Deviation-based Dynamic Time Warping for Clustering Human Sleep

Chiying Wang[1], Sergio A. Alvarez[2], Carolina Ruiz[1] and Majaz Moonis[3]
[1]*Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA 01609 U.S.A.*
[2]*Department of Computer Science, Boston College, Chestnut Hill, MA 02467 U.S.A.*
[3]*Department of Neurology, U. of Massachusetts Medical School, Worcester, MA 01655 U.S.A.*

Keywords: Dynamic Time Warping, Deviation, Human Sleep, Clustering.

Abstract: In this paper, we propose two versions of a modified dynamic time warping approach for comparing discrete time series. This approach is motivated by the observation that the distribution of dynamic time warping paths between pairs of human sleep time series is concentrated around the path of constant slope. Both versions use a penalty term for the deviation between the warping path and the path of constant slope for a given pair of time series. In the first version, global weighted dynamic time warping, the penalty term is added as a post-processing step after a standard dynamic time warping computation, yielding a modified similarity metric that can be used for time series clustering. The second version, stepwise deviation-based dynamic time warping, incorporates the penalty term into the dynamic programming optimization itself, yielding modified optimal warping paths, together with a similarity metric. Clustering experiments over synthetic data, as well as over human sleep data, show that the proposed methods yield significantly improved accuracy and generative log likelihood as compared with standard dynamic time warping.

1 INTRODUCTION

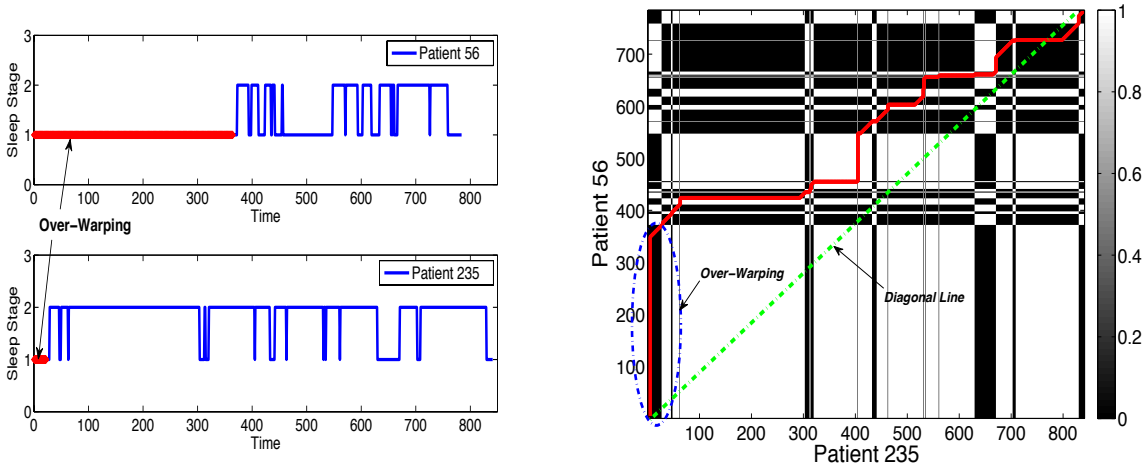
Human sleep patterns are closely associated with overall health and quality of life, making the scientific study of sleep an important pursuit. Sleep stage transitions (Kishi et al., 2008) and bout durations (Chu-Shore et al., 2010) are essential indicators in characterizing the structure of sleep. Typical patterns of human sleep have been found (Bianchi et al., 2010), yet sleep microstructure varies across individuals, being affected by age, circadian rhythms (Dijk and Lockley, 2002), and other factors.

A substantial challenge in modeling the dynamics of sleep is the scarcity of key dynamical events such as stage transitions within sleep sequences. This scarcity yields small samples over which dynamical models are to be trained, leading to high uncertainty in parameter estimates. An approach known as dynamical modeling-clustering (CDMC) was proposed (Alvarez and Ruiz, 2013) to address this challenge. CDMC reduces model variance through selective aggregation of instances during a clustering phase, so that models are learned over collections of dynamically similar instances rather than individual instances. The technique of initialization using clustering by Dynamic Time Warping (DTW) similarity (Oates et al., 2001) yields good convergence properties for CDMC (Wang et al., 2014).

Despite promising results, standard DTW as a similarity measure for unsupervised clustering of time series suffers from certain problems. One of these is the over-warping problem shown in Fig. 1. Over-warping refers to unnatural alignment of dissimilar segments in two time series. In Fig. 1, a subsequence of length over 300 in one patient is matched by dynamic time warping to a subsequence of length less than 10 in another. The result is unacceptable, yet the standard dynamic time warping distance between the two segments is zero. Explicitly penalized DTW has been developed to address over-warping. For example, (Clifford et al., 2009) proposes variable penalty DTW, which reduces nondiagonal moves during alignment. However, this approach is heavily dependent on a user-defined penalty function and thus difficult to apply in practice.

An additional concern with DTW is time complexity. Variants of DTW have been proposed that focus on improving efficiency by globally constraining the warping path to a predefined geometric region such as the Sakoe-Chiba band (Sakoe and Chiba, 1978) or the Itakura parallelogram (Itakura, 1975). However, the use of global constraints alone can lead

Wang, C., Alvarez, S., Ruiz, C. and Moonis, M.
Deviation-based Dynamic Time Warping for Clustering Human Sleep.
DOI: 10.5220/0005729200880095
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 4: BIOSIGNALS*, pages 88-95
ISBN: 978-989-758-170-0



(a) Over-Warping Occurrence in Time Series.

(b) Over-Warping Path in Dynamic Time Warping.

Figure 1: Over-warping of sleep stage sequences using dynamic time warping. (Left) Use of standard dynamic time warping inappropriately matches dissimilar segments (a long one versus a short one) in two sequences. (Right) The same over-warping problem described in terms of warping search area. The area circled by the dashed line indicates a large deviation of the standard warping path from the diagonal path of constant slope. The bar graph on the right indicates local cost measure and the background in the right figure shows the local cost marix of two discrete time series (patient 56 and patient 235) in dynamic time warping computation (see section 2.1).

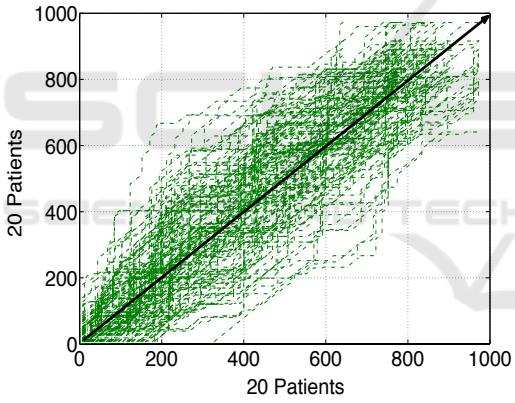


Figure 2: Optimal time warping paths (dashed lines) between 20 pairs of human sleep recordings. DTW optimal paths are usually close to the diagonal line from bottom-left to top-right in the warping space. The varying boundary of the distribution suggests the desirability of adaptively identifying warping areas locally instead of using a predefined global constraint.

to the over-warping problem described above. Some researchers (Ratanamahatana and Keogh, 2004) have argued that the effect of warping band width on the quality of the results is greatly domain dependent and that a narrow band might be valuable. The distribution of optimal warping paths between pairs of sleep time series in Fig. 2 (from the present authors' own work) likewise suggests that the use of local search constraints would be desirable.

Main Contributions of the Present Paper

1. We propose two novel DTW variants, global weighted dynamic time warping (gwDTW) and stepwise deviated dynamic time warping (sdDTW), that penalize deviations of the warping path from the path of constant slope. This overcomes the over-warping issue in Fig. 1, while retaining the efficiency advantages of approaches based on global constraints such as the Sakoe-Chiba band (Sakoe and Chiba, 1978) and Itakura parallelogram (Itakura, 1975), and without relying on domain dependent specifics as in variable penalty DTW (Clifford et al., 2009).
2. We apply the proposed modified DTW for clustering initialization within the combined dynamical modeling-clustering (CDMC) framework (Alvarez and Ruiz, 2013) over human sleep time series, and show that this approach better captures the dynamics of human sleep.

Organization of the Paper. Section 2 reviews standard DTW, and describes the proposed deviation-based dynamic time warping approach and its application to time-series clustering. Section 3 presents experimental results and analysis on time series clustering using deviation-based dynamic time warping. Section 4 describes conclusions and future work.

2 METHODS

We review standard dynamic time warping in section 2.1, as that technique will serve as a baseline. The proposed deviation-based dynamic time warping approach is described in section 2.2.

2.1 Dynamic Time Warping (DTW)

Dynamic time warping (DTW) is a classical dynamic programming algorithm for measuring the similarity of two time series (e.g., (Müller, 2007)). It performs an optimal alignment between two time series by nonlinearly warping their time dimensions. DTW has been applied to speech recognition (e.g., (Sakoe and Chiba, 1978) and (Itakura, 1975)), time series classification (Jeong et al., 2011), and unsupervised time series clustering (Oates et al., 2001).

The following are the essentials of standard DTW, as described in (Müller, 2007).

We consider two time sequences $X = (x_1, x_2, \ldots, x_N)$ of length $N \in \mathbb{N}$ and $Y = (y_1, y_2, \ldots, y_M)$ of length $M \in \mathbb{N}$, with individual values x_i, y_j in some feature space $\mathcal{F}$.

A **local cost measure** is a function

$$c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0} \tag{1}$$

The value of $c(x_i, y_j)$ is small if x_i, y_j are close to each other, and otherwise not. For discrete time series, one can use a cost matrix to define the values $c(x, y)$ for all pairs of values x, y; the simplest possibility is to use the identity matrix, that is, to let $c(x, y) = 0$ if $x = y$, otherwise $c(x, y) = 1$.

A **warping path** between X and Y is a sequence

$$p = (p_1, p_2, \ldots, p_L) \tag{2}$$

where $p_l = (n_l, m_l) \in [1 : N] \times [1 : M]$ for $l \in [1 : L]$ (the diagonal line from bottom-left to top-right in Fig. 1(b)), and $\max\{N, M\} \leq L \leq N + M$. It must satisfy the following three conditions:

- *Boundary condition*: $p_1 = (1, 1)$ and $p_L = (N, M)$ are the start and end points respectively.
- *Monotonicity condition*: horizontal and vertical components increase monotonically: $n_1 \leq n_2 \leq \ldots n_L$ and $m_1 \leq m_2 \leq \ldots m_L$.
- *Step size condition*: for each $l < L$, the difference $p_{l+1} - p_l$ is one of $(1, 0), (0, 1), (1, 1)$.

The **total cost** of a warping path p between X and Y is

$$\Phi_p(X, Y) = \sum_{l=1}^{L} c(x_{n_l}, y_{m_l}) \tag{3}$$

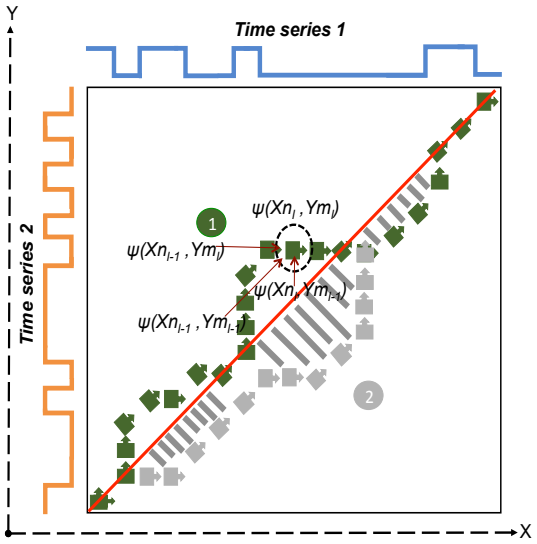


Figure 3: Deviation (e.g., gray shaded area) of warping path (squares with directional arrows) from path of constant slope (solid red line). Given two warping paths (1 and 2), the path with smaller deviation (the green one) is better.

An **optimal warping path** p^* is one having minimum total cost $\Phi_{p^*}(X, Y)$ among all warping paths from p_1 to p_L. $\Phi_{p^*}(X, Y)$ is referred to as the *DTW distance* between sequences X and Y.

2.2 Proposed Approach

2.2.1 Deviation Measure

We address the standard DTW concerns of overwarping and time complexity described in the Introduction, by penalizing nondiagonal moves in the search for an optimal warping path. This is done by using the measure of deviation discussed below.

Deviation refers to the area Δ_p (shaded areas in Fig. 3) bounded by the warping path p between two time series and the diagonal path of constant slope.

Deviation is computed by the procedure described in Algorithm 1. The following are the main steps:

- Initialize the deviation Δ_p to be 0. (step 1)
- Calculate the slope k and the intercept b of the diagonal path defined as $y = k * x + b$. (step 2-3)
- Repeat until the end point p_L is reached. (step 5)
  - Add absolute vertical distance between current point p_i and diagonal to the deviation Δ_p, if p_i and p_{i-1} differ vertically. (step 6-7)
  - update the counter, i. (step 8)
- Return the deviation Δ_p (step 9)

Algorithm 1: Deviation Calculation.

Input: A warping path $p = \{p_1, p_2, \ldots, p_L\}$, each $p_i = (n_i, m_i)$; L: total number of points in path p.
Output: The deviation Δ_p of p from the diagonal line through p_1 and p_L in warping space.
ComputeDeviation(p)

1. $\Delta_p = 0$
2. $k = (M-1)/(N-1)$
3. $b = M - k*N$
4. i = 2
5. **While** $(p_i \neq p_L)$
6. **if** $(m_i \neq m_{i-1})$
7. $\Delta_p = \Delta_p + |m_i - (k \cdot n_i + b)|$
8. $i = i+1$
9. **return** Δ_p

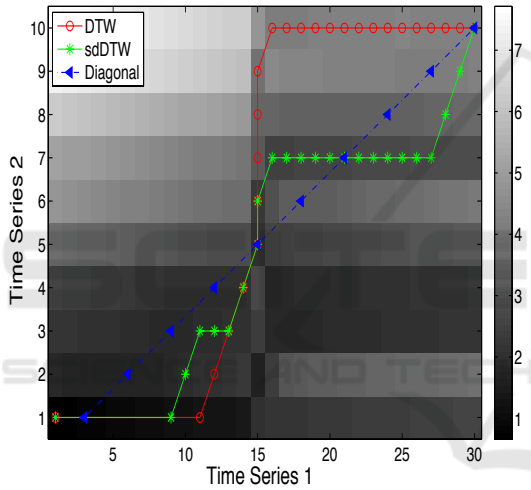


Figure 4: Optimal warping paths for standard DTW and stepwise deviation-based DTW (sdDTW). Standard DTW allows large deviations in searching for a path of minimum total cost. sdDTW aligns time series closer to the diagonal line. Background shading indicates local cost in sdDTW, which increases with distance to the diagonal.

2.2.2 Deviation-based Dynamic Time Warping

Global Weighted Dynamic Time Warping: The global weighted dynamic time warping ($gwDTW$) distance between sequences X and Y is defined as

$$gwDTW(X,Y) = \lambda_{gw} \cdot \Phi_{p^*}(X,Y) + (1-\lambda_{gw}) \cdot \sqrt{\Delta_{p^*(X,Y)}} \quad (4)$$

where $\Phi_{p^*}(X,Y)$ is standard DTW distance, and the deviation $\Delta_{p^*(X,Y)}$ (Algorithm 1) is added as a post-processing penalty to standard DTW. $p^*(X,Y)$ is the optimal path between X,Y in standard DTW. λ_{gw} controls the balance between standard DTW cost and the deviation $\Delta_{p^*(X,Y)}$. λ_{gw} ranges from 0 to 1. The square root of the deviation $\Delta_{p^*(X,Y)}$ is used because it scales linearly with Euclidean distance.

Stepwise Deviation-based Dynamic Time Warping: The stepwise deviation-based dynamic time warping ($sdDTW$) distance between sequences X and Y is defined as

$$sdDTW(X,Y) = \Psi_{p^*}(X,Y) \quad (5)$$

where $\Psi_{p^*}(X,Y)$ is the minimum total cost of a warping path p^* in Eq. 3 obtained by replacing the local cost measure in Eq. 1 by the modified measure $\varphi(x,y)$ in Eq. 6. Thus, sdDTW optimal warping paths are minimizers of a different cost measure than standard DTW paths.

$$\varphi(x,y) = \lambda_{sd} \cdot c(x,y) + (1-\lambda_{sd}) \cdot \sqrt{\Delta(x,y)} \quad (6)$$

where $c(x,y)$ denotes the original local cost measure in equation 1. $\Delta(x,y)$ is the deviation of a position (x,y) relative to the diagonal path of constant slope for sequences X and Y. See section 2.2.1. λ_{sd} is a parameter that determines the relative weights of the standard local cost measure and the deviation.

We use dynamic programming as in standard DTW to compute the sdDTW optimal warping path, based on the **modified accumulated cost matrix**

$$\bar{D}(n,m) = \Psi(X(1:n), Y(1:m)) \quad (7)$$

where $X(1:n) = (x_1, \ldots, x_n)$ and $Y(1:m) = (y_1, \ldots, y_m)$. $n \in [1:N]$ and $m \in [1:M]$. That is $X(1:n)$ and $Y(1:m)$ are subsequences of X and Y. The procedure is as follows:

- Initially, $\bar{D}(n,1) = \sum_{k=1}^{n} \varphi(x_k, y_1)$ and $\bar{D}(1,m) = \sum_{k=1}^{m} \varphi(x_1, y_k)$.
- Iteratively, take the minimum accumulated cost from three immediately adjacent directions: $\bar{D}(n-1,m) + \varphi(x_n, y_m)$, $\bar{D}(n,m-1) + \varphi(x_n, y_m)$, $\bar{D}(n-1,m-1) + \varphi(x_n, y_m)$.
- Until the final position (N,M) is reached. $\bar{D}(N,M)$ is the optimal dynamic time warping distance with respect to stepwise deviation-based dynamic time warping.

As illustrated in Fig. 4, optimal warping paths for the sdDTW distance metric exhibit less over-warping than the corresponding standard DTW paths.

2.2.3 Deviation-based Dynamic Time Warping Clustering

Deviation-based dynamic time warping clustering (dDTWC) performs unsupervised agglomerative hierarchical clustering of time series using the deviation-based DTW approaches in section 2.2.2 to calculate

distances. The proposed approach is described in pseudocode in Algorithm 2. The main steps are:

- Initially each time series instance X is in its own cluster (steps 1-2).
- Repeat until only k clusters remain (steps 3-6):
  - Merge the closest clusters, C and C'; the distance between two instances (X and Y) is defined by **gwDTW** or **sdDTW** in section 2.2.2; the distance between two clusters is the average distance between pairs of instances.
- Return clustering of dataset in k clusters (step 7).

3 EXPERIMENTAL EVALUATION

Deviation-based DTW clustering as described in section 2.2.3 was compared with clustering using the standard DTW distance metric. For all dynamic time warping computations, the local cost measure in Eq. 1 was defined as $c(x,y) = 1$ if the elements x and y are different, otherwise $c(x,y) = 0$. The weight values λ_{gw} and λ_{sd} in Eqs. 4 and 5, respectively, were determined empirically in order to maximize mean accuracy over a sample of labeled synthetic data generated as in section 3.2 (but separate from the synthetic data sample used for performance evaluation in section 3.2.3): λ_{sd} was set to 0.67. λ_{gw} was set to 0.83. All experiments were performed in MATLAB® (*The MathWorks*, 2015).

Two sets of experiments were carried out, corresponding to synthetic data and human sleep data, respectively. Details specific to each of these are described in sections 3.2.2 and 3.3.2 below.

3.1 Statistical Significance

Pairwise comparisons of median classification accuracy values (see section 3.2.2) of gwDTW and sdDTW clustering against the accuracy of standard DTW clustering (for synthetic data) and of negative log likelihood values of gwDTW and sdDTW clustering against that of standard DTW clustering (for human sleep data) were carried out by a non-parametric two-sided Wilcoxon rank sum test, since a Lilliefors normality test rejected normality at the $p < 0.05$ significance level in each case. A Bonferroni correction was performed jointly on the accuracy and log likelihood Wilcoxon p-values to ensure a familywise error rate less than 0.05.

3.2 Synthetic Markov Mixture Data

3.2.1 Dataset Generation

A synthetic dataset of discrete sequences was generated as in (Alvarez and Ruiz, 2013), from two distinct Markov models, each with two states. The two models differ in their transition probability matrices. Self-transition probabilities of 0.6 in one model and 0.8 in the other were selected. The probabilities of transitioning between states were 0.4 and 0.2, respectively. One of the two models is selected randomly and used to generate a sequence of the desired length, L. This process is iterated until a predetermined number of sequences, N, is obtained. The present paper uses the values $N = 100$ and $L = 300$ in all trials.

3.2.2 Experimental Procedure

Clustering Classification Accuracy. Supervised classification via clustering was performed with the generating model label as the classification target. Each cluster was associated with the class c that occurs most frequently among its members. The evaluation metric was classification accuracy, equal to the fraction of labeled instances (X,c) that are assigned to a cluster in which c is the majority class. Statistical hypothesis testing was performed using a Wilcoxon rank sum test to compare median accuracies as described in section 3.1. Experimental procedure was as in the following pseudocode:

Experimental Procedure, Synthetic Data Classification:

```
begin
  for i := 1 to TrialNum
    SD = generateSyntheticDataset(N, L);
    Accuracy(1, i) = evaluateBy DTW(SD);
    Accuracy(2, i) = evaluateBy gwDTW(λ_gw, SD);
    Accuracy(3, i) = evaluateBy sdDTW(λ_sd, SD);
  end
  Perform Wilcoxon rank sum test over Accuracy.
end
```

Note:

- *generateSyntheticDataset* followed the description in section 3.2.
- *evaluateBy***DTW** refers to the clustering procedure in section 2.2.3 and clustering evaluation by classification accuracy (see above); likewise for *evaluateBy***gwDTW** and *evaluateBy***sdDTW**.
- The total number of sequences, N, was set to 100.
- The length of a sequence, L, was set to be 300.
- The number of trials, $TrialNum$, was set to 100.

Algorithm 2: Deviation-Based DTW Clustering (dDTWC).

Input: An unlabeled time series dataset D = $\{X \mid X$ is a time series$\}$; a positive integer, k, the desired number of clusters; a predefined local cost measure $c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0}$ where $\mathcal{F}$ is the feature space in which the time series in D take their values. **dDTW** denotes the total cost measure associated with c, which is defined by Eq. 4 in the case of gwDTW and by Eq. 5 in the case of sdDTW.
Output: A partition of D into k clusters
dDTWC(D, k, d)

1. for each i, let $C_i =$ a cluster that contains only the i-th time series in D
2. $s =$ the number of time series in D (initial number of clusters)
3. **while** $s > k$
4. $(i^*, j^*) = \arg \min_{i,j \in \{1,\ldots,s\}} \bar{c}(C_i, C_j)$ (where $\bar{c}$ is mean cost for instance pairs in the two clusters)
   $= \arg \min_{i,j \in \{1,\ldots,s\}} \left\{ \frac{\sum_{X \in C_i, Y \in C_j} \mathbf{dDTW}(X,Y,c)}{|C_i| \cdot |C_j|} \right\}$
5. Merge C_{i^*} and C_{j^*} to reduce the number of clusters to $s-1$
6. $s = s - 1$
7. **return** $\{C_1, \ldots, C_k\}$

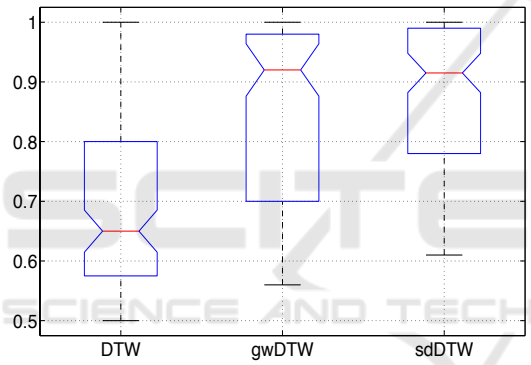


Figure 5: Clustering accuracies using standard DTW, gwDTW, and sdDTW as similarity measures over hidden Markov mixture data. Non-overlapping notches indicate significant difference in medians ($p < 0.05$). gwDTW and sdDTW are significantly more accurate than standard DTW.

3.2.3 Synthetic Data Results

This section evaluates performance of clustering over synthetic data using globally weighted DTW (gwDTW) (Eq. 4) or stepwise deviation-based DTW (sdDTW) as the similarity measure, as compared with standard DTW similarity.

Clustering accuracies over the synthetic dataset appear in Fig. 5. Both gwDTW and sdDTW perform significantly better than standard DTW, proving the benefit of incorporating deviation into the DTW computation for clustering of synthetic time series data. Median accuracies appear in Table 1.

Table 1: Median accuracies of clusterings based on DTW, gwDTW, and sdDTW. Asterisks denote Bonferroni-corrected statistical significance of differences with standard DTW in Wilcoxon rank sum test ($p < 0.05$).

DTW	gwDTW	sdDTW
0.65	0.92*	0.91*

3.3 Human Sleep Data

3.3.1 Datasets

A collection of 244 fully anonymized human polysomnographic recordings was extracted from polysomnographic overnight sleep studies performed in the Sleep Clinic at Day Kimball Hospital in Putnam, Connecticut, USA. Each polysomnographic recording is split into 30-second epochs. Lab technicians staged each 30-second epoch into one of the sleep stages Wake, stage 1, stage 2, stage 3, and REM (Rapid Eye Movement). Three versions of the human sleep dataset are considered, depending on whether these stage labels are grouped in some way:

- (W5) uses the five standard stage labels Wake, 1, 2, 3, REM.
- (WNR) uses the three stage labels Wake, NREM (stages 1, 2, and 3), REM.
- (WDL) uses the three stage labels Wake, Deep (stage 3), Light (stages 1,2,REM).

3.3.2 Experimental Procedure

Unsupervised clustering was performed over the human sleep datasets described in section 3.3.1.

The collective dynamic modeling clustering algorithm (Alvarez and Ruiz, 2013) was used for clustering, with two-state hidden semi-Markov chain models as the dynamical models. Initial cluster labels were computed by deviation-based DTW clustering as described in Algorithm 2, with either gwDTW (Eq. 4) or sdDTW (Eq. 5) as the distance metric. Clustering driven by the standard DTW distance metric was used as a basis for comparison.

Generative negative log likelihood was used to measure the quality of model fit for unsupervised clustering. Given a hidden semi-Markov model, M, built over a group of sequences such as human sleep sequences, the generative negative log-likelihood $-\log(P(s|M))$ of a sequence, s, is a measure of the probability that the sequence, s, would be produced by the model, M. Lower negative log-likelihood values (higher generative probabilities) imply a better model fit. The goal of clustering was to minimize the generative negative log likelihood. Comparison of median negative log likelihoods for different models was measured by a Wilcoxon rank sum test as described in section 3.1.

Experimental Procedure, Sleep Data Clustering:

```
begin
  D1, D2, D3 = W5, WNR, WDL datasets (3.3.1)
  m1, m2, m3 = DTW, gwDTW, sdDTW
  for j = 1 to 3
    for k = 1 to 3
      (M, nlogll(Dj, mk, s1...s244)) = CDMC(Dj, mk)
    end
    Perform pairwise Wilcoxon rank sum tests on
      nlogll(Dj, m1...m3, s1...s244)
  end
end
```

Notes:

- The W5, WNR, WDL datasets are as in section 3.3.1.
- DTW refers to clustering using standard DTW as the similarity metric; gwDTW and sdDTW refer to the deviation-based clustering techniques described in section 2.2.3.
- $CDMC(Dj,mk)$ refers to CDMC clustering (Alvarez and Ruiz, 2013) with semi-Markov cluster models, using the given method, mk, for clustering initialization, and is assumed to return a set of dynamical models together with negative generative log likelihoods nlogll(Dj,mk,sl) for all input sequences, sl, $l = 1,\cdots 244$.

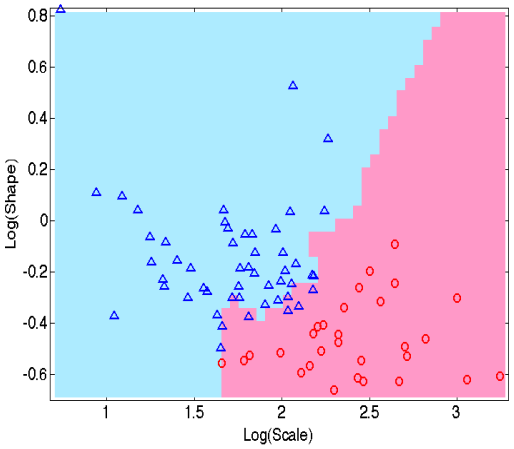


Figure 6: Visualization of clusters over human sleep dataset using gwDTW as similarity measure. Coordinates are Weibull shape and scale parameters for Wake stage. Red circles and blue triangles denote gwDTW clusters; background colors represent DTW clusters.

3.3.3 Human Sleep Data Results

Fig. 6 shows the two CDMC clusters (circles and triangles) with coordinates equal to the Weibull scale and shape parameters for the wake stage in the WNR dataset. gwDTW clusters better capture the boundary between the natural Weibull dynamical clusters in the human sleep dataset, as compared with standard DTW clusters.

Model fit was significantly better for both global weighted DTW (gwDTW) clustering and stepwise deviation-based DTW (sdDTW) clustering as compared with standard DTW-driven clustering, as shown in Table 2. This shows that deviation-based DTW is superior to standard DTW as a similarity metric for initialization of CDMC clustering over human sleep data, as well as for standalone clustering over synthetic data as shown in section 3.2.3.

Table 2: Median negative log likelihoods of gwDTW, sdDTW, and standard DTW clusterings over WNR, WDL, and W5 human sleep datasets in section 3.3.1. Asterisks indicate Bonferroni-corrected significance of differences with standard DTW in Wilcoxon rank sum test ($p < 0.05$).

	DTW	**gwDTW**	**sdDTW**
WNR	150.7	148.2*	147.9*
WDL	159.4	157.9*	158.1*
W5	194.1	192.3*	191.9*

4 CONCLUSIONS

This paper proposes two versions of a modified dynamic time warping (DTW) approach for comparing

discrete time series such as human sleep sequences: global weighted dynamic time warping (gwDTW) and stepwise deviation-based dynamic time warping (sdDTW). Both versions penalize deviations from the path of constant slope in the warping space, yielding the efficiency advantages of DTW approaches based on global constraints such as the Itakura parallelogram or the Sakoe-Chiba band, while better accounting for local deviations. gwDTW adds a deviation-based term to the standard DTW distance metric. sDTW adds a deviation term into the local cost function that drives the DTW dynamic programming optimization itself, yielding an improved warping path together with a similarity metric. Both gwDTW and sdDTW lead to significantly better clustering results than DTW in a classification task over labeled synthetic semi-Markov data, as well as in unsupervised clustering of human sleep data. The authors learned of an interesting "salient feature" approach to constrained DTW (Candan et al., 2012) after completing the work reported in the present paper. The salient feature approach extracts features of the input sequences that are then used to define locally adaptive constraints on the warping path. In future work, it would be desirable to pursue a performance comparison of the salient feature approach of (Candan et al., 2012) with that of the present paper.

ACKNOWLEDGEMENTS

The authors thank the anonymous referees for comments that helped improve the legibility of the paper, and for making us aware of (Candan et al., 2012).

REFERENCES

Alvarez, S. A. and Ruiz, C. (2013). Collective probabilistic dynamical modeling of sleep stage transitions. In *Proc. Sixth International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS 2013)*, Barcelona, Spain.

Bianchi, M. T., Cash, S. S., Mietus, J., Peng, C.-K., and Thomas, R. (2010). Obstructive sleep apnea alters sleep stage transition dynamics. *PLoS ONE*, 5(6):e11356.

Candan, K. S., Rossini, R., Wang, X., and Sapino, M. L. (2012). sDTW: computing DTW distances using locally relevant constraints based on salient feature alignments. *Proceedings of the VLDB Endowment*, 5(11):15191530.

Chu-Shore, J., Westover, M. B., and Bianchi, M. T. (2010). Power law versus exponential state transition dynamics: application to sleep-wake architecture. *PLoS ONE*, 5(12):e14204.

Clifford, D., Stone, G., Montoliu, I., Rezzi, S., Martin, F.-P., Guy, P., Bruce, S., and Kochhar, S. (2009). Alignment using variable penalty dynamic time warping. *Analytical Chemistry*, 81(3):1000–1007.

Dijk, D. J. and Lockley, S. W. (2002). Invited review: Integration of human sleep-wake regulation and circadian rhythmicity. *Journal of Applied Physiology*, 92(2):852–862.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72.

Jeong, Y.-S., Jeong, M. K., and Omitaomu, O. A. (2011). Weighted dynamic time warping for time series classification. *Pattern Recognition*, 44(9):2231–2240.

Kishi, A., Struzik, Z. R., Natelson, B. H., Togo, F., and Yamamoto, Y. (2008). Dynamics of sleep stage transitions in healthy humans and patients with chronic fatigue syndrome. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 294(6):R1980–R1987.

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Oates, T., Firoiu, L., and Cohen, P. R. (2001). Using dynamic time warping to bootstrap HMM-based clustering of time series. In *Sequence learning: Paradigms, algorithms, and applications*, pages 35–52. Springer-Verlag.

Ratanamahatana, C. A. and Keogh, E. (2004). Making time-series classification more accurate using learned constraints. In *Proceedings of SIAM International Conference on Data Mining (SDM '04)*, pages 11–22.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49.

Wang, C., Alvarez, S. A., Ruiz, C., and Moonis, M. (2014). Semi-Markov modeling-clustering of human sleep with efficient initialization and stopping. In *Proc. Seventh International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2014)*, Barcelona, Spain.