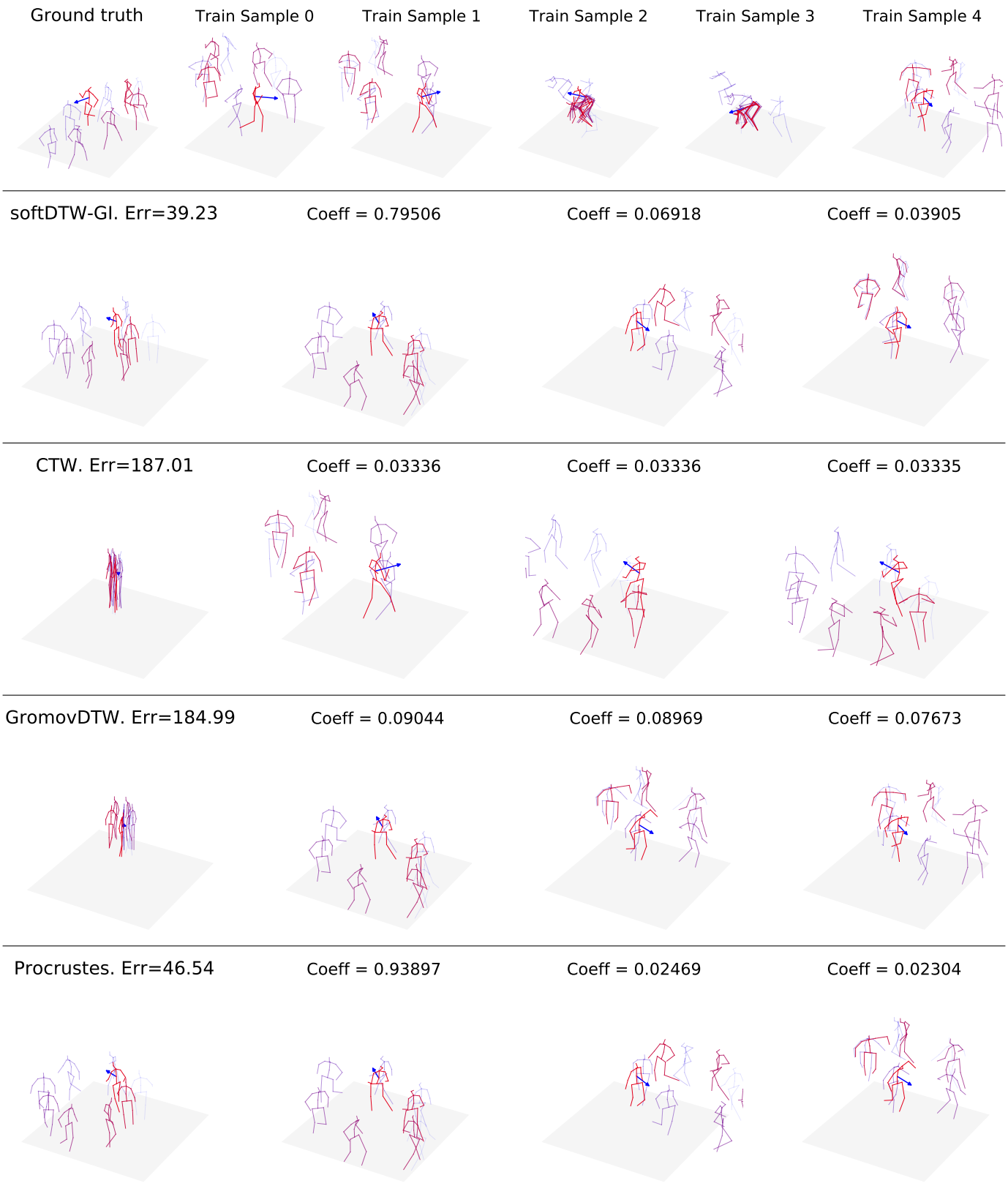

Fig. 6. Examples of the forecasted subseries. **(first row)** The first sample is the ground-truth $\hat{\mathbf{y}}_{T'\rightarrow}$ for the subject $S1$ and then training samples $\mathbf{x}^{(i)}_{T'\rightarrow}$ are depicted. **(from second to last row)** Predictions for the methods `softDTW-GI`, `CTW`, `GromovDTW` and `L2+Procrustes` (the other methods are given in 7). In the first column the prediction $\hat{\mathbf{y}}_{T'\rightarrow}$ is depicted along with the error $\|\hat{\mathbf{y}}_{T'\rightarrow} - \mathbf{y}_{T'\rightarrow}\|_2$. In the other columns, we illustrate the first 3 neighbors *w.r.t.* the method $\mathbf{x}^{(i)}_{\rightarrow T'}$ associated with their coefficients $a_d\left(\mathbf{y}_{\rightarrow T'}, \mathbf{x}^{(i)}_{\rightarrow T'}\right)$. For each movement an arrow indicates the orientation of the subject. The beginning of the movement is displayed in shaded blue while the end is displayed in bold red.

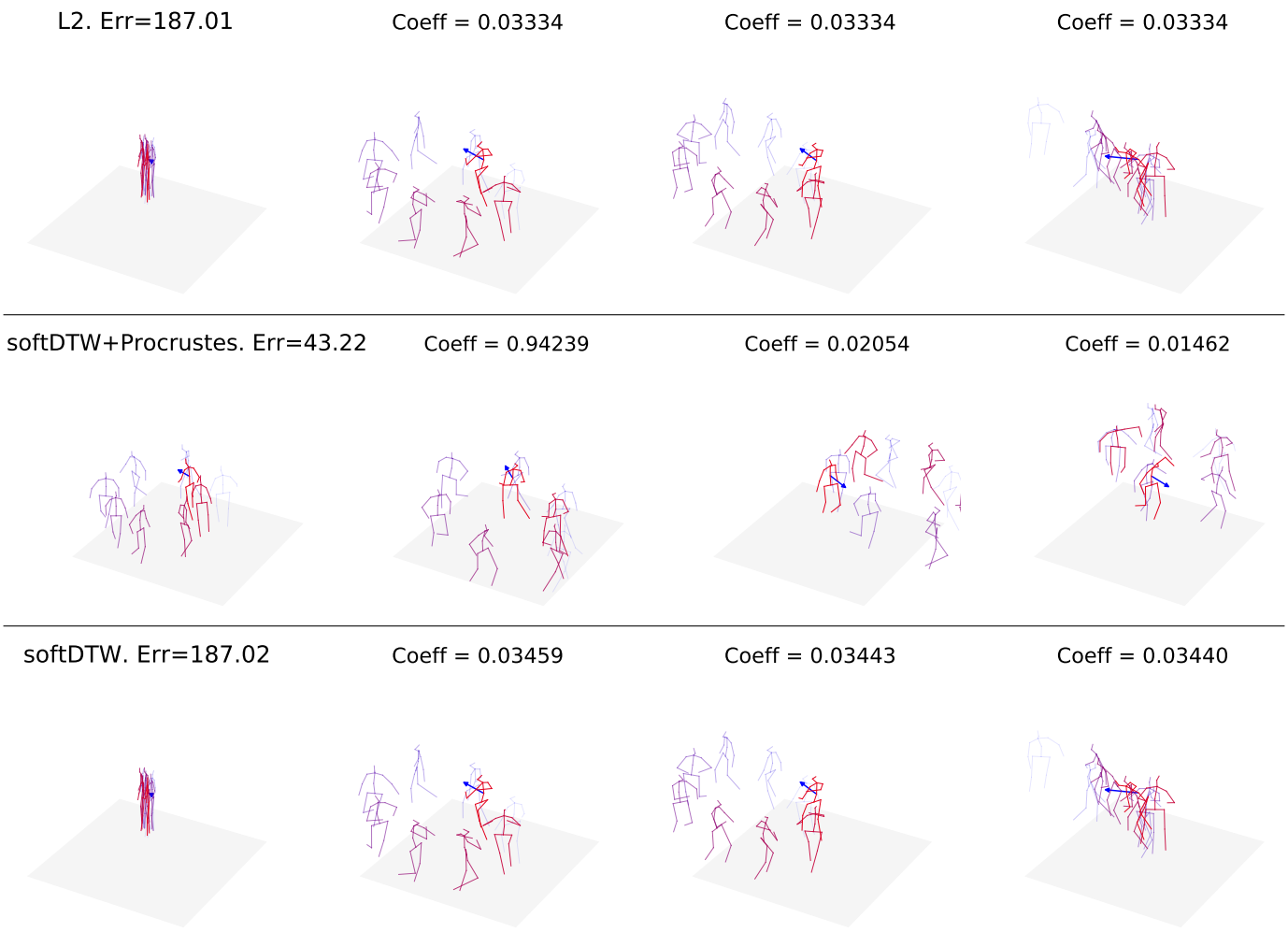

Fig. 7. Examples of the forecasted subseries for the methods *L2*, *softDTW* and *softDTW+Procrustes*

possible changes in signal acquisition (*e.g.* rotations of the camera), we randomly rotate the train subjects *w.r.t.* the z-axis. We consider the movements of type "Walking", "WalkDog" and "WalkTogether" for the training set and "Walking" for the test set. Top row of Figure 6 illustrates samples of movements $\mathbf{x}^{(i)}_{\rightarrow T'}$ resulting from this procedure and the resulting dataset is provided as supplementary material.

2) Competing structured prediction methods: Since the motions are in 3d, we look for global transformations of the features $f_i : \mathbb{R}^3 \rightarrow \mathbb{R}^3$. We use softDTW-GI as our similarity measure and its associated map f_i as described in Equation (7). We compare softDTW-GI to 7 baselines, that correspond to different pairs of time series similarity measure and feature space invariances. The first two baselines, denoted L2 and softDTW, do not encode any feature space invariance and are based on ℓ_2 and softDTW similarities respectively. We also consider a Procruste baseline [12] defined as:

$$d(\mathbf{y}_{\rightarrow T'}, \mathbf{x}^{(i)}_{\rightarrow T'}) = \min_{\mathbf{P}\in\mathbb{V}_{3,3}, \mathbf{b}\in\mathbb{R}^3} \|(\mathbf{x}^{(i)}_{\rightarrow T'}\mathbf{P}^T + \mathbf{b}) - \mathbf{y}_{\rightarrow T'}\|_2^2 \tag{16}$$

The corresponding transformation f_i is the affine map based on the optimal $\mathbf{P}^\star, \mathbf{b}^\star$ found by the previous problem. We denote this baseline L2+Procrustes. Another baseline is computed by first registering series using the Procrustes procedure defined above and then using the similarity measure $d(\mathbf{y}_{\rightarrow T'}, \mathbf{x}^{(i)}_{\rightarrow T'}) = DTW_\gamma(\mathbf{y}_{\rightarrow T'}, \mathbf{x}^{(i)}_{\rightarrow T'}\mathbf{P}^{\star T} + \mathbf{b}^\star)$. It is denoted softDTW+Procrustes. Finally, we also compare with GromovDTW [20] and CTW [16]. Note that the methods L2, softDTW, GromovDTW and CTW do not provide a transformation of the features of $\mathbf{x}^{(i)}_{\rightarrow T'}$ **onto** those of $\mathbf{y}_{\rightarrow T'}$ and, as such, we set $f_i = \mathrm{id}$ for all of these methods.

Method	Average test error
L2	183.11 +/- 3.90
softDTW [24]	183.12 +/- 3.90
CTW [16]	183.11 +/- 3.90
GromovDTW [20]	181.28 +/- 3.71
L2+Procrustes	46.33 +/- 0.21
softDTW+Procrustes	43.16 +/- 0.06
softDTW-GI (our)	**39.58 +/- 0.34**

TABLE I
AVERAGE ERROR ON TESTS SUBJECTS FOR THE TIME SERIES FORECASTING ON THE HUMAN3.6M DATASET

3) Results: Qualitative and quantitative results are provided in Figure 6, 7 and Table I respectively. We evaluate, for each test subject, the ℓ_2 reconstruction loss $\|\mathbf{y}_{T'\rightarrow} - \hat{\mathbf{y}}_{T'\rightarrow}\|_2$

between the ground truth time series and its prediction. Table I displays the average loss on the test subjects based on the best hyperparameter found using the timestamps $[T_0, T']$. Figure 6 and 7 present examples of reconstructed movements for the different methods on one test subject as well as the 3 highest coefficients a_d with the corresponding neighbors.

We observe from the quantitative study that softDTW, L2, CTW and GromovDTW lead to the worst reconstruction losses while L2+Procrustes, softDTW+Procrustes and softDTW-GI lead to the best ones. The results for the first four methods can be explained by the fact that none of them can use an explicit spatial transformation of the feature $f_i : \mathbb{R}^3 \to \mathbb{R}^3$ for the prediction and thus only a simple weighted average of the time-series $\mathbf{x}^{(i)}_{T'\to}$ is realized. This is illustrated for CTW, GromovDTW, softDTW and L2 in Figure 6 and 7, where we can see that the prediction tends to shrink. We can also see that softDTW+Procrustes is superior to a simple L2+Procrustes which highlights the importance of temporal realignment. More importantly, the performances of softDTW-GI is also better than softDTW+Procrustes and L2+Procrustes which shows that our joint realignment of time and space has an advantage over a two-step procedure such as softDTW+Procrustes which first finds the feature transformation and then aligns series in time.

Moreover, one can observe qualitatively that GromovDTW and CTW seem to uniformly average all the different motions to compensate for the lack of reprojection f_i. On the contrary, by capturing the possible spatial variability, L2+Procrustes and softDTW+Procrustes perform reasonably well qualitatively but the predicted movement is slightly less accurate than the one of softDTW-GI. This is due to the fact that L2+Procrustes or softDTW+Procrustes mainly chooses the movement corresponding to the first nearest neighbor ($a_d[1] = 0.94$) while softDTW-GI is able to average other dynamics ($a_d[1] = 0.79$). It is somehow a natural conclusion since the optimal transformations found by the Procrustes analysis supposes a trivial one-to-one correspondence of the timestamps (*i.e.* $\mathbf{y}_{\to T'}(t)$ corresponds to $\mathbf{x}^{(i)}_{\to T'}(t)$ at **the same time** t) and do not consider the temporal shifts between them. In this way, the method L2+Procrustes leads to unrealistic transformations when the dynamics of movements are not the same. Note that the two-step procedure softDTW+Procrustes is only slightly more precise as the feature realignment is independent of the temporal realignment since both are not optimized jointly. On the opposite, softDTW-GI method leads to the best qualitative results, highlighting the benefits of our approach over methods that whether discard the temporal variability of the movements (L2+Procrustes) or its spatial variability (softDTW).

E. Cover song identification

Cover song identification is the task of retrieving, for a given query song, its covers (*i.e.* different versions of the same song) in a training set. State-of-the-art methods either rely on anchor matches in the songs and/or on temporal alignments. In most related works, chroma or harmonic pitch class profile (HPCP) features are usually chosen, as they capture harmonic characteristics of the songs at stake [38].

For this experiment, we use the covers80 dataset [39] that consists in 80 cover pairs of pop songs and we evaluate the performance in terms of recall. Since the selection of features is not our main focus, we choose to extract chroma energy normalized statistics (CENS, [40]) over half a second windows. We compare variants of our method to a baseline that consists in a DTW alignment between songs transposed to the same key using the Optimal Transposition Index (OTI, [41]). This OTI computes a transposition based on average energy in each semitone band.

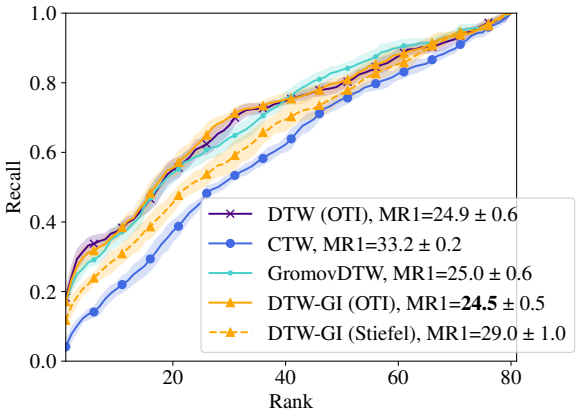

Fig. 8. Cover song identification using the covers80 dataset. Methods are compared in terms of recall and results are averaged over 10 train / test set draws. For each method, the shaded area corresponds to one standard deviation around the mean value.

Figure 8 presents recall scores for compared methods as well as the mean rank of the first correctly identified cover (MR1) that is a standard evaluation metric for the task (used in MIREX[1] for example). Presented results show that DTW-GI with feature space optimization on the Stiefel manifold (dashed line) is outperformed by the baseline relying on OTI. This is because a very common transformation in this setting is when cover songs are played in different keys, which is captured by the OTI transposition strategy. Interestingly enough, the flexibility of our DTW-GI framework allows us to use this OTI strategy. Using the BCD optimization scheme presented in Section III-B, we are able to compute the optimal transposition index along the alignment path (instead of computing it on averaged features) at each iteration of the algorithm. This leads to a significant improvement of the performance and illustrates both the versatility of our method and the importance of performing joint feature space transformation and temporal alignment. Note also that CTW, as it does not allow to take such prior information about the form of the feature space registration into account, clearly underperforms in this case.

V. Conclusion and Perspectives

We propose in this paper a novel similarity measure that can compare time series across different spaces in order to tackle both temporal and feature space invariances. This work extends the well-known Dynamic Time Warping algorithm to deal with time series from different spaces thanks to the introduction

[1]https://www.music-ir.org/mirex/wiki/2019:Audio_Cover_Song_Identification

of a joint optimization over temporal alignements and space transformations. In addition, we provide a formulation for the computation of the barycenter of a set of times series under our new geometry, which is, to the best of our knowledge, the first barycenter formulation for a set of heterogeneous time series. Another important special case of our approach allows for performing temporal alignment of time series with invariance to rotations in the feature space.

We illustrate our approach on several datasets. First, we use simulated time series to study the computational complexity of our approach and illustrate invariance to rotations. Then, we apply our approach on two real-life datasets for human motion prediction and cover song identification where invariant similarity measures are shown to improve performance.

Extension of this work will consider scenarios where features of the series do not lie in a Euclidean space, which would allow covering the case of structured data such as graphs evolving over time, for example. Future works also include the use of our methods in more elaborated models where, following ideas from [42, 43], softDTW-GI could be used as a feature extractor in neural networks. It could also serve as a loss to train heterogeneous time series forecasting models [35, 24].

REFERENCES

[1] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[2] C.-Y. Chang, D.-A. Huang, Y. Sui, L. Fei-Fei, and J. C. Niebles, "D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3546–3555.

[3] V. Wegner Maus, G. Câmara, M. Appel, and E. Pebesma, "dtwsat: Time-weighted dynamic time warping for satellite image time series analysis in R," *Journal of Statistical Software*, vol. 88, no. 5, pp. 1–31, 2019.

[4] S.-F. Huang and H.-P. Lu, "Classification of temporal data using dynamic time warping and compressed learning," *Biomedical Signal Processing and Control*, vol. 57, p. 101781, 2020.

[5] H. Janati, M. Cuturi, and A. Gramfort, "Spatio-temporal alignments: Optimal transport through space and time," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 1695–1704.

[6] S. Ben-David, T. Lu, T. Luu, and D. Pál, "Impossibility theorems for domain adaptation," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 129–136.

[7] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[8] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[9] P. Battaglia, J. B. C. Hamrick, V. Bapst, A. Sanchez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. E. Dahl, A. Vaswani, K. Allen, C. Nash, V. J. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, "Relational inductive biases, deep learning, and graph networks," *arXiv*, 2018.

[10] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng, "Measuring invariances in deep networks," in *Neural Information Processing Systems*, 2009, pp. 646–654.

[11] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and Vision Computing*, vol. 10, no. 3, pp. 145 – 155, 1992.

[12] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 53, no. 2, pp. 285–321, 1991.

[13] S. Cohen and L. Guibasm, "The Earth mover's distance under transformation sets," in *IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1076–1083.

[14] D. Alvarez-Melis, S. Jegelka, and T. S. Jaakkola, "Towards optimal transport with global invariances," in *International Conference on Artificial Intelligence and Statistics*, 2019.

[15] D. Alvarez-Melis, Y. Mroueh, and T. Jaakkola, "Unsupervised hierarchy matching with optimal transport over hyperbolic spaces," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1606–1617.

[16] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," in *Advances in neural information processing systems*, 2009, pp. 2286–2294.

[17] F. Zhou and F. De la Torre, "Generalized time warping for multi-modal alignment of human motion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1282–1289.

[18] G. Trigeorgis, M. A. Nicolaou, S. Zafeiriou, and B. W. Schuller, "Deep canonical time warping," in *EEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5110–5118.

[19] H. Deng, W. Chen, Q. Shen, A. J. Ma, P. C. Yuen, and G. Feng, "Invariant subspace learning for time series data based on dynamic time warping distance," *Pattern Recognition*, vol. 102, p. 107210, 2020.

[20] S. Cohen, G. Luise, A. Terenin, B. Amos, and M. P. Deisenroth, "Aligning time series on incomparable spaces," in *International Conference on Artificial Intelligence and Statistics*, vol. 130, 2021, pp. 1036–1044.

[21] G. A. Ten Holt, M. J. Reinders, and E. Hendriks, "Multi-dimensional dynamic time warping for gesture recognition," in *Annual conference of the Advanced School for Computing and Imaging*, vol. 300, 2007, p. 1.

[22] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams," *Neurocomputing*, vol. 73, no. 1-3, pp. 366–380, 2009.

[23] Z. Zhang, R. Tavenard, A. Bailly, X. Tang, P. Tang, and T. Corpetti, "Dynamic time warping under limited warping path length," *Information Sciences*, vol. 393, pp. 91–107, 2017.

[24] M. Cuturi and M. Blondel, "Soft-DTW: a differentiable loss function for time-series," in *International Conference on Machine Learning*, 2017, pp. 894–903.

[25] M. Cuturi, J.-P. Vert, O. Birkenes, and T. Matsui, "A kernel for time series based on global alignments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2007, pp. II–413.

[26] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Mathematical Programming*, vol. 142, no. 1-2, pp. 397–434, 2013.

[27] M. Jaggi, "Revisiting frank-wolfe: Projection-free sparse convex optimization." in *International Conference on Machine Learning*, 2013.

[28] J. Lawrence, J. Bernal, and C. Witzgall, "A purely algebraic justification of the kabsch-umeyama algorithm," *Journal of Research of the National Institute of Standards and Technology*, vol. 124, pp. 1–6, 2019.

[29] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, no. 3, pp. 678 – 693, 2011.

[30] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar *et al.*, "Tslearn, A Machine Learning Toolkit for Time Series Data," *J. Mach. Learn. Res.*, vol. 21, no. 118, pp. 1–6, 2020.

[31] M. Kochurov, S. Kozlukov, R. Karimov, and V. Yanush, "Geoopt: Adaptive riemannian optimization in PyTorch," 2019, https://github.com/geoopt/geoopt.

[32] G. Becigneul and O.-E. Ganea, "Riemannian adaptive optimization methods," in *Proceedings of the International Conference on Learning Representations*, 2019.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035.

[34] J. Kruskal and M. Wish, *Multidimensional Scaling*. Sage Publications, 1978.

[35] V. Le Guen and N. Thome, "Shape and time distortion loss for training deep time series forecasting models supplementary material," in *Neural Information Processing Systems*, 2019.

[36] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[37] H. Coskun, F. Achilles, R. DiPietro, N. Navab, and F. Tombari, "Long short-term memory kalman filters: Recurrent neural estimators for pose regularization," in *International Conference on Computer Vision*, Oct 2017.

[38] H. Heo, H. J. Kim, W. S. Kim, and K. Lee, "Cover song identification with metric learning using distance as a feature." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2017.

[39] D. P. Ellis and C. V. Cotton, "The 2007 labrosa cover song detection system," 2007.

[40] M. Müller, F. Kurth, and M. Clausen, "Audio matching via chroma-based statistical features." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2005.

[41] J. Serra, E. Gómez, and P. Herrera, "Transposing chroma representations to a common key," in *Proceeding sof the IEEE Conference on The Use of Symbols to Represent Music and Multimedia Objects*, 2008, pp. 45–48.

[42] X. Cai, T. Xu, J. Yi, J. Huang, and S. Rajasekaran, "Dtwnet: a dynamic time warping network," in *Neural Information Processing Systems*, 2019, pp. 11 636–11 646.

[43] B. Iwana, V. Frinken, and S. Uchida, "DTW-NN: a novel neural network for time series recognition using dynamic alignment between inputs and weights," *Knowledge-Based Systems*, vol. 188, 1 2020.