Multivariate Time Series Data Mining: PCA-based Measures for Similarity Search

L. Karamitopoulos[1], **G. Evangelidis**[1], **D. A. Dervos**[2]
[1] Department of Applied Informatics, University of Macedonia, 156 Egnatia St., 54006, Thessaloniki, Greece
[2] Department of Information Technology, Alexander Technological Educational Institute of Thessaloniki, P.O. Box 141, 57400 Sindos, Greece

Abstract - *In this paper, we discuss the application of Principal Component Analysis (PCA), for the purpose of determining a similarity/distance measure among multivariate time series. We review several PCA-based measures that have been proposed by researchers from diverse scientific fields and we extend the well-known statistic in the Statistical Process Control community, SPE, in order to define a novel distance measure. We conducted experiments on four datasets, which have been used extensively in the literature, and we provide the results of their performance with respect to classification accuracy. Experiments indicate that there is no measure that can be clearly considered as the most appropriate one for any dataset, and that the newly proposed measure is a promising option for similarity search.*

Keywords: Similarity Search, Principal Component Analysis, Time Series, Similarity Measure, Data Mining.

1 Introduction

Technological advances in automated monitoring systems and sensor devises have facilitated the generation of huge amounts of data in the form of time series. A time series is a collection of observations made sequentially through time. At each time point one or more measurements may be monitored corresponding to one or more attributes under consideration. The resulting time series is called univariate or multivariate respectively. Multivariate time series appear frequently in several diverse applications. Examples include human motion capture [1], geographical information systems [2], statistical process monitoring [3], or intelligent surveillance systems [4]. The need for analyzing efficiently this volume of information led to the adjustment of data mining techniques in order to incorporate the temporal nature of data. At the core of these techniques lies the concept of similarity since most of them require searching for similar patterns. For instance, it is of interest to form clusters of objects that move similarly by analyzing data from surveillance systems or classify current operating conditions in a manufacturing process into one of several operational states. In the field of computer graphics, an animator needs to search efficiently a motion database for similar motions to a desired one [5].

There is a vast literature in univariate time series similarity search, mainly focused on the interrelated issues of representation [6], distance/similarity measure [7] and indexing [8]. However, the case of multivariate time series has not been extensively explored with respect to these issues. In this case, the similarity is sought among objects that consist of p-dimensional time series, that is, there are p attributes of consideration measured sequentially through time. Most of the papers concentrate on indexing multidimensional time series and provide an appropriate representation scheme and/or a similarity measure. In addition to that, most of the research interest lays on trajectories, which usually consist of 2 or 3 dimensional time series. The authors of [9] and [10] suggest similarity measures based on the Longest Common Subsequence (LCSS) model, whereas a modified version of the Edit Distance for real-valued series is provided in [11]. Vlachos et al. [12] propose an indexing framework that supports multiple similarity/distance functions, without the need to rebuild the index. Several researchers approach similarity search by applying a measure and/or an indexing method on transformed data. Kahveci et al. [13] propose to convert a p-dimensional time series of length n to a univariate time series of length np by concatenation, and then apply a representation scheme for the purpose of dimensionality reduction. On the other hand, Lee et al. [14] propose a scheme for searching a database, which, in the pre-processing phase includes the representation (e.g. DFT) of each one of the p time series separately. Cai & Ng [15] approximate and index multidimensional time series with Chebyshev polynomials and prove the Lower Bounding Lemma for this representation. That is, that the true distance between two time series is lower-bounded by the distance in the space of Chebyshev coefficients. In the previous three papers, the Euclidean distance is applied as a distance measure. Finally, Bakalov et al. [16] extend the *Symbolic Aggregate Approximation* (SAX) [6] and the corresponding distance measure for multivariate time series.

In this paper, we investigate the application of *Principal Component Analysis* (PCA) on multivariate time series for the purpose of defining a similarity/distance measure. PCA is a well-known statistical approach that can be used to reduce the dimensionality of a multivariate dataset by condensing a large number of interrelated variables into a smaller set of variates, while retaining as much as possible of the variation present in the original dataset [17]. We review several PCA-based similarity/distance measures that have been proposed in the literature from many diverse fields. Moreover, we extend the well-known statistic in the Statistical Process Control (SPC) community SPE or Q, in order to provide a

distance measure that has never been tested in the context of similarity search. Experiments were performed on four datasets that have been extensively utilized in the literature. PCA-based similarity search is more complicated and usually requires expensive computations, however, it may improve similarity search providing at the same time useful information for post hoc analysis.

In Section II, we briefly provide the background on PCA and we discuss its implications in similarity search. Section III discusses PCA-based similarity/distance measures proposed from diverse fields and introduces a distance measure based on statistical process control theory. In Section IV, we describe the datasets, the methods and the results of the conducted experiments. Finally, conclusions and future work is presented in Section V.

2 PCA-Based Similarity Search

2.1 Background on PCA

Principal Component Analysis is applied on a multivariate dataset, which can be represented as a matrix $X_{n\times p}$. In case of time series, n represents their length (number of time instances), whereas p is the number of variables being measured (number of time series). Each row of X can be considered as a point in p-dimensional space. The objective of PCA is to determine a new set of orthogonal and uncorrelated composite variates $Y_{(j)}$, which are called principal components:

$$Y_{(j)} = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p \qquad (1)$$

where j = 1, 2, …, p. X_i denotes the ith variable. The coefficients a_{ij} are called component weights. Each principal component is a linear combination of the original variables and it is derived in such a manner that its successive component will account for a smaller portion of variation in X. Therefore, the first principal component accounts for the largest portion of variance, the second one account for the largest portion of the remaining variance subject to being orthogonal to the first one and so on. Hopefully, the first (k) components will retain most of the variation present in all of the original variables (p). Thus, an essential dimensionality reduction may be achieved by projecting the original data on the new k-dimensional space, as long as, $k << p$.

The derivation of the new axes (components) is based on Σ, where Σ denotes the covariance matrix of X. Alternatively, this derivation could be based on the correlation matrix, which is equivalent to perform PCA on standardized variables (i.e. variables with mean equal to zero and standard deviation equal to one). Each eigenvector of Σ provides the component weights a_{ij} of the $Y_{(j)}$ component, while the corresponding eigenvalue, denoted λ_j, provides the variance of this component.

Intuitively, PCA transforms a dataset X by rotating the original axes of a p-dimensional space and deriving a new set of axes (components), as in Fig. 1. The component weights represent the angles between the original and the new axes. In particular, the component weight a_{ij} is the cosine of the angle between the ith original axis and the jth component [18]. The values of $Y_{(j)}$ calculated from equation (1) provide the coordinates of the original data in the new space.

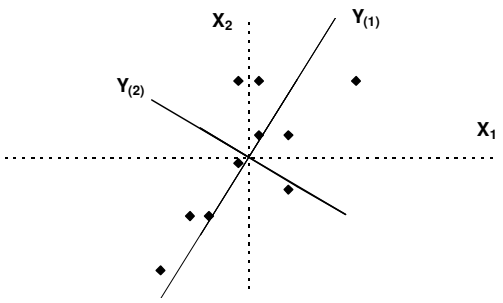

Fig. 1. A multivariate time series consisting of two variables (X_1 and X_2) and ten time instances. Dots represent the time instances, while solid lines represent the principal components that have been derived by PCA. A dimensionality reduction can be achieved, if only the first component $Y_{(1)}$ is retained and data is projected on it.

Conclusively, the application of PCA on a multivariate dataset $X_{n\times p}$ results in two matrices, in particular the matrix of component weights $A_{p\times p}$ and the matrix of variances $\Lambda_{p\times 1}$. In addition to that, the matrix of the new coordinates $Y_{n\times p}$ of the original data can be calculated from A, since $Y = X \cdot A$.

2.2 Implications of PCA in Similarity Search

PCA is applied for dimensionality reduction, and thus it can be considered as a representation scheme. Consequently, data is transformed into a new space and similarity should be based on at least one of the produced matrices, mentioned in the previous paragraph, $A_{p\times p}$, $\Lambda_{p\times 1}$, and $Y_{n\times p}$. The central concept is that, if two multivariate time series are similar, their PCA representation will be similar, that is, the produced matrices will be close enough. Searching similarity based on $A_{p\times p}$, means to compare the angles of principal components derived from two multivariate time series, whereas searching based solely on $Y_{n\times p}$ is useless, since these values are coordinates in different spaces. The matrix $\Lambda_{p\times 1}$ contains information about the shape of the time series and it may be used in conjunction with $A_{p\times p}$ for further distinguishing power. However, this matter needs to be considered on the specific characteristics of each application.

Regarding data volume, PCA representation may achieve essential data reduction, as long as, the number of time instances n is much greater than the number of variables p. Note that a dataset $X_{n\times p}$ is represented by the matrices $A_{p\times p}$ and $\Lambda_{p\times 1}$. Moreover, a further data reduction can be achieved, if only k components are retained, where $k << p$. There are several criteria for determining the number of components to retain, such as the scree graph or the cumulative percentage of total variation [17]. According to the latter criterion, one could select that value for k, for which the first k components retain more than 90% of the total variation present in the original data.

Another implication of PCA-based similarity search is that two multivariate time series with equal number of variables but different lengths can be compared directly, since the dimensions of the produced matrices A and Λ are independent of the lengths (p×p and p×1 respectively).

Another issue that needs to be clarified is the existence of various types of distortions in original data and how these are handled through PCA representation. More specifically, there are four major distortions that ought to be considered, namely, offset translation, amplitude scaling, time warping and noise. Offset translation refers to the case where there are differences in the values of two time series, while the general shape remains similar. PCA representation automatically takes this fact into consideration, since it is based on covariances, which are not affected by the magnitude of the values. This is a potential disadvantage of PCA, if similarity search is to be based also on the magnitude of the values. Amplitude scaling refers to the case where there are differences in the scaling of the values of two time series, while the general shape remains similar. In this case, PCA representation can be based on the correlations among variables, instead of the covariances. This is an alternative way of deriving the principal components, which produce slightly different results, but not essentially different in the context of dimensionality reduction. Time warping distortion may appear globally or locally. In the case of global time warping (i.e. two multivariate time series evolve in different rates), PCA representation is expected to be similar, since the shorter time series can be considered as a systematic random sample of the longer one, resulting to similar covariance matrix. Intuitively, the existence of local time warping distortions may be captured by the covariances of the corresponding variables. Yang & Shahabi [19] provide empirical evidence that PCA addresses the time warping distortion. The distortion of noise is intrinsically handled by PCA, since the discarded principal components account mainly for variations due to noise.

The final implication of applying PCA is that the temporal nature of data is not taken into account while deriving the principal components, since this procedure is based on the covariances among variables.

3 PCA-based Measures

3.1 A Review on PCA-based Measures

There are several PCA-based measures that have been proposed in order to compare two objects, which are in the form of multivariate time series. The main idea is to derive the principal components for each one and then to compare the produced matrices.

Suppose that we have two multivariate time series denoted X_{nxp} and Y_{nxp}. Applying PCA on each one results in the matrices of component weights A_X and A_Y, and variances Λ_X and Λ_Y respectively. All the following measures assume that the number of variables p is the same for all series. This is a rational assumption, since these series are usually generated by the same process within a specific application.

One of the earliest measures has been proposed by Krzanowski [20]. This measure (2) is applicable to time series, although originally it was not applied on such type of data. The proposed approach is to retain k principal components (k << p) and compare the angles between all the combinations of the first k components of the two objects.

$$Sim_{PCA}(X,Y) = trace(A_X^T A_Y A_Y^T A_X) = \sum_{i=1}^{k}\sum_{j=1}^{k} cos^2\theta_{ij}, \quad 0 \le Sim_{PCA} \le k \tag{2}$$

where θ_{ij} is the angle between the ith principal component of X and the jth principal component of Y.

Johannesmeyer [21] modified the previous measure by weighting the angles with the corresponding variances as in (3).

$$S_{PCA}^{\lambda} = \sum_{i=1}^{k}\sum_{j=1}^{k}(\lambda_{Xi}\cdot\lambda_{Yj}\cdot cos^2\theta_{ij}) / \sum_{i=1}^{k}\lambda_{Xi}\cdot\lambda_{Yi}, \quad 0 \le S_{PCA}^{\lambda} \le 1 \tag{3}$$

Singhal & Seborg [22] extend the previous measure by incorporating an extra term in (3), which expresses the distance between the original values of the two objects. This distance factor (S_{dist}) can be particularly useful in case the two objects have similar principal components but the values of their variables are essentially different. In order to find the distance between the two objects, it is required to set one of them as the reference dataset. Then, the Mahalanobis distance of a dataset from the reference is computed as in (4). Y is assumed to be the reference dataset.

$$\Phi = \sqrt{(\bar{x}-\bar{y})^T \Sigma_Y^{*-1} (\bar{x}-\bar{y})} \tag{4}$$

$\bar{x}$ and $\bar{y}$ are vectors that contain the mean values of the variables that consist the datasets X and Y, whereas, Σ_Y^{*-1} is the pseudo-inverse of the covariance matrix of Y.

Assuming Gaussian distribution, the authors propose as a distance factor the probability that the distance is at least Φ units (5).

$$S_{dist} = \sqrt{\frac{2}{\pi}}\int_{\Phi}^{\infty} e^{-z^2/2} dz, \quad 0 \le S_{dist} \le 1 \tag{5}$$

Note that, although S_{dist} represents distance between two objects, it is a similarity measure. As the distance Φ increases, the corresponding probability decreases.

Finally, the proposed measure (6) is the weighted summation of two similarity measures.

$$SF = w_1 S_{PCA}^{\lambda} + w_2 S_{dist}, \quad 0 \le SF \le 1 \ \ and \ \ w_1 + w_2 = 1 \tag{6}$$

Yang & Shahabi [19] propose a similarity measure, Eros, which is based on the acute angles between the corresponding components from two objects X and Y (7). Contrary to the previous measures, all components are retained from each object and their variances form a weight vector w. More specifically, the variances obtained from all the objects in a database are aggregated into one weight vector, which is updated when objects are inserted or

removed from the database. Finally, the authors provide lower and upper bounds for this measure.

$$Eros(X,Y,w) = \sum_{i=1}^{p} w(i) \cdot |cos\theta_i|, \quad 0 \le Eros \le 1 \tag{7}$$

Otey & Parthasarathy [23] define a distance measure in terms of three dissimilarity functions that take into account the differences among the original values, the angles between the corresponding components and the difference in variances. For the first term, authors propose to use either the Euclidean (8) or the Mahalanobis distance (9).

$$D_d(X,Y) = \sqrt{(\mu_X - \mu_Y) \cdot (\mu_X - \mu_Y)^T} \tag{8}$$

$$D_d(X,Y) = \sqrt{(\mu_X - \mu_Y) \cdot \Sigma_{XY}^{-1} \cdot (\mu_X - \mu_Y)^T} \tag{9}$$

where μ_X and μ_Y are the vectors that contain the mean values of the variables that consist the datasets X and Y, whereas Σ_{XY} is the covariance matrix of the combination of datasets X and Y.

The second term is defined as the summation of the acute angles between the corresponding components, given that all components are retained (10).

$$D_r(X,Y) = trace(cos^{-1}(|A_X^T A_Y|)), \quad 0 \le D_r \le p\pi/2 \tag{10}$$

The third term accounts for the differences in the distributions of the variance over the derived components. Consider the random variable V_X having the probability mass function

$$p_X = P(V_X = i) = \lambda_i^X / trace(\Lambda_X).$$

$P(V_X = i)$ represents the proportion of the variance in the direction of the ith principal component.

The difference between the distributions V_X and V_Y is defined as the symmetric relative entropy (11).

$$D_v(X,Y) = \frac{1}{2}(H(p_X \| p_Y) + H(p_Y \| p_X)) \tag{11}$$

The proposed distance measure can be defined in the following two forms:

$$D_\Pi = D_d \cdot D_r \cdot D_v \tag{12}$$

$$D_\Sigma = \beta_0 + \beta_d \cdot D_d + \beta_r \cdot D_r + \beta_v \cdot D_v. \tag{13}$$

Otey & Parthasarathy refer to (12) as their basic formulation, and to (13) as a more flexible form that allows the terms to be weighted differently according to the needs of a given application.

Li & Prabhakaran [24] propose a similarity measure for recognizing distinct motion patterns in motion streams in real time. This measure, which is called k Weighted Angular Similarity (kWAS), can be obtained by applying singular value decomposition on the transformed datasets, X^TX and Y^TY, and retaining the first k components. kWAS is based on the acute angles between the corresponding components weighted by the corresponding eigenvalues (14).

$$\Psi(X,Y) = (1/2)\sum_{i=1}^{k}((\sigma_i / \sum_{i=1}^{n}\sigma_i + \lambda_i / \sum_{i=1}^{n}\lambda_i)|u_i \cdot v_i|), \quad 0 \le \Psi(X,Y) \le 1 \tag{14}$$

where σ_i and λ_i are the ith eigenvalues corresponding to ith eigenvectors u_i and v_i of the matrices X^TX and Y^TY.

When the original datasets are mean centered, the above procedure is equivalent to applying PCA on the original data. The eigenvectors u_i and v_i are the corresponding principal components, while the eigenvalue-based weight in (14) is equal to the one obtained, if σ_i and λ_i are replaced by the variances of the corresponding components. The absolute value implies that the cosine of the acute angles is computed.

In the context of Statistical Process Control, Kano et al. [25] propose a distance measure for the purpose of monitoring process and identifying deviations from normal operating conditions. This measure is based on the Karhunen-Loeve expansion, which is mathematically equivalent to PCA. However, it involves applying eigenvalue decomposition twice during its calculation, which is the most computationally expensive part. At this stage of our research, we decided not to include it in our experiments.

3.2 The Proposed Measure

In this paper, we propose a distance measure that is based on a well-known statistic in multivariate Statistical Process Control (SPC), namely the Squared Prediction Error (SPE) or Q statistic [26]. The calculation of SPE requires applying PCA on a dataset X and retaining the first k principal components. In SPC, this dataset may represent a time period of normal operating conditions of a process. When values that correspond to a new time instance arrive, they are projected on the plane derived by PCA, in order to obtain their new coordinates as in (15).

$$t = y \cdot A \tag{15}$$

where $A_{p\times k}$ is the matrix of principal components, $t_{1\times k}$ and $y_{1\times p}$ are vectors, which consist of the projected and the original values respectively.

In order to obtain the error that this projection introduces to the new values, we calculate the predicted values of y (16).

$$\hat{y} = t \cdot A^T \tag{16}$$

Then, the SPE statistic is defined as in (17). It represents the squared perpendicular distance of a new observation (the values of a new time instance) from the plane. A high value of SPE means that the projection model

is not valid for that observation.

$$SPE = \sum_{j=1}^{p} (y_j - \hat{y}_j)^2 \quad (17).$$

In our case, the objective is to compare two datasets $X_{n\times p}$ and $Y_{m\times p}$ for the purpose of similarity search. We propose to apply PCA on X and calculate SPE for each time instance of Y. The summation of these values can be used as a distance measure between X and Y (18).

$$SPE_{dist} = \sum_{i=1}^{m} \sum_{j=1}^{p} (y_{ij} - \hat{y}_{ij})^2 \quad (18)$$

For all previously discussed measures, we can apply PCA on objects and store the component weight and variance matrices offline. Then, we compare a query object with any one in the database by computing a similarity/distance measure. The advantage of SPE_{dist} is that PCA, which is the most computationally expensive task, is not applied on this query object.

A similar approach can be found in the work of Barbic et al. [27], who propose a technique for the purpose of segmenting motion capture data into distinct motions. However, the authors utilize the squared error of the projected values and not the predicted values, as we propose in our work. Moreover, they focus on an application that involves one multivariate time series, which should be segmented.

4 Experiments

4.1 Datasets

The experiments have been conducted on three real-world datasets and one synthetically created dataset, which have been used extensively in the literature.

The first dataset relates to Australian Sign Language (AUSLAN), which contains sensor data gathered from 22 sensors placed on the hands (gloves) of a native AUSLAN speaker. There are 95 distinct signs, each one performed 27 times. In total, there are 2 565 signs in the dataset. More technical information can be found in [28].

The second dataset, HUMAN GAIT, involves the task of identifying a person at a distance. Data are captured using a Vicon 3D motion capture system, which generates 66 values at each time instance. 15 persons participated in this experiment and were required to walk in 3 sessions, at 4 different speeds, 3 times for each speed. In total, there are 540 walk sequences. More technical information can be found in [29].

The third dataset relates to EEG data that arises from a large study to examine EEG correlates of genetic predisposition to alcoholism. It contains measurements from 64 electrodes placed on the scalp and sampled at 256 Hz (3.9-msec epoch) for 1 second. There are three versions of this dataset, the small, the large and the full, according to the volume of the original data included [30]. We utilized the large dataset, which contains data for 10 alcoholic and 10 control subjects. Each subject was exposed to 3 different stimuli, 10 times for each one. In total, there are 600 EEG's.

Finally, the transient classification benchmark (TRACE) is a synthetic dataset designed to simulate instrumentation failures in a nuclear power plant [31]. There are 4 process variables, which generate 16 different operating states, according to their co-evolvement through time. There is an additional variable, which initially takes on the value of 0, until the start of the transient occurs and its value changes to 1. We retain only that part of data, where the transient is present. For each state, there are 100 examples. In total, there are 1600 examples.

4.2 Method & Measures

In order to compare the performance of the proposed similarity/distance measures, we perform one-nearest neighbor classification and evaluate it by means of classification error rate.

We use 9-fold cross validation for the datasets AUSLAN and HUMAN GAIT taking into account all the characteristics of the experiments, while creating the subsets. The observed differences in the error rates among the various methods were statistically tested. Due to the small number of subsets and to the violation of normality assumption in some cases, Wilcoxon Signed-Rank tests were performed at 5% significance level. For the EEG and TRACE dataset, we use the train and test datasets provided by the authors.

The similarity measures that were tested on our experiments are SimPCA (2), SλPCA (3), Eros (7), DΠ (12), kWAS (14), and SPEdist (18).

SF (6) and DΣ (13) are not included in these experiments due to the fact that they involve tuning parameters for each dataset.

In order to calculate the DΠ (12), we use the definition and the corresponding conventions of the relative entropy H as in [32], since it is not provided by authors.

$$H(p_X \| p_Y) = \sum_x p_X \cdot log(p_X / p_Y) \quad (19)$$

Regarding Eros (7), the weight vector w was computed by averaging the variances of each component across the objects of the training dataset and normalizing them so that $\Sigma w_i = 1$, for i = 1, 2… p. The authors propose alternative ways of computing the weight vector, which have not been tested here.

All other measures require determining the number of components k to be retained. We have conducted classification for consecutive values of k, until the improvement in classification rate was not practically significant. For the maximum value of k that has been used, at least 95% of the total variation is retained for all objects in all datasets except from the EEG. In this case, at least 80% of the total variation is retained for all objects, however, the classification rate does not improve significantly for values of k greater than 12.

For comparison reasons, we also included in the experiments the Euclidean distance. Since this measure requires datasets of equal number of time instances, we decided to apply linear interpolation on the original datasets and set the length of the time series equal to the corresponding mean length (Table 1). The transformed

datasets were utilized only when Euclidean distance was selected for the classification task.

Principal Component Analysis is performed on the covariance matrices. For comparison reasons, the similarity measures DΠ (12), kWAS (14), and SPEdist (18) were computed on the mean centered values.

All the necessary codes and experiments were developed in MATLAB, whereas the statistical analysis was performed in SPSS.

TABLE 1. DESCRIPTION OF DATASETS

Dataset	# of variables	mean length	# of objects	# of classes
AUSLAN	22	57	2565	95
HUMAN GAIT	66	133	540	15
EEG	64	256	600	2
TRACE	4	250	1200	16

4.3 Results

The classification rates are presented in the form of percentages in Table 2. Although we have run the experiments (when required) for various values of k (the number of principal components retained), we present only the best classification accuracies. Moreover, these values of k should not be considered as the best ones, since slightly different ones may result in practically comparable performances.

TABLE 2. CLASSIFICATION ERROR RATES (%)

Dataset / Measure	ASL	HUMAN GAIT	EEG	TRACE
Euclidean	13.8	6.3	30.2	3.9
Sim_{PCA}	(1) 12.0	(4) 1.3	(12) 0.3	(1) 50.3
S^{λ}_{PCA}	(4) 11.5	(3) 18.5	(10) 16.5	(3) 31.4
Eros	9.7	3.7	14.8	21.4
kWAS	(4) 9.4	(5) 3.5	(12) 26.2	(3) 32.9
D_{Π}	83.8	40.6	-	19.75
SPE_{dist}	(4) 8.0	(4) 3.5	(8) 2.0	(2) 48.3

Numbers in parentheses indicate the number of principal components retained. Lack of number indicates measures that exploit all components.

First, we will compare similarity/distance measures with respect to each dataset separately. For AUSLAN dataset, it seems that SPEdist produces at least similar results to Eros and kWAS. Statistically testing their differences across the specific subests, SPEdist produces better results than all ($p<0.05$). The low performance of DΠ suggests that a different form of the proposed ones in [23] should be investigated, with respect to this dataset. Regarding the HUMAN GAIT dataset, SimPCA, Eros, kWAS and SPEdist seems to provide the best results. Statistically testing their differences across the specific subests, SimPCA produces better results than all, whereas the performance of Eros, kWAS and SPEdist is statistically similar ($p>0.05$). For EEG dataset, SimPCA and SPEdist seems to provide considerably better results than other measures, with classification error rates of 0.3% and 2% respectively, when the next best performing measure, Eros, presents a classification error rate of 14.8%. Finally, for the TRACE dataset, which consist of only 4 variables, Euclidean distance, a non-PCA-based measure, performs essentially better than all measures with 3.9% classification error rate. The next best performing measures are DΠ and Eros with classification error rates of 19.75% and 21.4% respectively.

5 Conclusion – Future work

In this paper, we discussed the application of Principal Component Analysis (PCA) on multivariate time series datasets for the purpose of similarity search. More specifically, there were three main contributions. First, we reviewed several PCA-based similarity/distance measures that have recently proposed from diverse fields, not necessarily within data mining context. Second, we proposed a distance measure, which does not require for the query object to be PCA-represented. This measure seems to be promising with respect to classification accuracy. Third, we provided comparative results from experiments performed on four widely utilized datasets by applying several of the proposed measures.

Future work will be focused on conducting extensive experiments on more datasets. In addition to that, we will continue review PCA-based similarity/distance measures, since this technique attracts the attention from several diverse fields. Principal Component Analysis has not been extensively explored in the context of similarity search in multivariate time series and hence, it has the potential to offer more in the Data Mining field.

6 References

[1] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231-268, Mar. 2001.

[2] L. Chapman and J. E. Thornes, "The use of geographical information systems in climatology and meteorology." *Progress in Physical Geography*, vol. 27, no. 3, pp. 313-330, Sept. 2003.

[3] M. Kano, K. Nagao, S. Hasebe, I. Hashimoto, H. Ohno, R. Strauss, and B. R. Bakshi, "Comparison of multivariate statistical process monitoring methods with applications to the Eastman challenge problem," *Computers & Chemical Engineering*, vol. 26, no. 2, pp. 161-174, Feb. 2002.

[4] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: a review," *IEE Proc. Vision Image and Signal Processing*, vol. 152, no. 2, pp. 192-204, Apr. 2005.

[5] K. Forbes and E. Fiume, "An Efficient Search Algorithm for Motion Data Using Weighted PCA," in *Proc. ACM SIGGRAPH/ Eurographics Symp. on Computer Animation*, Los Angeles, CA, 2005, pp. 67-76.

[6] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proc. 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, San Diego, CA, 2003, pp. 2-11.

[7] D. Gunopoulos and G. Das, "Time series similarity measures," *Tutorial notes 6th ACM SIGKDD Int. Conf.*

on Knowledge discovery and data mining, Boston, MA, 2000, pp. 243-307.

[8] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in *Proc.4th Int. Conf. FODO*, Evanston, IL, 1993, pp. 69-84.

[9] M. Vlachos, M. Hadjieleftheriou, D. Gunopoulos and E. Keogh, "Indexing multidimensional time-series," *The VLDB Journal*, vol. 15, no. 1, pp. 1-20, Jan. 2006.

[10] D. Buzan, S. Sclaroff, and G. Kollios, "Extraction and clustering of motion trajectories in video," in *Proc. 17th ICPR*, Boston, MA, 2004 vol. 2, pp. 521-524.

[11] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proc. ACM SIGMOD Int. Conf. on Management of data*, Baltimore, MD, 2005, pp. 491-502.

[12] M. Vlachos, M. Hadjieleftheriou, D. Gunopoulos, and E. Keogh, "Indexing multidimensional time-series with support for multiple disatance measures," in *Proc. 9th ACM SIGKDD*, Washington, D.C., 2003, pp. 216-225.

[13] T. Kahveci, A. Singh, and A. Gurel, "Similarity searching for multi-attribute sequences," in *Proc. 14th SSDBM, 2002*, Edinburg, Scotland, 2002, pp. 175-184.

[14] S. L. Lee, S. J. Chun, D. H. Kim, J. H. Lee, and C. W. Chung, "Similarity search for multidimensional data sequences," in *Proc. ICDE*, San Diego, CA, 2000, pp. 599-608.

[15] Y. Cai and R. Ng, "Indexing spatio-temporal trajectories with Chebyshev polynomials," in *Proc. ACM SIGMOD*, Paris, France, 2004, pp. 599-610.

[16] P. Bakalov, M. Hadjieleftheriou, E. Keogh, and V. J. Tsotras, "Efficient trajectory joins using symbolic representations," in *Proc. 6th Int. Conf. on Mobile data management*, Ayia Napa, Cyprus, 2005, pp. 86-93.

[17] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer, 2004, ch. 1.

[18] J. C. Gower, "Multivariate Analysis and Multidimensional Geometry," *The Statistician*, vol. 17, no. 1, pp. 13-28, 1967.

[19] K. Yang and C. Shahabi, "A PCA-based similarity measure for multivariate time series," in *Proc. 2nd ACM MMDB*, Washington, D.C., 2004, pp. 65-74.

[20] W. Krzanowski, "Between-groups comparison of Principal Components," *JASA*, vol. 74, no. 367, pp. 703-707, Sept. 1979.

[21] M.C. Johannesmeyer, "Abnormal situation analysis using pattern recognition techniques and historical data," M.S. thesis, UCSB, Santa Barbara, CA, 1999.

[22] A. Singhal and D. E. Seborg, "Clustering multivariate time-series data," *Journal of Chemometrics*, vol. 19, no. 8, pp. 427-438, Aug. 2005.

[23] M. E. Otey & S. Parthasarathy, "A dissimilarity measure for comparing subsets of data: application to multivariate time series," in *Proc. ICDM Workshop on Temporal Data Mining*, Houston, TX, 2005,

[24] C. Li & B. Prabhakaran, "A similarity measure for motion stream segmentation and recognition," in *Proc.6th Int. Workshop MDM/KDD*, Chicago, IL, 2005, pp. 89-94.

[25] M. Kano, K. Nagao, H. Ohno, S. Hasebe, and I. Hashimoto, "Dissimilarity of process data for statistical process monitoring," in *Proc. IFAC Symp. ADCHEM*, Pisa, Italy, 2000, vol. I, pp. 231-236.

[26] J. Kresta, J. F. MacGregor, and T. E. Marlin, "Multivariate statistical monitoring of process operating performance," *The Canadian Journal of Chemical Engineering*, vol. 69, pp. 35-47, Feb. 1991.

[27] J. Barbic, A. Safonova, J. Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard, "Segmenting motion capture data into distinct behaviors," in *Proc. Graphics Interface Conf*, London, Ontario, Canada 2004, pp. 185-194.

[28] M. W. Kadous, "Temporal Classification: extending the classification paradigm to multivariate time series." Ph. D. Thesis, School of Computer Science and Engineering, University of New South Wales, 2002.

[29] R. Tanawongsuwan and A. Bobick, "Performance analysis of time-distance gait parameters under different speeds," in *Proc. 4th Int. Conf. AVBPA*, Guilford, UK, 2003, pp. 715-724.

[30] H. Begleiter (1999). The UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science.

[31] D. Roverso, "Plant diagnostics by transient classification: the Aladdin approach," *International Journal of Intelligent Systems*, vol. 17, no. 8, pp. 767-790, Aug. 2002.

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons Inc., 1991, ch.2.