Expert Systems with Applications 39 (2012) 12814–12823

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Correlation based dynamic time warping of multivariate time series

Zoltán Bankó, János Abonyi *

University of Pannonia, Department of Process Engineering, P.O. Box 158, H-8200, Veszprem, Hungary

ARTICLE INFO

Keywords:
Dynamic time warping
Principal component analysis
Multivariate time series
Segmentation
Similarity

ABSTRACT

In recent years, dynamic time warping (DTW) has begun to become the most widely used technique for comparison of time series data where extensive a priori knowledge is not available. However, it is often expected a multivariate comparison method to consider the correlation between the variables as this correlation carries the real information in many cases. Thus, principal component analysis (PCA) based similarity measures, such as PCA similarity factor (SPCA), are used in many industrial applications.

In this paper, we present a novel algorithm called correlation based dynamic time warping (CBDTW) which combines DTW and PCA based similarity measures. To preserve correlation, multivariate time series are segmented and the local dissimilarity function of DTW originated from SPCA. The segments are obtained by bottom-up segmentation using special, PCA related costs. Our novel technique qualified on two databases, the database of signature verification competition 2004 and the commonly used AUSLAN dataset. We show that CBDTW outperforms the standard SPCA and the most commonly used, Euclidean distance based multivariate DTW in case of datasets with complex correlation structure.

1. Introduction

A time series is a sequence of values measured as a function of time. These kinds of data are widely used in the fields of process engineering (Singhal & Seborg, 2005), medicine (Tormene, Giorgino, Quaglini, & Stefanelli, 2009), bioinformatics (Aach & Church, 2001), chemistry (Abonyi, Feil, Németh, & Árva, 2005a), finance (Rada, 2008) and even for tornado prediction (Mcgovern, Rosendahl, Brown, & Droegemeier, 2011). The increasing popularity of knowledge discovery and data mining tasks for discrete data has indicated the growing need for similarly efficient methods for time series databases. These tasks share a common requirement: a (dis) similarity measure has to be defined between the elements of a given database. Moreover, the results of a data mining application from simple clustering and classification to complex decision-making systems are highly dependent on the applied dissimilarity measure.

Dissimilarity of multivariate time series can be approached from two different perspectives. The first option is to compare the variables directly and determine their weight based on a training database. Although this approach obviously has its advantages and can provide acceptable results, it is often not as effective as one would expect.

The reason of this unexpected inaccuracy is that the multivariate time series are usually much more than the collection of univariate time series as they are not only described by the variables but their relation. This relation is the correlation between the variables and it can be treated as a hidden process which carries the real description of a complex system. A classic example for this hidden process based approach is process monitoring in chemical plants: a high-density polyethylene plant requires to track more than 10 variables. Under monitoring, the tracked signals (polymer production intensity, hydrogen input, ethylene input, etc.) are measured against their stored patterns to detect any sign of malfunctions. However, the signals should not be compared to their counterparts only because the deviation in one or more signals does not mean malfunction automatically. For this reason, multivariate monitoring and control schemes based on latent variable methods have been receiving increasing attention by industrial practitioners. {...} Several companies have enthusiastically adopted the methods and have reported many success stories. Applications have been reported where multivariate statistical process control, fault detection and diagnosis is achieved by utilizing the latent variable space, for continuous and batch processes, as well as, for process transitions as for example start ups and re-starts (Kourti, 2005). Motivated by the results above, modifications of general SPCA (Krzanowski, 1979) were developed for various purposes (Gunther, Conner, & Seborg, 2008; Johannesmeyer, Singhal, & Seborg, 2002; Yang & Shahabi, 2007).

Although PCA considers the time series as a whole, it does not take into account the alternations in the relationship between the variables. The main goal of this paper is to construct a dissimilarity measure which deals with the changes in the correlation structure of the variables as flexible as DTW allows.

* Corresponding author. Tel.: +36 88 624209.
E-mail address: abonyij@fmt.uni-pannon.hu (J. Abonyi).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.012

The increasing and less and less costly computational power made possible to create similarity measures without considering their indexing capabilities for tasks where the quality of the comparison is much more important than the speed of the calculation. Non-linear matching techniques such as DTW, longest common subsequences and symbolic approximation have been studied extensively for this purpose. DTW, which has long been known in the speech recognition community (Sakoe & Chiba, 1971), excelled these methods in popularity thanks to its adaptability (Giorgino, 2009) and efficiency. It was successfully applied variety of problems in various disciplines from signature verification (Kholmatov & Yanikoglu, 2005) to weather prediction (Mcgovern et al., 2011).

Concluding the above cited results, it is desirable to combine the strength of DTW and PCA similarity factor. We propose a new and intuitive method which is based on DTW aided PCA and segmentation, named correlation based dynamic time warping (CBDTW). The coherent parts of a multivariate time series define segments; therefore, segmentation is applied to address that the PCA similarity factor does not take into account the alternation of variables. These segments can be compared directly; however, DTW is used as it makes the presented dissimilarity measure invariant to phase shifts of the time axis and to the differences in the number of segments. Moreover, DTW is also capable of compensating the "locally elastic" shifts (local time warping) of time series.

The presented algorithm was qualified on two databases which differ from the correlation point of view: the database of signature verification competition 2004 and the AUSLAN dataset which is widely used by the time series data mining community (Frank & Asuncion, 2010). We show that CBDTW obviously outperforms the standard SPCA independently of the complexity of the correlation and it also surpasses the most commonly used Euclidean distance based multivariate DTW for the AUSLAN dataset which has a complex correlation structure of 22 variables.

The rest of the paper is organized as follows. Section 2 details the nomenclature of the proposed dissimilarity measure and segmentation. In Section 3, we introduce the novel similarity measure in full detail, while Section 4 conducts a detailed empirical comparison of the presented method with other techniques. Finally, in Section 5 we make our conclusions and suggestions for future work.

2. Theoretical background

X_n is an n-variable, m-element time series where x_i is the ith variable and $x_i(j)$ denotes its jth element:

$$\begin{aligned} &X_n = [x_1, x_2, x_3, \ldots, x_n], \\ &x_i = [x_i(1), x_i(2), \ldots, x_i(j), \ldots, x_i(m)]^T \end{aligned} \tag{1}$$

According to this notation a multivariate time series can be represented by a matrix in which each column corresponds to a variable and each row represents a sample of the multivariate time series at a given time:

$$\begin{bmatrix} X_n(1) \\ X_n(2) \\ \vdots \\ X_n(m) \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ x_1(2) & x_2(2) & \cdots & x_n(2) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(m) & x_2(m) & \ldots & x_n(m) \end{bmatrix} \tag{2}$$

Throughout this paper, the dissimilarity between X_n and Y_n is denoted by $d(X_n, Y_n)$, where $0 \leqslant d(X_n, Y_n)$, $d(X_n, Y_n) = d(Y_n, X_n)$ and $d(X_n, X_n) = 0$.

As it was mentioned before, the time series should be segmented to consider the alternation of the latent variable. Moreover, the segmentation has another advantage, i.e. it speeds up DTW which is computationally expensive ($O(m^2)$, where m is the length of the time series).

The ith segment of X_n is a set of consecutive time points, $S_i(a,b) = [X_n(a); X_n(a+1); \ldots; X_n(b)]$. The c-segmentation of time series X_n is a partition of X_n to c non-overlapping segments, $S^c_{X_n} = [S_1(1,a); S_2(a+1,b); \ldots; S_c(k,m)]$. In other words, a c-segmentation splits X_n to c disjoint time intervals, where $1 \leqslant a$ and $k \leqslant m$. The segmentation problem can be framed in several ways (Keogh, Chu, Hart, & Pazzani, 2001), but its main goal is always the same: finding homogenous segments by the definition of a cost function, $cost(S_i(a,b))$. This function can be any arbitrary function which projects the space of multivariate time series to the space of the non-negative real numbers. Usually, $cost(S_i(a,b))$ is based on the differences between the values of the ith segment and its approximation by a simple function f (constant or linear function, a polynomial of a higher but limited degree):

$$cost(S_i(a,b)) = \frac{1}{b-a+1} \sum_{l=a}^{b} d(X_n(l), f(X_n(l))) \tag{3}$$

Thus, the segmentation algorithms simultaneously determine the parameters of the models and the borders of the segments by minimizing the sum of the costs of the individual segments:

$$cost(S^c_{X_n}) = \sum_{i=1}^{c} cost(S_i(a,b)) \tag{4}$$

The segmentation cost of a time series can be minimized by dynamic programming, which is computationally intractable for many real datasets (Himberg, Korpiaho, Mannila, Tikanmaki, & Toivonen, 2001). Consequently, heuristic optimization techniques such as greedy top-down or bottom-up techniques are frequently used to find good but suboptimal c-segmentations:

- **Bottom-up**: Every element of X_n is handled as a segment. The costs of the adjacent segments are calculated and two segments with the minimum cost are merged. The merging cost calculation of adjacent segments and the merging are continued until some goal is reached.
- **Top-down**: The whole X_n is handled as a segment. The costs of every possible split are calculated and the one with the minimum cost is executed. The splitting cost calculation and splitting is continued recursively until some goal is reached.
- **Sliding window**: The first segment is started with the first element of X_n. This segment is grown until its cost exceeds a predefined value. The next segment is started at the next element. The process is repeated until the whole time series is segmented.

All of these segmentation methods have their own specific advantages and drawbacks. Accordingly, the sliding window method is the fastest one, however, it is not able to divide up a sequence into predefined number of segments. The applied method depends on the given task. Keogh et al. (2001) examined these heuristic optimization techniques in detail through the application of piecewise linear approximation. It can be said if real-time (on-line) segmentation is not required, the best result can be reached by bottom-up segmentation.

3. Correlation based dynamic time warping of multivariate time series

DTW allows us to select the local dissimilarity function (dissimilarity between the data points.[1]) Most frequently Euclidean

[1] For a comprehensive discussion of DTW, see Rabiner & Juang (1993).

distance is used for this purpose (Giorgino, 2009). It pairs the points of the n-dimensional space and compares them to each other, as it is shown in Fig. 1. This approach is useful when all of the variables have to be used, they are all measured in the same scale and there are no significant differences in their amplitudes and values. Even if such differences exist, z-normalization can be used.

Although correlation is slightly considered, there is a serious problem: DTW creates "singularities" because it tries to minimize the variance of the local dissimilarities between the points by warping the time axis. This property can prevent DTW to align trends if they are located slightly "higher" or "lower" than their corresponding pair. Derivative DTW (DDTW) was introduced (Keogh & Pazzani, 2001) to correct this behavior; however, DDTW has not been extended for multivariate time series.

The handicap of SPCA based methods is more obvious. They are not able to handle the alternations of the correlation structure which affect the hyperplanes, therefore in most real-life applications segmentation is required to create homogeneous segments from the viewpoint of the correlation structure. However, the segmentation raises another problem. Although in many real-life applications a lot of variables must be simultaneously tracked and monitored, most of the segmentation algorithms are used for the analysis of only one time-variant variable (Kivikunnas, 1998). Usage of only one variable for segmentation of multivariate time series is not precise enough when the correlation between the variables is an important factor. Moreover, the higher dimensional segmentation problems, such as surface simplification (Heckbert & Garland, 1997), have much better understanding than its multivariate relative.

3.1. Proposed algorithm and evolution method

To overcame the above mentioned problem of SPCA, the PCA based segmentation can be applied which we presented previously (Abonyi, Feil, Németh, & Árva, 2005b). Hotelling's T^2 statistics and the Q reconstruction error were used as the measure of the homogeneity of the segments, i.e. to construct the cost function. Fig. 2 shows these two measures in case of a 2-variable 11-element time series. The elements are represented by black ellipses and the gray spot marks the intersection of the axes of the principal components, i.e. the center of the space which was defined by these principal components. If the second principal component is ignored then two distances can be computed for each element. The first one is the squared Euclidean distance between the original data point and its reconstructed value using the most significant principal component only. The arrow noted by Q represents this lost information which can be computed for the jth data point of the times series X_n as follows:

$$Q(j) = (X_n(j) - \widehat{X}_n(j))(X_n(j) - \widehat{X}_n(j))^T = X_n(j)\left(I - U_{X_n,p}U_{X_n,p}^T\right)X_n(j)^T, \tag{5}$$

where $\widehat{X}_n(j)$ is the jth predicted value of X_n, I is the identity matrix and $U_{X_n,p}$ is the matrix of eigenvectors. These eigenvectors belong to the most important $p \leqslant n$ eigenvalues of covariance matrix of X_n, thus they describe the hyperplanes. Please note, the Q error based segmentation can be considered as the natural extension of Piecewise Linear Approximation which was presented by Keogh and Pazzani (1999). Both of them define the cost function based on the Euclidean distance between the original data and its reconstructed value from a lower dimensional hyperplane.

The second measure which can be used to construct the cost function is Hotelling's T^2 statistic. This shows the distance of each element from the center of the data, hence it signs the distribution of projected data. Its formula is the following for the jth point:

$$T^2(j) = Y_p(j)Y_p(j)^T, \tag{6}$$

where $Y_p(j)$ is the lower (p) dimensional representation of $X_n(j)$. The cost functions can be defined as:

$$\begin{aligned} cost_Q(S_i(a,b)) &= \frac{1}{b-a+1}\sum_{j=a}^{b} Q(j) \\ cost_{T^2}(S_i(a,b)) &= \frac{1}{b-a+1}\sum_{j=a}^{b} T^2(j) \end{aligned} \tag{7}$$

Using one of the above mentioned PCA based segmentation, the correlation based dynamic time warping of multivariate time series can be realized. The proposed method can be summarized as follows:

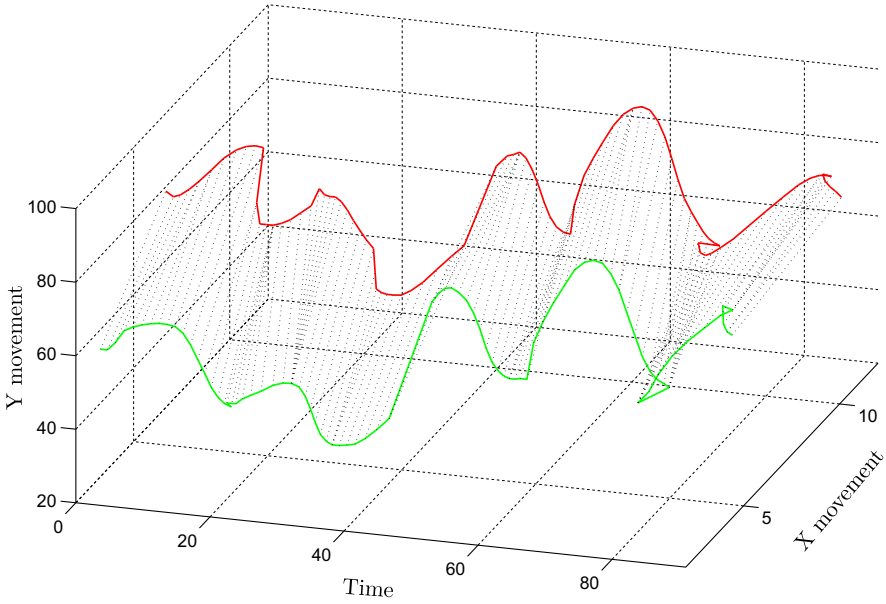


Fig. 1. Trajectories of two signatures compared with Euclidean distance based DTW.

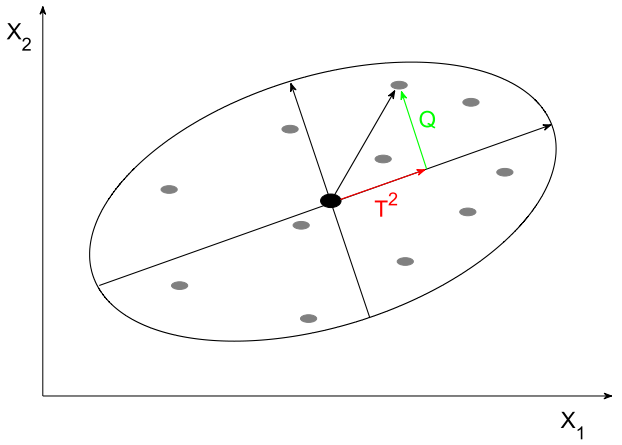


Fig. 2. Measures to use for PCA model based segmentation.

- Segment the time series of the given database based on correlation to create the most homogenous segments. The projection error or Hotelling's statistics can be used as basis of the cost function. This segmentation can be done off-line in most cases.
- Segment the query time series according to the database.
- Calculate the DTW dissimilarity between the query and the time series of the database. The local dissimilarity measure of DTW can be chosen arbitrarily. In this paper, it is derived from a covariance based similarity measure, i.e., $1 - SPCA$.

Validation of a similarity measure which is not optimized for a specific database cannot be carried out theoretically, experimental evaluation must be made on wide range of different datasets. Unfortunately, there is no similar classification/clustering page for multivariate time series as for univariate series (Keogh et al., 2011). So, to perform the validation of CBDTW, a modified leave-one-out k nearest neighbor search algorithm was used. For the pseudocode see Algorithm 1. In the first *for* loop the *precision* array is initialized which contains the values of precision at the given values of recall. The precision is calculated for every item of the database in the main *for* loop. The value of k (number of nearest neighbors) and r (number of required items from the same class) is set for the actual time series and the kNN search is performed in the *while* loop. If the number of retrieved items from the same class (c) is equal to r then the *precision* array is updated and the value of k and r are incremented. If the number of items from the same class is less than the required value (r), the algorithm looks for more neighbors (increments the value of k). This continues until the value of r is less than or equal to *r_items*. For simplicity and according to Yang and Shahabi (2004), the value of *r_items* was chosen to 10.

Using the *precision* array, a recall-precision graph can be plotted which is a common tool to measure and demonstrate the performance of the information retrieval (IR) systems (Frakes & Baeza-Yates, 1992). The precision expresses the proportion of the relevant sequences from the set of the retrieved items. Similarly, the recall is the number of the relevant elements in a database retrieved by the k nearest neighbor search. Note that the graph can be considered as the extension of the 1-NN search used in by Keogh et al. (2011).

Algorithm 1. The pseudocode of modified leave-one-out k nearest neighbor search for recall-precision diagrams

```
Input: N: the number of multivariate time series in the database
Input: k: the number of required nearest neighbors
Input: r_items: the number of relevant items
Output: precision: the array of the values of precision as the function of recall
/* Create the result array */
for (i = 1; i ⩽ r_items; i++) do
    precision(i) = 0
end
/* Calculate the precision for every element in dataset */
for (i = 1; i ⩽ N; i++) do
    /* Select the ith time series from the database */
    curr_item = sequences_in_database(i);
    /* Number of nearest neighbors */
    k = 1;
    /* Number of requested relevant items */
    r = 1;
    while (r ⩽ r_items) do
        /* Perform kNN search for curr_item, c is the number of the items from the k retrieved items with the same label as curr_item */
        c = knnsearch(curr_item, k);
        if (c == r) then
            precision[r] = precision[r] + c/k;
            r = r + 1;
        end
        k = k + 1;
    end
end
for (i = 1; i ⩽ 10; i++) do
    precision(i) = precision(i)/N;
end
```

Before Algorithm 1 was executed, two important parameters had to be selected for both datasets. The first one is the number of principal components (p) as the increasing number of principal components decreases the reconstruction error. If p equals to the number of the variables, the Q reconstruction error becomes zero and the values of T^2 statistics become the real distances in the whole dataset. On the other hand, if p is too small, the reconstruction error will be too large to characterize the covariance precisely. In these two extreme cases the segmentation is not based on the internal relationships of the variables, so simple equidistant segments can be detected. To avoid this, the value of p has to be selected carefully, i.e. the first few p eigenvalues should contain more than 90% of the total variance.

The other important parameter is the number of segments. It can be determined by different techniques such as using permutation test to determine whether the increase of the model accuracy with the increase of the number of segments is due to the underlying structure of the data or due to the noise (Vasko & Toivonen, 2002). For simplicity we applied a similar but much simpler method to determine the number of segments. Our method based on the weighted modeling error where the weight is the number of the segments. To get a clearer picture, the relative reduction rate of the modeling error is also used:

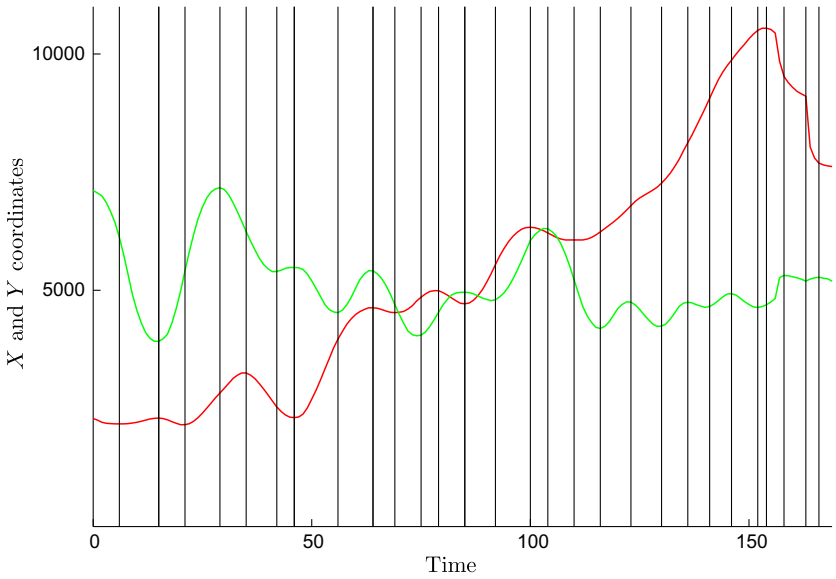


Fig. 3. Signature from the database of SVC2004. The vertical lines mark the major changes in covariance between the time series of the two coordinates x and y.

$$RR(S^c_{X_n}) = \frac{cost\left(S^{c-1}_{X_n}\right) - cost\left(S^c_{X_n}\right)}{cost\left(S^{c-1}_{X_n}\right)}, \quad (8)$$

where $RR(S^c_{X_n})$ is the relative reduction of error when c segments are used instead of $c-1$ segments.

4. Experimental results

In the following, the detailed empirical comparison of the presented method (CBDTW) with Euclidean-distance (EUC), Euclidean-distance based multivariate DTW, PCA similarity factor (SPCA) and its segmented version (SEGPCA) is presented. To demonstrate the scalability of the new method, we examined different number of segments and hyperplanes. Before the results are discussed, we present the applied parameters to help reproduction.

Euclidean distance is not able to handle time series with different length and DTW tends to show its best when the sequences have the same length (Ratanamahatana & Keogh, 2005). Hence, we have interpolated the time series to their average length in each database for these two methods. The PCA based methods do not require such an action; however, the average length of the AUSLAN dataset is only 57. Thus, to prepare all of its times series for segmentation, linear interpolation[2] was used to obtain sequences with a length of 300. Another important parameter is how to frame the segmentation. As previously mentioned, the best segmentation can be reached by the bottom-up algorithm (Keogh et al., 2001), thus we used this. The dissimilarity between the corresponding segments was arisen from the PCA similarity factor, that is $1 - SPCA$.

The applied constraints on the warping path of DTW, both global and local, also had to be defined. Ratanamahatana and Keogh (2004, 2005) stated that "*all the evidence suggests that narrow* (global) *constraints are necessary for accurate DTW*" which is obviously true for properly preprocessed datasets used in most data mining applications. However, sometimes there is no chance to do proper preprocessing and compensate the initial/ending shifts of the time series in real-time application due to the time or hardware limit. Moreover, the difference between using an optimized warping path (e.g., applying R–K band (Ratanamahatana & Keogh, 2004)) or no warping path is often not significant even for preprocessed datasets (Keogh et al., 2011). Thus, as a bullet proof solution, we did not use any global constraints. We also assumed that no extensive knowledge exists on the databases, thus the most simple and yet widely used local constraint was selected, i.e. type I[3] of Rabiner and Juang (1993).

Selecting databases for validation purposes is almost as hard as the creation of a new and useful similarity measure. There is no argue that the best data should come from the industrial world (production data, plant supervision data, real-time sensor information provided for ECUs, etc.); however, these kinds of data are rarely allowed to be published. Thus, according to Keogh and Kasetty (2003), two datasets were selected for validation purposes which are available on the Internet. An important aspect behind the selection of them was the difference between their correlation structure which can affect the efficiency of any PCA based similarity measure. If the underlying ("hidden") process can be seen easily and the correlation of the variables does not vary inner class then the PCA based similarity measures are not as effective as the conventional measures like Euclidean distance or its warped version.

4.1. SVC2004

The aforementioned first database was created for signature verification conference 2004 (Yeung et al., 2004). It has 40 sets of signature data and each set contains 20 genuine signatures from one signature contributor and 20 skilled forgeries from five other contributors. Although both genuine signatures and skilled forgeries are available, obviously the 800 genuine signatures were used for validation only. In each signature file, the signature is represented as a sequence of points. The first line stores the total number of points, averages 184. The signatures were collected on a WACOM Intuos tablet, hence seven parameters of the pen was measured under the enrollment process:

- X-coordinate – scaled cursor position along the x-axis.
- Y-coordinate – scaled cursor position along the y-axis.
- Time stamp – system time at which the event was posted (not used in this paper).
- Button status – current button status (0 for pen-up and 1

[2] Using interp1 function of Matlab.

[3] Symmetric, non-normalizable local constraint.

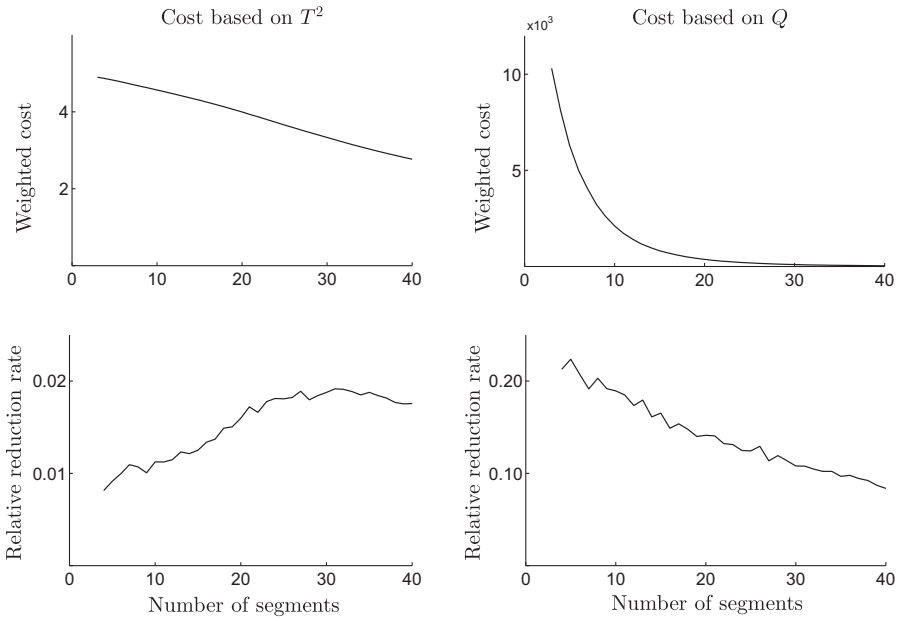


Fig. 4. The weighted costs and their relative reduction rates with number of segments in case of SVC2004 dataset using two hyperplanes.

for pen-down, not used in this paper).

- Azimuth – clockwise rotation of cursor about the z-axis.
- Altitude – angle upward toward the positive z-axis.
- Pressure – adjusted state of the normal pressure.

Please note, the time stamps and the button status were not used in this paper. The reason behind this decision is very simple: time stamps and button status do not advance the accuracy considerably, except when special handwriting recognition techniques are used such as stroke detection, equidistant resampling, etc. Furthermore, the usage of time stamps requires an extra interpolation step which makes harder to reproduce the results.

SVC2004 is an ideal dataset for any segmentation based method because the correlation alternates many times between the variables as it was illustrated in Fig. 3. However, the hidden process is not as hidden as one can expect. Only one of the coordinates can be used to represent the whole signature due to the fact that the variables usually change in the same way at the same place for a given participant.

For validation, the number of segments and the number of principal components had to be chosen in advance for the PCA based methods. The number of hyperplanes was determined by using the desired accuracy (loss of variance) of the PCA models. The first one, two, three and four eigenvalues describe 79.84%, 99.04%, 99.91% and 99.99% of the total variance respectively. To demonstrate the dependency of PCA based methods on the tightness of the representation both 2 and 4 dimensional hyperplanes were used for such techniques.

The number of the segments was achieved by plotting the weighted costs ($cost_{T^2}/c$ and $cost_Q/c$, where c is the number of segments) and their relative reduction rates as the function of the number of segments. Fig. 4 shows these plots. Considering how dif-

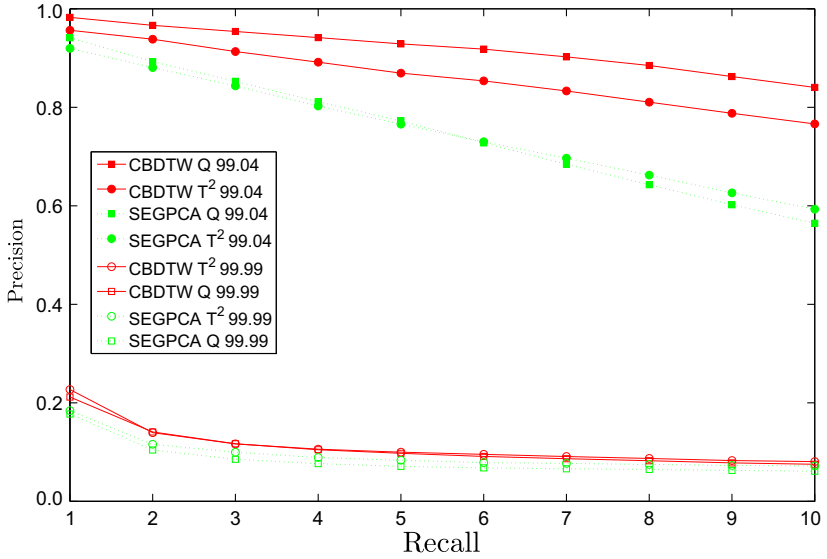


Fig. 5. The recall-precision graph of SVC2004 dataset using SEGPCA (dashed lines) and CBDTW (solid lines) with two (filled marks) and four (empty marks) hyperplanes. Circle and square marks used to show whether T^2 or Q based segmentation is used.

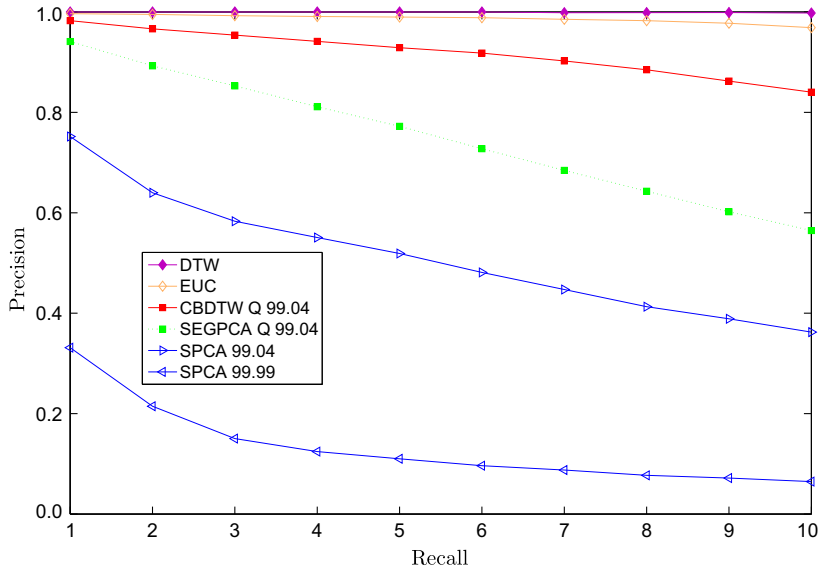


Fig. 6. The recall-precision graph of SVC2004 dataset using all dissimilarity measures.

ferent each signature is—i.e. how they differ from correlation structure point of view—, it is not surprising that there is no optimal segmentation from the viewpoint of relative error reduction, i.e. there is no brake point in the relative reduction rates. However, there is a practical limit on the number of segments ($\lfloor 184/(p+1) \rfloor$), thus it was chosen to 20 which is suitable both 2 and 4 principal components.

The results of Algorithm 1 using CBDTW and SEGPCA on SVC2004 database are shown in Fig. 5. It clearly seems that the third and fourth hyperplanes do not add any useful information to the first and second, but makes the classification less effective because the PCA similarity factor weights the eigenvectors equally. It is also has to be noted that independently of the type of the segmentation cost and the number of retained principal components, CBDTW provide more precise results than SEGPCA.

Finally, Algorithm 1 was executed for all other methods. In the light of the fact that the hidden process is not "hidden enough", it is not surprising that Euclidean distance and its warped version outperformed all other methods as it can be seen in Fig. 6. From CBDTW point of view, the relations of the PCA based techniques are much more interesting: the graph of SEGPCA shows that this technique really excelled the segmentation free, standard SPCA and the application of CBDTW improved its results even further.

4.2. AUSLAN

The high quality version of Australian language sign dataset (AUSLAN) collected by (Kadous, 2002) has been also selected for validation purposes. It was provided by (Keogh et al., 2011) and both the normal and high quality version can be downloaded from the UCI machine learning repository (Frank & Asuncion, 2010) as of 2012.

The high quality dataset contains 95 signs and each of them has 27 examples, totals 2565. Two 5DT gloves as well as two Flock-

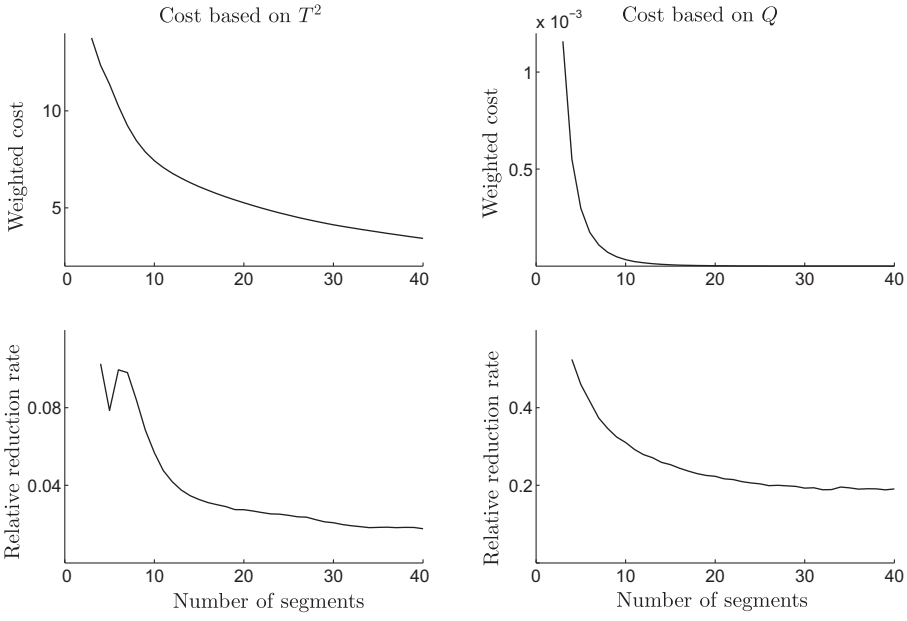


Fig. 7. The weighted costs and their relative reduction rates with number of segments in case of AUSLAN dataset using two hyperplanes.

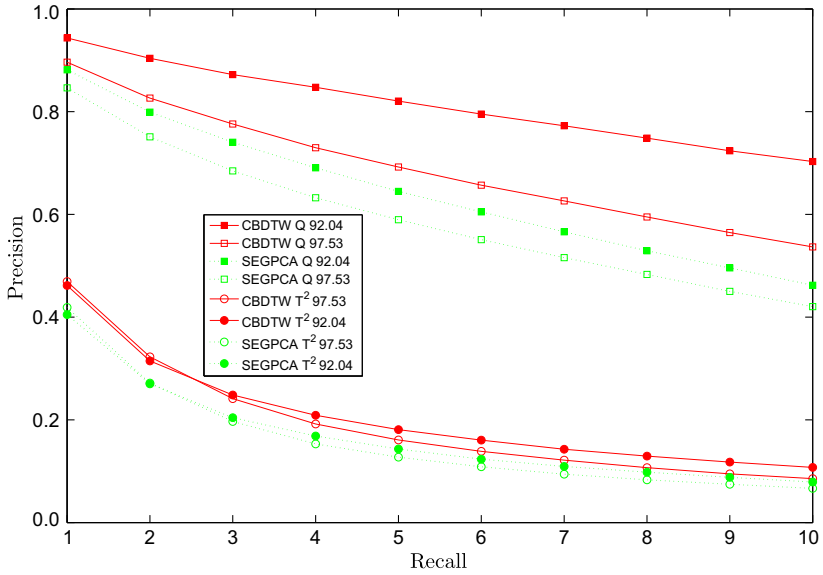


Fig. 8. The recall-precision graph of AUSLAN dataset using SEGPCA (dashed lines) and CBDTW (solid lines) with one (filled marks) and two (empty marks) hyperplane(s). Circle and square marks used to show whether T^2 or Q based segmentation is used.

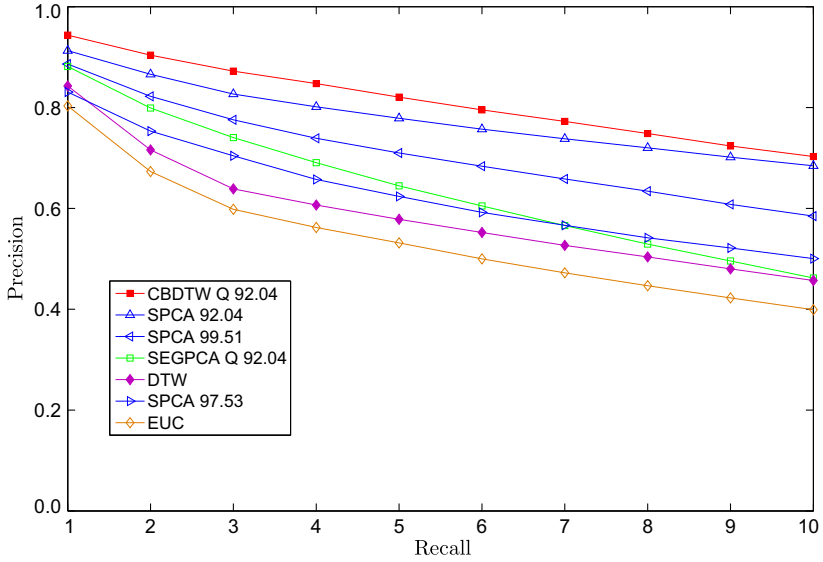


Fig. 9. The recall-precision graph of AUSLAN dataset using all dissimilarity measures.

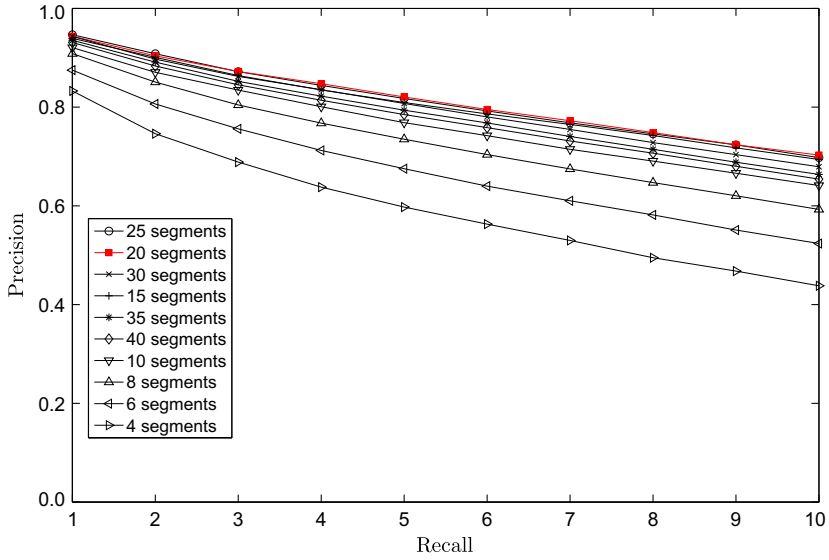
of-Birds magnetic position trackers with refresh rate of 100 Hz were used to acquire the signs of a native signer. The signals of 11 channels were recorded from each hand. The position trackers measured the x, y, z coordinates, the roll, pitch and yaw for each hand. The gloves also provided the finger bend data from the five fingers. Position and orientation were defined to 14-bit accuracy, giving position information with a typical positional error less than one centimeter and angle error less than one half of a degree. Finger bend was measured with 8 bits per finger.

AUSLAN is a much more complex dataset than SVC2004. It has 22 variables, and lots of them are 0 most of the time. This yields to the underlying process is much more hidden than it was for SVC2004. In addition, the average length of the time series is only 57, so some preprocessing steps were necessary. It is obvious that correlation based dynamic time warping requires 10–20 segments at least for effective warping and this cannot be guaranteed with this average length because a segment has 0 $cost_Q$ until the number of its elements are not equal or exceed to the number of the principal components. Hence, the sequences were interpolated[4] to a sufficiently large, fixed length which was chosen to 300. Please note, this interpolation is also used for SEGPCA, but not executed for the standard PCA similarity factor.

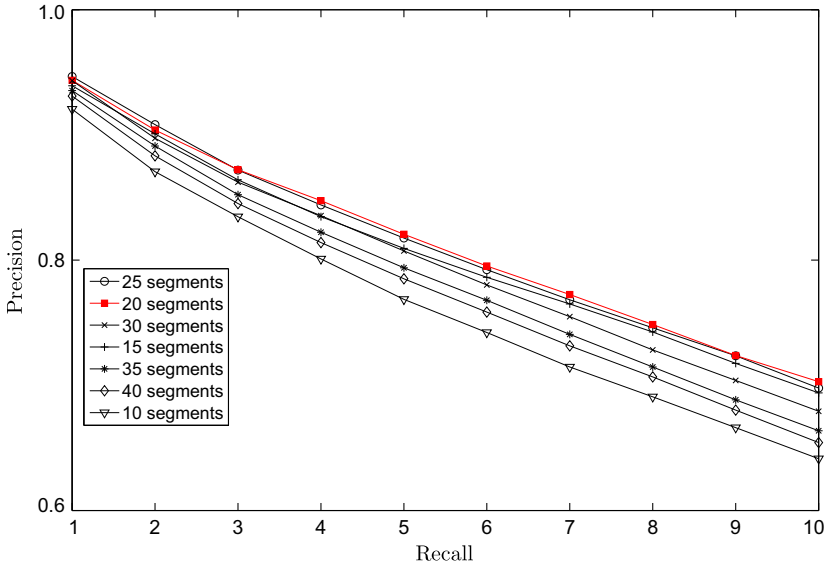
The first one, two, three and four eigenvectors describe 92.04%, 97.53%, 98.95% and 99.51% of the total variance, thus the first one eigenvector was also selected – besides the first two and four eigenvectors – to provide basis for PCA based methods. The weighted error and its relative reduction rate using two hyperplanes can be seen in Fig. 7. According to this, 20 segments have seemed as appropriate as it was for SVC2004.

The precision-recall graphs of SEGPCA and CBDTW are plotted using one and two hyperplanes in Fig. 8. The graphs of four hyper-

[4] Using interp1 function of Matlab.



(a) Using 4, 6, 8, 10, 15, 20, 25, 30, 35, 40 segments



(b) Focusing on the best results only

Fig. 10. The recall-precision graph of AUSLAN dataset for CBDTW using different number of segments.

planes based methods are located close to the last four graphs, so they were omitted to maintain readability. Although it can be seen that Fig. 8 represents similar results as Fig. 5 (the graphs can be divided into two groups); however, what makes the difference in this case between the two groups is not the number of hyperplanes but the type of the segmentation cost.

The results of the validation are shown in Fig. 9. The high values of PCA similarity factor based methods show that the correlation obviously describes the classes. Considering the fact how many variables had to be tracked to observe the underlying processes (i.e. 22 variables had to be monitored to record the signs which can be expressed almost as good by using only 1 "hidden" variable) it is not surprising at all that Euclidean distance and its time warped extension are not performed as well as they did for SVC2004.

For AUSLAN, CBDTW proved its superiority over all other methods; however, it also interesting to know how CBDTW scales with the number of the segments. Thus, we executed CBDTW again using different number of segments. The results can be seen in Fig. 10.

5. Conclusion and future work

In this paper, we presented a novel similarity measure for highly correlated multivariate time series. Our method based on covariance driven segmentation and dynamic time warping. We utilized two homogeneity measures as cost function for segmentation. These homogeneity measures correspond to the two typical applications of PCA models. The Q reconstruction error can be used to segment the time series according to the direct change of the correlation among the variables, while the Hotelling's T^2 statistics can be utilized to segment the time series based on the drift of the center of the operating region. The dissimilarity between the segments were derived from the simple PCA similarity factor. Finally, we applied DTW to compensate the time shifts and make the presented dissimilarity measure more accurate.

To prove that CBDTW can outperform PCA similarity factor in any environment, CBDTW was tested on two datasets which differ from correlation point of view. AUSLAN has 22 variables with a

complex correlation structure. It was selected to simulate the "typical" industrial data, i.e. large number of variables, whose correlation structure cannot be revealed without the application of PCA. The algorithm was also tested on the dataset of SVC2004 in which, contrary to AUSLAN, the correlation between the variables is obvious.

The recall-precision graphs showed superiority of CBDTW over PCA similarity factor – irrespective of the complexity of the hidden process – and it even outperforms the Euclidean distance based dynamic time warping when a high number of variables with complex correlation structure has to be handled. The results also show that the proposed algorithm can replace the standard PCA similarity factor in many areas such as distinguishing and clustering typical operational conditions and analyzing product grade transitions of process systems. Moreover, CBDTW can be used for data mining purposes when indexing is not required.

Acknowledgements

This work was supported by the TAMOP-4.2.1/B-09/1/KONV-2010-0003 and TAMOP- 4.2.2/B-10/1-2010-0025 projects and the E.ON Business Services Kft. The authors also thank Prof. Eamonn J. Keogh for providing us the preprocessed version of the AUSLAN dataset.

References

Aach, J., & Church, G. M. (2001). Aligning gene expression time series with time warping algorithms. *Bioinformatics, 17*, 495–508.

Abonyi, J., Feil, B., Németh, S., & Árva, P. (2005a). Modified Gath–Geva clustering for fuzzy segmentation of multivariate time-series. *Fuzzy Sets and Systems, 149*, 39–56.

Abonyi, J., Feil, B., Németh, S., & Árva, P. (2005b). Principal component analysis based time series segmentation. In *IEEE international conference on computational cybernetics*.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall.

Frank, A., & Asuncion, A. (2010). *UCI machine learning repository*. Irvine, CA: University of California - School of Information and Computer Science. <http://archive.ics.uci.edu/ml>.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in *r*: The dtw package. *Journal of Statistical Software, 31*, 1–24.

Gunther, J. C., Conner, J. S., & Seborg, D. E. (2008). Fault detection and diagnosis in an industrial fed-batch cell culture process. *Biotechnology Progress, 23*, 851–857.

Heckbert, P. S., & Garland, M. (1997). Survey of polygonal surface simplification algorithms. In *Proceedings of the 24th international conference on computer graphics and interactive techniques multiresolution surface modeling course*.

Himberg, J., Korpiaho, K., Mannila, H., Tikanmaki, J., & Toivonen, H. (2001). Time series segmentation for context recognition in mobile devices. In *ICDM* (pp. 203–210).

Johannesmeyer, M. C., Singhal, A., & Seborg, D. (2002). Pattern matching in historical data. *Aiche Journal, 48*, 2022–2038.

Kadous, M. W. (2002). Temporal classification: Extending the classification paradigm to multivariate time series. Ph.D. Thesis School of Computer Science and Engineering, University of New South Wales.

Keogh, E. J., Chu, S., Hart, D., & Pazzani, M. J. (2001). An online algorithm for segmenting time series. In *ICDM* (pp. 289–296).

Keogh, E. J., & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery, 7*, 349–371.

Keogh, E. J., & Pazzani, M. (1999). Scaling up dynamic time warping to massive datasets. In *3rd European conference on principles and practice of knowledge discovery in databases (PKDD'99)* (Vol. 1704, pp. 1–11).

Keogh, E. J., & Pazzani, M. J. (2001). Dynamic time warping with higher order features. In *Proceedings of the 2001 SIAM international conference on data mining*.

Keogh, E. J., Zhu, Q., Hu, B., Hao, Y., Xi, X., Wei, L., & Ratanamahatana, C. A. (2011). The UCR time series classification/clustering homepage. <www.cs.ucr.edu/~eamonn/time_series_data/>. Riverside CA. University of California – Computer Science and Engineering Department.

Kholmatov, A., & Yanikoglu, B. (2005). Identity authentication using improved online signature verification method. *Pattern Recognition Letters, 26*, 2400–2408.

Kivikunnas, S. (1998). Overview of process trend analysis methods and applications. In *ERUDIT workshop on applications in pulp and paper industry, ERUDIT*.

Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing, 19*, 213–246.

Krzanowski, W. (1979). Between-groups comparison of principal components. *Journal of the American Statistical Society, 74*, 703–707.

Mcgovern, A., Rosendahl, D. H., Brown, R. A., & Droegemeier, K. K. (2011). Identifying predictive multi-dimensional time series motifs: An application to severe weather prediction. *Data Mining and Knowledge Discovery, 22*, 232–258.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Rada, R. (2008). Expert systems and evolutionary computing for financial investing: A review. *Expert Systems with Applications, 34*, 2232–2240.

Ratanamahatana, C. A., & Keogh, E. J. (2004). Making time-series classification more accurate using learned constraints. In *SDM*.

Ratanamahatana, C. A., & Keogh, E. J. (2005). Three myths about dynamic time warping data mining. In *SDM*.

Sakoe, H., & Chiba, S. (1971). A dynamic programming approach to continuous speech recognition. In *Proceedings of the seventh international congress on acoustics* (Vol. 3, pp. 65–69), Budapest.

Singhal, A., & Seborg, D. E. (2005). Clustering multivariate time-series data. *Journal of Chemometrics, 19*, 427–438.

Tormene, P., Giorgino, T., Quaglini, S., & Stefanelli, M. (2009). Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine, 45*, 11–34.

Vasko, K. T., & Toivonen, H. T. T. (2002). Estimating the number of segments in time series data using permutation tests. In *ICDM '02: Proceedings of the 2002 IEEE international conference on data mining (ICDM'02)*. Washington, DC, USA: IEEE Computer Society (p. 466).

Yang, K., & Shahabi, C. (2004). A PCA-based similarity measure for multivariate time series. In *MMDB '04: Proceedings of the 2nd ACM international workshop on multimedia databases* (pp. 65–74). ACM Press.

Yang, K., & Shahabi, C. (2007). An efficient *k* nearest neighbor search for multivariate time series. *Information and Computation, 205*, 65–98.

Yeung, D. -Y., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T., & Rigoll, G. (2004). SVC2004: First international signature verification competition. In *Proceedings of the International Conference on Biometric Authentication (ICBA)* (Vol. 3072, pp. 16–22).